Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action

Pauline T. Kim*

The growing use of predictive algorithms is increasing concerns that they may discriminate, but mitigating or removing bias requires designers to be aware of protected characteristics and take them into account. If they do so, however, will those efforts be considered a form of discrimination? Put concretely, if model-builders take race into account to prevent racial bias against Black people, have they then engaged in discrimination against white people? Some scholars assume so and seek to justify those practices under existing affirmative action doctrine. By invoking the Court's affirmative action jurisprudence, however, they implicitly assume that these practices entail discrimination against white people and require special justification. This Article argues that these scholars have started the analysis in the wrong place. Rather than assuming, we should first ask whether particular race-aware strategies constitute discrimination at all. Despite rhetoric about colorblindness, some forms of race consciousness are widely accepted as lawful. Because creating an algorithm is a complex, multi-step process involving many choices, tradeoffs and judgment calls, there are many different ways a designer might take race into account, and not all of these strategies entail discrimination against white people. Only if a particular strategy is found to discriminate is it necessary to scrutinize it under affirmative action doctrine. Framing the analysis in this way matters, because affirmative action doctrine imposes a heavy legal burden of justification. In addition, treating all race-aware algorithms as a form of discrimination reinforces the false notion that leveling the playing field for disadvantaged groups somehow disrupts the entitlements of a previously advantaged group. It also mistakenly suggests that prior to considering race, algorithms are neutral processes that uncover some objective truth about merit or desert, rather than properly understanding them as human constructs that reflect the choices of their creators.

DOI: https://doi.org/10.15779/Z387P8TF1W

[*] Daniel Noyes Kirby Professor of Law, Washington University School of Law, St. Louis, MO. This research was partially supported by a National Science Foundation-Amazon grant, number IIS-1939677. I owe thanks to Allia Howard, Maryl Evans, Yaseen Morshed, Maggie Rick, and Sara Hubaishi for excellent research assistance. I am also grateful to Daniel Harawa, Travis Crum, Kyle Rozema, Scott Baker, Kimberly Norwood, Kevin Collins, Peggie Smith, Eugene Vorobeychik, Andrew Estornell, Sanmay Das, Patrick Fowler, C.J. Ho, Brendan Juba, Manish Raghavan, Solon Barocas, Andrew Selbst, Deborah Widiss, Deborah Hellman, Catherine Powell, Ari Waldman and participants at the 2020 Colloquium on Scholarship in Employment and Labor Law, the 2021 Privacy Law Scholars Conference, the Race, Tech and Law Convening, and Washington University School of Law faculty workshops for many helpful conversations about the issues discussed in this Article.

INTRODUCTION

It is now widely recognized that algorithms can discriminate against disadvantaged groups. As reliance on these tools to make decisions about people increases, there are growing concerns that they will reproduce or worsen inequality in domains like housing, employment, credit, and criminal law enforcement.[1] Numerous empirical studies have documented instances of machine learning algorithms producing race- or gender-biased results,[2] such that the question is no longer whether algorithms can discriminate, but what to do about it. Data scientists and machine learning experts are working to devise technical solutions to prevent discrimination,[3] proposing competing methods for

[1] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (employment); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017) (employment); Kristin Johnson, Frank Pasquale & Jennifer Chapman, *Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation*, 88 FORDHAM L. REV. 499 (2019) (credit); Rashida Richardson, Jason M. Schwartz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15 (2019) (criminal law enforcement); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017) (criminal law enforcement); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291 (2020) (criminal law enforcement); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633 (2017) (immigration).

[2] *See, e.g.*, Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMC'NS ACM 44 (2013); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROCS. MACH. LEARNING RSCH. 1 (2018); Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove & Aaron Rieke, *Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes*, 3 PROCS. ACM ON HUM.-COMPUT. INTERACTION 1 (2019); Amit Datta, Anupan Datta, Jael Makagon, Deirdre K. Mulligan & Michael Carl Tschantz, *Discrimination in Online Advertising a Multidisciplinary Inquiry*, 81 PROCS. MACH. LEARNING RSCH. 1 (2018); Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=Gg58888u2U5db3W3CsuKrD0LD_VQJReQ [https://perma.cc/4FS8-DTFF].

[3] For a small sampling of work in this area, see, for example, Irene Y. Chen, Fredrik D. Johansson & David Sontag, *Why Is My Classifier Discriminatory?*, ARXIV :1805.12002 [CS, STAT] (2018), discussing data collection; Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV:1808.00023 [CS] (2018); Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, ITCS (2012), discussing task specific metrics; Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and Removing Disparate Impact*, ACM 259 (2015); Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, ARXIV:1610.02413 [CS] (2016); Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Wwithout Discrimination*, 33 KNOWLEDGE AND & INFO. SYS. 1 (2012); Toshihiro Kamishima et al., Shotaro Akaho & Jun Sakuma, *Fairness-Aware Learning Tthrough Regularization Approach*, 2011 IEEE 11TH INT'L CONF. ON DATA MINING WORKSHOPS 643 (2011); Michael Kearns, Seth Neel, Aaron Roth & Zhiwei Steven Wu, *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*, PROCS. 35TH INT'L CONF. ON MACH. LEARNING ARXIV:1711.05144 [CS] (2018); Zachary C. Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, 32ND CONF. ON NEURAL INFO. PROCESSING SYS. ARXIV:1711.07076 [CS, STAT] (2018); Joshua R. Loftus, Chris Russell, Matt J. Kusner & Ricardo Silva, *Causal Reasoning for Algorithmic Fairness*, ARXIV:1805.05859 [CS] (2018); Jialu Wang, Yang Liu &

ensuring algorithmic fairness. Although there is considerable disagreement over how best to define fairness, consensus has emerged on one point—namely, that simply blinding a model to sensitive characteristics like race or sex will not prevent these tools from having discriminatory effects.[4] Not only can biased outcomes still occur, but discarding demographic information makes bias harder to detect,[5] and, in some cases, could make it worse.[6]

In order to mitigate or prevent algorithmic bias, designers must be aware of and take into account protected characteristics. Because building fair algorithms requires explicit consideration of race, scholars have begun to question whether these strategies are legal under antidiscrimination law.[7] The concern is that by taking race into account, these efforts will themselves be considered a form of intentional discrimination forbidden by law.[8] To put it concretely, if model builders take race into account to prevent racial bias against Black people, have they then engaged in discrimination against white people?[9] What strategies can they employ to reduce discriminatory impacts on historically marginalized racial groups without running afoul of the law?

Some researchers have assumed that the law prohibits any consideration of race in decision-making. If true, many of the de-biasing strategies developed by computer scientists would be doomed to practical irrelevance. More recently,

Caleb Levy, *Fair Classification with Group-Dependent Label Noise*, CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 526 (2021); Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez & Krishna P. Gummadi, *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment*, INT'L WORLD WIDE WEB CONF. COMM. 1171 (2017).

[4] *See, e.g.*, Corbett-Davies & Goel, *supra* note 3; Dwork et al., *supra* note 3; Hardt et al., *supra* note 3; Kamishima et al., *supra* note 3; Loftus et al., *supra* note 3; Anya E. R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257 (2020); Yang & Dobbie, *supra* note 1; Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

[5] *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ashesh Rambachan, *Algorithmic Fairness*, 108 AEA PAPERS & PROCS. 22 (2018).

[6] *See, e.g.*, Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019) (finding that excluding characteristics, like race, from mortgage data led to increased pricing gaps).

[7] *See* Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, 797 ACM 8–9 (2017); Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389, 1392 (2019).

[8] *See* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1094–95 (2019); Sandra G. Mayson, *Bias in, Bias Out*, 128 YALE L.J. 2218, 2230, 2262–63 (2019).

[9] Throughout this Article, I use hypotheticals involving measures taken to reduce bias against Black people and the potential legal claims by white plaintiffs challenging those efforts. I do so primarily for ease of reference, and not to suggest that challenges of addressing racial bias are solely a Black-white issue. Bias in predictive algorithms can affect other racial groups and legal challenges to race-conscious remedies have not been brought exclusively by white plaintiffs, and thus, the analysis here extends to situations involving other forms of race or ethnic bias.

scholars like Jason Bent[10] and Daniel Ho and Alice Xiang[11] have sought to defend race-aware algorithms under the Supreme Court's affirmative action doctrine.[12] Implicit in their arguments is the assumption that by taking race into account, these de-biasing strategies constitute a form of disparate treatment or racial classification that would be unlawful unless justified under the Court's affirmative action doctrine.

This Article argues that these scholars have started the analysis in the wrong place. Rather than assuming that any race-aware algorithm requires special justification under affirmative action doctrine, we should *first* ask whether taking account of race constitutes discrimination at all. Under current law, not all race-conscious efforts to mitigate bias trigger legal scrutiny. Only *after* a particular strategy has been found to constitute disparate treatment or a racial classification does the heightened scrutiny applied to affirmative action plans kick in.

This point is often overlooked because the Court's affirmative action jurisprudence is sometimes assumed to impose a requirement of strict colorblindness. In fact, as numerous scholars have pointed out, the law does not categorically prohibit race consciousness.[13] Both private and government decision-makers routinely use information about race in ways that trigger no particular legal concern. Practices such as collecting demographic information or using racial characteristics in suspect profiles are so commonplace that they are rarely remarked upon, let alone subject to legal challenge.[14] And courts have found some race-conscious actions, like an employer's efforts to improve the racial diversity of its applicant pool, do not constitute discrimination and are legally permissible.[15] What triggers the special scrutiny articulated in the Court's affirmative action cases is not mere race-awareness, but specific ways race is used that constitute disparate treatment or racial classifications.[16]

[10] Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 825–41, 852 (2020).

[11] Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, 2020 U. CHI. L. REV. ONLINE 134, 136 (2020).

[12] Other scholars, like Sandra Mayson and Anupam Chander, characterize any attention to race in the model-building process as "algorithmic affirmative action" without discussing the legality of these strategies. *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1040–42 (2017) (using "affirmative action" in its broadest sense to include any proactive practices to correct deficiencies in equality of opportunity); Mayson, *supra* note 8, at 2262 (using the term algorithmic affirmative action to describe and asses the normative desirability of different strategies without considering their legality).

[13] *See, e.g.*, Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV. 1115, 1115 (2016); Justin Driver, *Recognizing Race*, 112 COLUM. L. REV. 404, 404 (2012); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 819 (2020) ("[T]he doctrine's resistance to the use of racial classifications is not categorical."); Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 505–06 (2003).

[14] *See infra* Part III.B.

[15] *See infra* Part III.A.

[16] The question of when race-conscious action requires justification is framed somewhat differently depending upon the source of law. For example, under Title VII of the Civil Rights Act of

Not all efforts to redress racial inequities amount to disparate treatment or racial classifications. When fairness considerations lead a decision-maker to revise its processes or remove unnecessary barriers that harm disadvantaged groups, it has not engaged in discrimination. Its actions do not involve making decisions about individuals by preferring one group over another. Instead, they simply discard arbitrary obstacles in order to level the playing field for all. Similarly, many efforts to eliminate problematic features that cause bias in algorithms are more accurately characterized as non-discriminatory efforts to remove unfairness, rather than "reverse discrimination" that must meet the stringent requirements imposed by affirmative action doctrine.[17]

This point is obscured by the tendency to assume that algorithms have a fixed form, rather than recognizing them as malleable and contingent on the choices made by their creators. In popular and legal discourse, the algorithm is imagined as an objective thing, as if a correct solution exists to every prediction problem and considerations of group fairness somehow represent a deviation from the "true" solution.[18] In fact, however, the model-building process is a complex one, involving multiple decisions. None of them are inevitable, and every one potentially impacts fairness.[19] The designers must make difficult choices each step of the way, involving tradeoffs, subjective judgments and the weighing of values. Each of these choices can be consequential in shaping the final model and the results it produces.

These observations lead to two important implications relevant to the legality of race-aware algorithms. First, that there is no single, definitive model that exists prior to taking racial equity concerns into account, and therefore, no clear baseline against which outcomes under a racially de-biased model can be compared. Given the numerous choices involved in the model-building process, multiple solutions will exist for any given prediction problem. Those competing models may perform equally well and yet produce different predictions in

1964, affirmative action plans require legal justification when they result in disparate treatment. Thus, white plaintiffs challenging such plans must show that the decision-maker took an adverse action against them because of their race. Under the Equal Protection Clause, the focus is on racial classifications. It is the use of racial classification by a government actor that triggers strict scrutiny.

[17] Scholars who have invoked the idea of algorithmic affirmative action have not been entirely clear about what strategies their analysis encompasses, although they generally seem to lump together any awareness of race in the model-building process. For example, although Bent acknowledges that fairness strategies can come into play at different points in the process and take a variety of forms, in his legal discussion he subsumes them into a generic "race-aware model" and concludes that any such model constitutes a prima facie violation of discrimination law. Bent, *supra* note 10, at 823–25.

[18] *Cf.* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U. CAL. DAVIS L. REV. 653, 661 (2017) (describing how legal scholars treat machine learning "as a fully formed black box").

[19] *See* Barocas & Selbst, *supra* note 1, at 729; Lehr & Ohm, *supra* note 18, at 677–93; Deven R. Desai & Joshua A. Kroll, *Trust but Verify: A Guide to Algorithms and the Law*, 31 HARV. J. L. & TECH. 1, 4 (2017) (noting the tendency of both critics and advocates to "stray into uncritical deference" to algorithms as "infallible science").

individual cases.[20] Because there is no single "correct" model against which to compare a de-biased model, specific individuals cannot necessarily claim that they had some entitlement or settled expectation that was disrupted by efforts to reduce racial bias.

The second implication is that efforts to make a model less biased could involve taking race into account in many different ways. Exactly when and how a given de-biasing strategy does so is critically important for judging its legality. This Article argues that some strategies, for example, addressing data limitations or reconsidering how a problem is defined, do not amount to disparate treatment or involve racial classifications at all. Consequently, arguments about whether they meet the demanding standards of affirmative action doctrine are beside the point. No particular legal scrutiny is warranted because they do not constitute discrimination in the first place.

Thus, in contrast to scholars who defend race-aware algorithms as lawful under the Court's affirmative action cases, I argue that it is important to recognize that some de-biasing strategies do not constitute discrimination at all. The difference between these two approaches is not merely semantic. From a doctrinal perspective, defending a strategy under affirmative action doctrine entails a heavy burden of justification, making a race-aware model presumptively unlawful unless a demanding legal standard is met. Even if the standard can be met, as a practical matter, this additional burden may discourage developers from voluntarily trying to identify and address sources of bias.

On a conceptual level, characterizing race-aware strategies as non-discriminatory rather than justifiable under affirmative action doctrine also matters. The affirmative action frame reinforces the false notion that any steps taken to reduce bias or level the playing field for disadvantaged groups inherently harms white people and therefore requires special justification. It also plays into a common misconception that algorithms are neutral and objective tools that precisely measure merit or desert, rather than entirely human constructs that reflect the choices of their creators.

The affirmative action frame is particularly inapt in the context of criminal law enforcement, which has occupied a good portion of the debates around algorithmic fairness. Unlike the typical settings for affirmative action challenges, which involve distributing resources or opportunities, criminal law enforcement entails punitive sanctions and a cascading set of damaging collateral consequences.[21] Given that communities of color are disproportionately targeted

[20] *See* Charles T. Marx, Flavio du Pin Calmon & Berk Ustun, *Predictive Multiplicity in Classification*, ARXIV:1909.06677 [CS, STAT] 1 (2020).

[21] *See, e.g.*, *id.*; Eisha Jain, *Prosecuting Collateral Consequences*, 104 GEO. L.J. 1197 (2016); Gabriel J. Chin, *Race, The War on Drugs, and the Collateral Consequences of Criminal Conviction*, 6 J. GENDER RACE & JUST. 255 (2002); Michael Pinard, *Collateral Consequences of Criminal Convictions: Confronting Issues of Race and Dignity*, 85 N.Y.U. L. REV. 457 (2010); Michael Pinard, *Criminal Records, Race and Redemption*, 16 N.Y.U. J. LEGIS. & PUB. POL'Y 963 (2013).

by police and prosecutors[22] and are *over*-represented, not under-represented in the system, it makes little sense to judge efforts to reduce racial bias in this context as if they somehow discriminated against white defendants.

Before developing these arguments in detail, a couple of preliminary caveats are necessary. First, although algorithmic biases based on sex, age, disability, and other protected characteristics are also concerns, this Article centers the discussion on race. Issues surrounding race are both highly salient and politically fraught in American society. This country's long history of slavery, segregation, racially exclusionary immigration policies, differential policing, and private discrimination remains visible in the stark racial disparities that persist in health, education, housing, employment, financial stability, and incarceration. These inequities make addressing racial discrimination particularly pressing, but at the same time, U.S. law is deeply and particularly suspicious of the use of race in decision-making.[23] Thus, race poses the most challenging instance for determining the legality of strategies intended to reduce or remove bias from algorithms.

Second, the term "race-aware algorithms" is admittedly a misnomer. Computers do not have awareness or consciousness the way humans do, nor do they act with intentionality in any sense relevant to antidiscrimination law.[24] I use the term "race-aware algorithm" as shorthand for the state of mind of the humans who create the algorithm. It refers to designers who are conscious of racial considerations when making choices in building a model—hence, I also refer to "race-conscious model-building." While racial considerations may come into play at many points, one particular choice concerns whether a model will have access to information about race at the moment it makes predictions about new cases. This specific type of strategy raises distinctive issues, and, to that extent, I specifically note when models use race at prediction time.

This Article proceeds as follows. Part I briefly canvasses the evidence of algorithmic bias and the technical responses that have developed in response. Part II discusses in greater detail the complexities of the model-building process and the implications for evaluating the legality of race-conscious interventions to remove bias. In Part III, I analyze existing antidiscrimination law, focusing first on Title VII as an example of statutory prohibitions and then on constitutional doctrine developed under the Equal Protection Clause. This

[22] *See* MICHELLE ALEXANDER, THE NEW JIM CROW: MASS INCARCERATION IN THE AGE OF COLORBLINDNESS 32 (2010).

[23] For example, under the Equal Protection Clause, racial classifications are subject to strict scrutiny. *See, e.g.*, Loving v. Virginia, 388 U.S. 1 (1967); Adarand Constructors, Inc. v. Pena, 515 U.S. 200 (1995). In comparison, sex classifications face a less demanding intermediate level of scrutiny. *See, e.g.*, Craig v. Boren, 429 U.S. 190 (1976); United States v. Virginia, 518 U.S. 515 (1996). Classifications based on age and disability are not subject to any heightened level of review. *See, e.g.*, Massachusetts Bd. Ret. v. Murgia, 427 U.S. 307 (1976) (age); City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432 (1985) (disability).

[24] Huq, *supra* note 8, at 1089.

analysis shows that race-conscious decision-making is not categorically prohibited, nor does it automatically trigger heightened legal scrutiny. Instead, whether a particular form of race consciousness is lawful or not depends on when and how race is taken into account. Part IV applies these insights to a handful of algorithmic de-biasing strategies, arguing that many do not constitute disparate treatment at all, while others are likely legally impermissible. In between lies a gray area of legal uncertainty, but even there, strong arguments exist that some strategies that involve taking race into account at prediction time do not constitute disparate treatment or racial classifications. In Part V, I consider why this matters, arguing that for both doctrinal and rhetorical reasons it is important to distinguish non-discriminatory uses of race, which operate to remove existing sources of bias, from the types of affirmative action plans that are perceived as entailing special preferences for certain groups. I also briefly consider whether the changed composition of the Supreme Court affects any of the legal analyses herein.

I.
ALGORITHMIC BIAS AND TECHNICAL RESPONSES

A growing literature highlights ways that predictive algorithms can systematically disadvantage subordinated groups. Safiya Noble and Ruha Benjamin have documented how the algorithms that power online searches reproduce racism and other forms of inequity, reinforcing oppression of marginalized groups.[25] Other scholars have shown that recommender systems deliver employment and housing ads to online audiences skewed along race and gender lines,[26] or suggest that people with African-American-associated names have criminal records when they do not.[27] Many additional examples exist. A recruitment algorithm systematically downgraded women candidates for computer programming positions because it was trained using a dataset

[25] *See, e.g.*, SAFIYA UMOJA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); RUHA BENJAMIN, RACE AFTER TECHNOLOGY: ABOLITIONIST TOOLS FOR THE NEW JIM CODE (2019).

[26] *See* Muhammad Ali et al., *supra* note 2, at 1 (finding significant skew along race and gender lines in delivery of employment and housing ads on Facebook). *See also* Piotr Sapiezynski, Avijit Ghosh, Levi Kaplan, Alan Mislove & Aaron Rieke, *Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, ARXIV:1912.07579 [CS] 8 (2019) (reporting experimental results showing that neutral targeting criteria can still result in Facebook ads being delivered to audiences biased along lines of gender, race, age and political views); Ava Kofman & Ariana Tobin, *Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement*, PROPUBLICA (Dec. 13, 2019), https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement [https://perma.cc/3K5H-Z4PD] (providing examples of biased delivery of job advertisements on Facebook).

[27] *See* Sweeney, *supra* note 2, at 46–47.

composed primarily of men.[28] A selection algorithm disfavored women and racial minorities for medical school admission based on past discriminatory practices.[29] Facial recognition systems made far more mistakes in identifying people with darker skin.[30] A tool allocating health care directed greater resources to white patients than Black patients with the same level of need.[31] An algorithm used to inform bail decisions over-predicted recidivism risks for Black suspects as compared with white suspects who had been arrested.[32]

This growing body of evidence of the risks of algorithmic discrimination has shifted the conversation from *whether* algorithms can discriminate to *what to do about it*. While the legal literature has debated whether or how existing antidiscrimination laws apply to automated decision tools,[33] computer scientists have focused on developing methods to remove bias and ensure that algorithms are fair.[34] These efforts are complicated by the ambiguity surrounding the meaning of "bias." At a general level, "bias" can refer to any algorithm that produces a disparate impact. Racially skewed outcomes can occur for different reasons, however, and the underlying cause may affect judgments about whether they are normatively unfair or legally impermissible.

The notion of bias in algorithms encompasses both statistical bias and societal bias.[35] Statistical bias can result when the data used to train the model are unrepresentative of the population or contain systematic errors. It can also occur if the data encode human biases, such as supervisor evaluations of work performance or caseworker assessments of gang involvement that are shaped by implicit biases. These types of data problems undermine the accuracy of a model and can harm already disadvantaged groups without justification.

However, algorithms may also produce skewed predictions because they reflect societal bias, accurately reproducing real differences between racial groups. Sometimes, the choice of the target variable may lead a model to capture

[28] Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://perma.cc/H4NM-48QP].

[29] Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657, 657 (1988).

[30] Buolamwini & Gebru, *supra* note 2, at 1.

[31] Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 448–49 (2019).

[32] Angwin et al., *supra* note 2.

[33] *See, e.g.*, Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Barocas & Selbst, *supra* note 1; Kim, *supra* note 1; James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164 (2017); Michael Selmi, *Algorithms, Discrimination and the Law*, 82 OHIO STATE L.J. 611 (2021); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395 (2018).

[34] *See supra*, note 3 and sources cited therein.

[35] *See* Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour & Kristian Lum, *Algorithmic Fairness: Choices, Assumptions, and Definitions*, 8 ANN. REV. STATS. & ITS APPLICATION 141, 144 (2021).

existing patterns of disadvantage and segregation. For example, an algorithm designed to recommend applicants based on what happened to similar applicants in the past will discriminate if that selection process was biased.[36] The model is accurate in the sense that it performs the defined prediction task well, but its outputs are shaped by pre-existing biases. Similarly, a model may predict a higher risk of loan default for Black borrowers because they in fact earn less money due to discrimination in the labor market. Although the model may be statistically sound, its use raises questions about whether it is fair to rely on accurate predictions that rest on discrimination by others.[37]

In this Article, I do not try to resolve these types of normative questions. Instead, I focus on a different set of questions. If the people designing or deploying a predictive algorithm wish to avoid or reduce disparate impacts on historically subordinated groups, what steps are they legally permitted to take? If they discover that a model has an unintended racial impact, what can they do in response? Given this focus, I use the term "bias" broadly to refer to any observed racially disparate impact regardless of the cause. And I refer to efforts to remove or reduce that impact as strategies for de-biasing or mitigating bias in the model, without making any assumptions or judgments about the reasons the bias occurs.

Computer scientists have proposed a wide range of strategies for de-biasing algorithms, generating a rich literature on algorithmic fairness and offering competing strategies for achieving fairness goals. One of the difficulties they have confronted is that no consensus exists on how to define fairness or what constitutes non-discrimination. Researchers have offered multiple ways of formalizing these concepts,[38] but these definitions are often incompatible, such that it is not possible to simultaneously satisfy them all.[39]

There is, however, one point on which there is consensus. Merely blinding an algorithm will not prevent bias.[40] Because race is often correlated with other personal characteristics or behaviors, any reasonably rich dataset will contain features that, either singly or in combination, can act as stand-ins. For example,

[36] *See* Lowry & Macpherson, *supra* note 29, at 657 (reporting that a computer program used to screen applicants to medical school discriminated against women and racial minorities because it reproduced past biased decisions by a selection panel).

[37] *See* Deborah Hellman, *Big Data and Compounding Injustice*, J. MORAL PHIL. (forthcoming 2021) (arguing that when data reflect social inequities that result from injustice, predictions that rely on that data can compound injustice).

[38] *See, e.g.*, Huq, *supra* note 8, at 1115 (referencing 21 different definitions of fairness); Mayson, *supra* note 8, at 2226; Mitchell et al., *supra* note 35, at 147–53.

[39] *See* Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIO. METHODS & RSCH. 3 (2021); Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV:1609.05807 [CS, STAT] (2016).

[40] *See, e.g.*, Corbett-Davies & Goel, *supra* note 3, at 2; Dwork et al., *supra* note 3, at 218, 226; Hardt et al., *supra* note 3, at 1; Kamishima et al., *supra* note 3, at 643; Loftus et al., *supra* note 3, at 1–2; Yang & Dobbie, *supra* note 1, at 315; Kroll et al., *supra* note 4, at 22.

due to patterns of residential segregation, zip code can often be used as a proxy for race. Removing race as a variable will not prevent biased outputs if an algorithm can still rely on zip code to make predictions. As a result, some have argued that both race and all proxies for it should be eliminated from predictive models.[41] However, it is not a simple matter to remove all proxies for race. It is not always intuitively obvious which features can act as a proxy and some of those variables may be relevant to the predicted outcome even though they correlate with race. Because race influences so many aspects of American life, it may be impossible in some situations to remove its correlates and still have a meaningful model.[42] In short, strategies that center on removing race or its proxies from models are of limited utility.

As a result, technical efforts to prevent algorithms from discriminating inevitably need to take race into account. At the outset, information about race is necessary to assess whether a training dataset contains biases or is unrepresentative of the population to be predicted. Beyond concerns about data quality, many other strategies to reduce or remove bias require explicitly taking race into account at some point in the model-building process. In addition, information about race is necessary to audit the impact of algorithms because they can have unexpected consequences when deployed in real-world settings. The critical point is that efforts to diagnose and remove racial bias from an algorithm require an awareness of race.

II.
THE COMPLEX PROCESS OF MODEL-BUILDING

In *Employing AI*, Charles Sullivan asks the reader to engage in a thought experiment.[43] "Imagine," he writes, that "a company . . . effectively delegat[es] all its hiring decisions to a computer. It gives the computer only one instruction: 'Pick good employees.'"[44] The computer, which he names Arti, is given all available data, including traditional human resources data, the employer's operational data, and whatever personal data can be scoured from the Internet. Sullivan then considers what would happen if Arti "go[es] rogue,"[45] selecting employees on the basis of race or sex. His purpose in proposing this thought experiment is to scrutinize existing discrimination doctrine and to expose some of its inadequacies.

While perhaps a useful construct for interrogating current doctrine, Arti also exposes some common misconceptions about algorithms. At the most basic

[41] *See, e.g.*, Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014) (arguing that race-based predictive models are inaccurate).

[42] Yang & Dobbie, *supra* note 1, at 298 (arguing in the context of the criminal system that it is infeasible to build an algorithm with no race-correlated inputs "due to the influence of race in nearly every aspect of American life today").

[43] Sullivan, *supra* note 33, at 395.

[44] *Id.*

[45] *Id.* at 402.

level, an algorithm is nothing more than a series of instructions. The debate around algorithmic discrimination has mostly focused on a particular type of algorithm, machine-learning algorithms, where the instructions are not specified by the programmer in advance. Instead, a predictive model is developed by applying mathematical tools to extract patterns from an existing dataset called the training data. Those observed patterns are then used to make inferences about what will happen in future cases. In the popular imagination, an algorithm is a well-defined tool such that once one identifies a goal like "hire good employees," computer scientists can find the "correct" version of the algorithm. In fact, there is typically no single best solution to the prediction problem.

David Lehr and Paul Ohm have pointed out that legal scholars also sometimes have a "naturalized" view of predictive models that ignores "the intricate processes of machine learning."[46] As they put it, "[o]ut of the ether apparently springs a fully formed 'algorithm,' or 'model,' ready to catch criminals, hire employees, or decide whom to loan money."[47] An algorithm, however, results from a process involving multiple steps, each requiring the designer to make choices about how to build the model.

Importantly, there is no inevitable destination, no uniquely definitive model that represents the "correct" solution to the problem. Instead, each choice along the way involves weighing tradeoffs and exercising judgment. Depending upon the designer's choices, different models will result, and these models may produce different predictions for the same individual.

Lehr and Ohm catalog the multiple steps involved in building machine learning models. First, there is the question of problem formulation,[48] which involves translating high-level goals like "pick good employees" into an optimization problem that a computer can solve. This translation process requires some human to decide what it means to be a "good" employee. The designer must choose whether a "good employee" is someone who is highly productive, will stay on the job for a long time, is creative, or has strong interpersonal skills. There is no "correct" definition, but the designer's choices will in turn affect what the model looks like and which individuals it predicts are good prospects.

These types of questions arise in other contexts as well.[49] For example, should the risk of re-offending be measured by future arrests? Or only convictions? And for any offense or only felonies? Similarly, one must decide when a "default" on a loan has occurred. After one missed payment? Two? Or a dozen? Very often the quality sought to be predicted cannot be directly

[46] Lehr & Ohm, *supra* note 18, at 661.

[47] *Id.* at 668.

[48] *Id.* at 672–76. *See also* Samir Passi & Solon Barocas, *Problem Formulation and Fairness*, PROCS. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 39 (2019) (arguing that formulating data science problems inherently involves ethical and normative questions).

[49] *See, e.g.*, Eaglin, *supra* note 1, at 75–77 (explaining that the task of defining "recidivism" involves subjective choices that relate to important policy decisions).

measured. Someone must decide what observable metric best approximates it by defining a "target variable" that serves as a proxy for the desired quality. So, for example, an algorithm's designers might seek to optimize gross sales or number of years on the job, even though these quantities cannot entirely or accurately capture what makes a "good employee."

In addition to defining the target variable, designers must decide what data sources to use to train a model. They could select existing datasets or collect the data themselves. In choosing a dataset, they must consider factors such as the number of observations included, the number and types of features captured, the reliability of the data, and whether it is representative of the target population. Different datasets will differ on these dimensions, meaning that the designers must make tradeoffs. They may need to weigh, for example, whether to rely on a large dataset with a limited number of features, or a dataset with highly granular information, but few observations of under-represented groups.

After selecting a dataset, decisions must be made about how to utilize the data. Designers must decide what to do about missing or obviously incorrect information. Should they omit those observations from the model-building process or impute values for them? Similar choices must be made about outliers in the data. Extreme values may provide valuable information or may represent exceptional cases that will distort predictions unless excluded. Once these decisions have been made, the designer must select a subsample of the data to "train" the model. From exposure to training data, the model "learns" the optimal prediction rules. The predictive model that results is then applied to the remaining data, the "test data," to gauge its accuracy. The training data is typically chosen at random from the full dataset, but the designer must decide whether to split the data 50/50 or in some other proportion. This decision turns on considerations such as the size of the whole dataset and the distribution of values of key variables within it. And, as discussed in more detail below, the variation between different random draws of the training data can affect the precise model that is generated.

The designer must also decide what type of algorithm to implement. There are different types of models such as logistic regression, random forest, neural networks, etc. Each uses different technical strategies for optimizing the prediction problem. The type of model chosen will again reflect certain tradeoffs. Some models may be inappropriate for predicting certain types of target variables; others may offer varying abilities to trade off different types of errors, or to adjust the parameters of the model. Once again, there is often no single best approach to employ; rather, designers must weigh the alternatives and exercise judgment in selecting the type of model.

While it would be possible to select a model and then just set it loose "in the wild," it would be highly irresponsible to do so. Designers typically "tune" the algorithm by adjusting its parameters, then assess the performance of the model and make further adjustments. Part of this process includes selecting the features to be included, which can affect the accuracy and performance of the model. Model-building is thus an iterative process, in which "an analyst provisionally assesses its performance and often chooses to then re-tune the algorithm, re-train it, and re-assess it. Such a cycle can occur multiple times."[50] And while this description suggests a step-by-step process of development, Lehr and Ohm caution that "much machine learning dances back and forth across [the] steps instead of proceeding through them linearly."[51]

Even after deployment, an algorithm is not a static thing. Its designers will want to observe its operation "in the wild" to determine whether it performs as expected. Real-world conditions may differ from the testing environment, and changing conditions or strategic responses by other actors in the system may degrade the model's accuracy or utility. The model development process thus entails evaluating its actual operations and making adjustments as necessary—perhaps skipping back and revisiting some of the choices made earlier in the process.

As the above sketch of the model-building process demonstrates, creating machine learning models involves an open-ended iterative process. Even with a well-defined objective—something far more precise than "pick good employees"—that process entails the exercise of judgment and the weighing of tradeoffs at many different decision points. Each of these choices shapes the final version of the model and influences the predictions it will make when deployed. In sum, there is no single solution to a prediction problem, but instead a multitude of possible models. Humans must choose *which* model to adopt, a decision that necessarily entails value choices and discretionary judgments.

From these observations follow two important implications relevant to the question of whether race-conscious model-building strategies are lawful. First, because there is no single "correct" model for any given problem, there also is no "true" prediction for any given individual. The choices made in creating a machine learning model will affect the distribution of predicted outcomes, such that a particular person might score highly enough to receive a benefit under one model, but not under another, even before any group fairness considerations are taken into account.

Variations in predicted outcome can result from relatively minor changes in the model-building process. For example, random draws of a training dataset can cause a meaningful amount of variation in the predicted outcomes for a given individual even though all else in the model-building process is the same.[52] Similarly, the inclusion or removal of a *single* person in the model's training data

[50] Lehr & Ohm, *supra* note 18, at 698.

[51] *Id.* at 669.

[52] Andrew Estornell, Sanmay Das, Patrick Fowler, Chien-Ju Ho, Brendan Juba, Pauline Kim & Yevgeniy Vorobeychik, *Individual Impacts of Group Fairness* (2022) (in progress).

can change the outcomes for some other individuals under the resulting model, and this effect occurs with "surprising frequency."[53]

If seemingly minor changes such as the random draw of a training dataset or the failure to include one observation can alter the outcome for some people, then other choices, such as feature selection or the type of model, are likely to have even more significant impacts on individual outcomes. Researchers refer to this as "predictive multiplicity" when competing models perform equally well but produce conflicting predictions in a substantial number of individual cases.[54] And foundational choices, like defining the target of prediction, will even more profoundly shift the way outcomes are distributed.

These observations matter for the law, because the absence of a definitive baseline model means that there is no single "correct" model against which interventions to reduce bias can be measured. Individual outcomes are not stable, but can vary depending upon small choices made in the model-building process. As a result, it is difficult to say for certain that a particular individual would have been selected absent considerations of racial equity and therefore has some settled expectation that was disrupted.

The second implication that follows from understanding the complexity of the model-building process is that there are many ways in which unfair bias can creep into a model. Consequently, there are multiple points at which a designer might make choices to try to remove or reduce racial bias. Different strategies will have different impacts on the final model and the outcomes it predicts. Thus, determining the legality of race-conscious de-biasing efforts will depend upon exactly when and how race is taken into account in the model-building process.

III.
ANTIDISCRIMINATION DOCTRINE AND RACE-CONSCIOUS DECISION-MAKING

Because strategies for building fair algorithms require explicit consideration of race, some researchers question whether they are legal under antidiscrimination law.[55] The concern is that by taking race into account, these efforts will themselves be considered a form of intentional discrimination forbidden by law. To put it more concretely, if model-builders take race into

[53] Emily Black & Matt Fredrikson, *Leave-One-Out Unfairness*, ACM 285 (2021). They find that "it occurs often enough to be a concern in some settings (i.e., up to 7% of data is affected); that it occurs even on points for which the model assigns high confidence; and is not consistently influenced by dataset size, test accuracy, or generalization error." *Id.*

[54] *See* Marx et al., *supra* note 20, at 1 (defining "predictive multiplicity" as "the ability of a prediction problem to admit competing models that assign conflicting predictions") (emphasis removed).

[55] *See, e.g.*, Bent, *supra* note 10, at 805–09 (questioning whether "algorithmic affirmative action" is legal); Sam Corbett-Davies et al., *supra* note 7, at 8–9 (raising concerns that taking race into account in a model will trigger strict scrutiny); Cofone, *supra* note 7, at 1427–31 (suggesting that it is against the law to include information about race in a model); Ho & Xiang, *supra* note 11, at 134 (arguing that algorithmic fairness strategies pose serious legal risks of violating equal protection).

account to prevent an algorithm from being biased against Black people, have they engaged in discrimination against white people? Contrary to what some have assumed, race consciousness in the model-building process does not automatically render an algorithm unlawful. Rather, its permissibility depends upon when and how race is taken into account.

Before considering which de-biasing strategies are lawful, this Section first explains existing antidiscrimination doctrine. The Supreme Court's race jurisprudence has been subject to extensive criticism, particularly by critical race scholars who argue that the Court's colorblind approach overlooks historical and systemic disadvantages imposed on the basis of race, and thereby enables and reinforces racial subordination.[56] My purpose here is not to add to this body of criticism nor to defend the Court's decisions. Instead, this Section analyzes existing law, taking established doctrine at face value and the Justices at their word when they explain their reasoning. Under current doctrine, ample room exists for certain types of race-conscious efforts to promote fairness.

The legal prohibition on race discrimination has many sources. Different statutes prohibit discrimination when lending money,[57] hiring workers,[58] selling or renting a home,[59] entering a contract,[60] or providing educational opportunities.[61] The Constitution also forbids race discrimination, but it only applies to state actors. Exploring the nuances of each potentially relevant law is not possible here. Instead, Part III.A analyzes one antidiscrimination statute, Title VII of the Civil Rights Act of 1964, which prohibits discrimination in employment and has particularly well-developed case law. Part III.B then examines the prohibition on race discrimination under the Equal Protection Clause of the Constitution.[62]

[56] For a small sampling of this extensive literature, see, for example, Devon W. Carbado, *Footnote 43: Recovering Justice Powell's Anti-Preference Framing of Affirmative Action*, 53 U. CAL. DAVIS L. REV. 1117 (2019); Kimberlé Williams Crenshaw, *Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law*, 101 HARV. L. REV. 1331 (1988); Neil Gotanda, *A Critique of "Our Constitution Is Color-Blind"*, 44 STAN. L. REV. 1 (1991); Ian Haney-López, *Intentional Blindness*, 87 N.Y.U. L. REV. 1779 (2012); Cheryl I. Harris, *Equal Treatment and the Reproduction of Inequality*, 69 FORDHAM L. REV. 1753 (2001). Others, such as Aziz Huq, contend that the Court's existing race jurisprudence is particularly unsuited to problems of discrimination in algorithmic decision-making. Huq, *supra* note 8, at 1101 (arguing that current equal protection doctrine is a poor fit because it poses questions not relevant to algorithmic decision-making).

[57] Equal Credit Opportunity Act, 15 U.S.C. § 1691(a)(1).

[58] Title VII of the Civil Rights Act of 1964, Pub. L. No. 88-372, 78 Stat. 255 (codified at 42 U.S.C. § 2000e et seq. (1964)).

[59] Title VIII of the Civil Rights Act of 1968 (Fair Housing Act), 42 U.S.C. §§ 3601–19.

[60] 42 U.S.C. § 1981.

[61] Title VI of the Civil Rights Act of 1964, Pub. L. No. 88-372, 78 Stat. 255 (codified at 42 U.S.C. § 2000d et seq. (1964)).

[62] U.S. CONST. amend. XIV, § 1.

A. Statutory Law

1. The Title VII Framework

Title VII prohibits discrimination in employment on the basis of race, sex, and other protected characteristics.[63] Employment discrimination cases generally fall into two types: disparate treatment or disparate impact.

The typical disparate treatment case involves intentional discrimination, requiring plaintiffs to show that they suffered less favorable treatment motivated by a protected characteristic. To prevail on a disparate treatment claim, plaintiffs must show that they suffered an adverse action taken "*because of*" their race or other protected characteristic.[64] Establishing causation is critical to proving disparate treatment, and there are two routes for doing so: (1) showing that the protected characteristics was a "motivating factor" for the adverse decision, or (2) demonstrating that it was a "but-for cause."[65] Pursuant to the first route, if the plaintiff shows that race was a motivating factor, the employer is liable, although it may avoid certain remedies by establishing an affirmative defense.[66] The second route, showing "but-for" causation, requires a plaintiff to demonstrate that the protected characteristic actually made a difference in the outcome.[67] It asks whether an adverse outcome for a worker would have come out differently if the protected characteristic had not been taken into account.

[63] 42 U.S.C. § 2000e-2 (1964) (prohibiting employment discrimination based on race, color, religion, sex, and national origin). Other federal statutes create additional protected characteristics. *See* Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101–12213 (disability); Age Discrimination in Employment Act of 1967, 29 U.S.C. §§ 621–634 (age); Genetic Information Nondiscrimination Act of 2008, 42 U.S.C. §§ 2000ff–2000ff-11 (genetic traits).

[64] 42 U.S.C. § 2000e-2(a)(1). Even though disparate treatment is often described as involving intentional discrimination, the prohibition against discrimination "because of" a protected characteristic can extend beyond cases involving invidious intent. *See, e.g.*, Katie Eyer, *The But-For Theory of Anti-Discrimination Law*, 107 VA. L. REV. 1624–25 (2021) (arguing that disparate treatment in the sense of differential treatment extends beyond intentional discrimination); Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357, 1357 (2009) (noting that Title VII doctrine deviates from requiring discriminatory intent in many situations that do not fall under disparate impact either).

[65] Bostock v. Clayton Cnty., 140 S. Ct. 1731, 1739–40 (2020).

[66] The "motivating factor" standard applies in so-called "mixed-motive" situations, where there is evidence that a mix of legitimate and illegitimate factors motivated an adverse decision. The employer is liable if the protected characteristic motivated the firing, although it can avoid paying damages and certain forms of injunctive relief if it demonstrates that it would have made the same decision absent consideration of the protected characteristic. 42 U.S.C. §§ 2000e-2(m), 2000e-5(g)(2)(B). The motivating factor standard is not available for retaliation claims. *See* Univ. Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 360 (2013). The same is true for age discrimination claims, which must be proven under the but-for causation standard. *See*, Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 174 (2009).

[67] *Bostock*, 140 S. Ct. at 1740. The "but-for" causation standard is imported from tort law, an interpretation of Title VII that has been subject to criticism. *See, e.g.*, Sandra F. Sperino, *Let's Pretend Discrimination Is a Tort*, 75 OHIO ST. L.J. 1107, 1112 (2014) (criticizing the use of but-for causation in discrimination cases). Although it is generally considered more demanding than the motivating factor test, some scholars have argued otherwise. *See, e.g.*, Eyer, *supra* note 64, at 1 (arguing that an expansive understanding of but-for causation is "potentially radical in its legal effects").

Unlike disparate treatment claims, disparate impact cases do not require proof of intent. The analysis instead focuses on the discriminatory effects of facially neutral practices.[68] Plaintiffs proceeding under a disparate impact theory establish a prima facie case by showing that an employment practice has a significant adverse effect on certain groups, for example, by screening out disproportionately more Black than white applicants for a particular job.[69] Employment practices that disparately impact disadvantaged racial groups are unlawful unless the employer can show that they are "job related . . . and consistent with business necessity."[70]

Disparate impact theory is relevant to predictive algorithms because these tools may disproportionately screen out racial minorities from employment opportunities, even if the employer did not intend to discriminate when adopting the tool. Although scholars debate how effective disparate impact theory will be in addressing algorithmic bias,[71] the risk of liability incentivizes employers to take steps to reduce biased outcomes. If doing so involves taking race into account, they may worry that they risk running afoul of disparate treatment law.

2. *The Supreme Court's Affirmative Action Cases*

When Title VII became effective in 1965, many employers had racially segregated workforces. Some firms had openly engaged in segregation or racial exclusion. In other cases, discriminatory intent was difficult to prove, but stark racial disparities left employers vulnerable to legal challenges under the disparate impact theory. Given the significant risks of legal liability, employers had strong incentives to scrutinize their own practices for discrimination and to voluntarily correct them. The lingering effects of past racial segregation, however, proved difficult to eradicate, in part due to low hiring and turnover

[68] The theory was first recognized in *Griggs v. Duke Power Co.*, 401 U.S. 424, 430–31 (1971), and was later codified as part of the Civil Rights Act of 1991.

[69] A prima facie case of disparate impact is typically established by showing that the selection rate for one group (e.g., Black applicants) is significantly different from the selection rate of another group (e.g., white applicants) using standard tests of statistical significance, such as two standard deviations. *See, e.g.*, Castenada v. Partida, 430 U.S. 482, 496 n.17 (1977) ("[I]f the difference between the expected value and the observed number is greater than two or three standard deviations, then the hypothesis that [observed number] was random would be suspect . . . ."); Jones v. City of Boston, 752 F.3d 38, 51 (1st Cir. 2014) (accepting a p-value of five percent as the threshold for statistical significance); Stagi v. Nat'l R.R. Passenger Corp., 391 Fed. Appx. 133, 138 (3d Cir. 2010) (suggesting that a finding of statistical significance with probability at or below 0.05 will typically be sufficient). Some courts and commentators also refer to the "four-fifths rule," which asks whether the selection rate for a disadvantaged group is less than 4/5 the selection rate of the most advantaged group. The "four-fifths rule," however, is not a legal rule, but a "rule of thumb" articulated by federal agencies to guide their priorities when enforcing antidiscrimination law. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 995 n.3 (1988) (noting that the 4/5 rule is not more than a rule of thumb for courts and has been criticized on technical grounds); *Jones*, 752 F.3d at 51 (noting that the 4/5 rule is "a helpful benchmark in certain circumstances" but generally not decisive); *Stagi,* 391 Fed. Appx. at 138 (noting the 4/5 rule "has come under substantial criticism, and has not been particularly persuasive").

[70] 42 U.S.C. § 2000e-2(k).

[71] *See supra* note 33 and sources cited therein.

rates. As a result, some employers undertook more active efforts to integrate their workforces—sometimes voluntarily, and sometimes under legal compulsion.

Employers' efforts to desegregate the workplace took many forms, but the most visible were challenged legally. White workers sued employers that adopted affirmative action plans, claiming that any preference given to Black workers was itself a form of racial discrimination forbidden by Title VII. Affirmative action plans that were implemented following a judicial finding of past intentional discrimination were generally upheld.[72] More difficult questions arose when an employer voluntarily adopted an affirmative action plan prior to any litigation.

The leading case addressing the lawfulness of voluntary affirmative action plans under Title VII is *United Steelworkers v. Weber*.[73] In light of a history of near-total exclusion of Black workers from craftwork positions,[74] the employer, Kaiser Aluminum & Chemical Corp., and the steelworkers union created a program to train its unskilled workers for skilled positions. Applicants were accepted into the program based on seniority, with the caveat that at least 50% of the positions had to be filled by Black workers until the proportion of Black skilled craftworkers at the plant (then 1.83%) roughly matched the percentage of Black workers in the local labor force (39%).[75] Brian Weber, a white worker who had more seniority than some of the Black workers accepted into the program, was not admitted and sued, alleging that the plan discriminated against him.

Because Weber was not admitted because of his race, he argued that he had been subjected to disparate treatment. The employer sought to defend the training program as a valid affirmative action plan. Analyzing the text, purpose, and historical context of Title VII, the Supreme Court in *Weber* concluded that the statute does not prohibit all voluntary race-conscious affirmative action. The goal of the Civil Rights Act, it noted, was "the integration of Blacks into the mainstream of American society,"[76] which required opening employment opportunities to them on an equal basis. The Court emphasized the importance of *voluntary* employer efforts to solve problems of racial discrimination. As it explained, Title VII was intended "as a spur or catalyst to cause 'employers and unions to self-examine and to self-evaluate their employment practices and to endeavor to eliminate' so far as possible, the last vestiges of" the country's history of racial segregation.[77]

Although some uses of race to promote equality might violate antidiscrimination law, the Court concluded that Kaiser's affirmative action plan

[72] *See, e.g.*, Loc. No. 93, Int'l Ass'n Firefighters v. City of Cleveland, 478 U.S. 501, 515 (1986); United States v. Paradise, 480 U.S. 149, 166 (1987).

[73] United Steelworkers v. Weber, 443 U.S. 193 (1979).

[74] Kaiser only hired persons with prior craft experience. Black workers were excluded from craft unions; thus, they were unable to present the proper credentials. *Id.* at 198.

[75] *Id.* at 198–99.

[76] *Id.* at 202.

[77] *Id.* at 204.

"f[ell] on the permissible side of the line,"[78] pointing to several relevant considerations. First, its purpose mirrored that of Title VII, "to break down old patterns of racial segregation and hierarchy."[79] In addition, the plan "d[id] not unnecessarily trammel the interests of white employees."[80] It did not disrupt settled expectations by, for example, requiring the discharge of white workers, nor did it create an "absolute bar" to their advancement.[81] Finally, the plan was temporary, and not intended to maintain a permanent racial balance in the workforce.[82]

The Supreme Court has only revisited the lawfulness of affirmative action programs under Title VII once more, this time in the context of sex. In *Johnson v. Transportation Agency*,[83] Paul Johnson sued when a promotion he sought was given to a female applicant, Diane Joyce. He alleged sex discrimination because the Agency had an affirmative action plan that took into account the sex of a qualified applicant when filling positions in which women were significantly underrepresented.[84]Applying the framework it had established in *Weber*, the Court found that the plan furthered the purposes of Title VII, did not disrupt legitimate, settled expectations of male employees, and was not intended to maintain a permanent racial or sexual balance.[85] It thus concluded that the affirmative action plan was permissible and rejected Johnson's claim of discrimination.

Weber and *Johnson* provide a legal framework for assessing voluntary affirmative action plans[86]; however, not everything an employer does that might

[78] *Id.* at 208.

[79] *Id.*

[80] *Id.*

[81] *Id.*

[82] *Id.*

[83] Johnson v. Transp. Agency, 480 U.S. 616, 616 (1987).

[84] *Id.* at 621. Unlike in *Weber*, the plan did not involve rigid numerical quotas, but the Court assumed, following the findings of the district court, that sex was "the determining factor" in the decision to promote Joyce. *Id.* at 616. One might question this conclusion given the facts. The interview panel rated Johnson a 75 and Joyce a 73, and it is doubtful that the difference was meaningful. In any case, the employer was not required to promote the person with the highest score, and the affirmative action plan did not require any particular number of female hires. *Id.* at 655. In deciding whom to promote, the Director considered numerous factors, including the severe underrepresentation of women in the relevant job category. *Id.* at 625.

[85] *Id.* at 637–40. The Court concluded that the Agency's affirmative action plan was justified under *Weber*, because it was intended to eliminate the egregious under-representation of women in skilled job positions. *Id.* at 636. "*None* of the 238 positions [were] occupied by a woman." *Id.* The Court also concluded that Johnson had "no legitimate, firmly rooted expectation" to the position that was disrupted by the affirmative action plan. *Id.* at 638. Further, the plan did not set aside any positions solely for women, or impose any fixed hiring quotas, but instead took a flexible, case-by-case approach. *Id.* at 639.

[86] After the Civil Rights Act of 1991 was passed, some advocates argued that the new subsection 703(m), which makes any adverse employment decision "motivated by" a protected characteristic unlawful, rendered all employer affirmative action plans unlawful. The Act, however, also made clear that its provisions did not affect the lawfulness of valid affirmative action plans. Civil Rights

be labeled "affirmative action" triggers this analysis. The term is not well-defined and has been applied to a broad range of activities that are aimed at redressing racial inequality, but which can be quite different in operation and effect. The *Weber* Court emphasized that its decision addressed only plans "that accord racial preferences in the manner and for the purpose" of Kaiser's particular plan.[87] As discussed in the next section, other employer plans or practices are not required to meet the *Weber*/*Johnson* requirements even if they might fall within an expansive notion of "affirmative action."

3. *Anti-Bias and Diversity Efforts*

The *Weber*/*Johnson* framework applies when an employer has engaged in disparate treatment and seeks to justify its actions on the grounds that they were taken pursuant to a valid affirmative action plan.

Courts, however, do not always find that an affirmative action or diversity plan causes disparate treatment. White or male workers sometimes allege discrimination when they lose out on an employment opportunity by pointing to an employer's affirmative action plan as evidence of a discriminatory motive. The mere existence of such a policy, however, is insufficient to prove that the employer engaged in disparate treatment. Rather, the plaintiff must establish a causal link between the policy and an adverse action.[88] If the plan requires rigid numerical goals and was applied to the hiring decision at issue, then the existence of the plan may raise an inference of discrimination.[89] In numerous other cases, however, the mere fact that an employer has an affirmative action plan, or has stated an interest in diversifying its workforce, does not by itself provide evidence of discriminatory intent.[90] In the absence of a clear connection to the

Act of 1991, Pub. L. No. 102-166, § 116, 105 Stat. 1071 (1991) (codified at 42 U.S.C. § 1981). In any case, the "motivating factor" provision does not appear to have affected how courts treat affirmative action or diversity plans.

[87] United Steelworkers v. Weber, 443 U.S. 193, 200 (1979).

[88] Rudin v. Lincoln Land Cmty. Coll., 420 F.3d 712, 722 (7th Cir. 2005) (citing Whalen v. Rubin, 91 F.3d 1041, 1045 (7th Cir. 1996)).

[89] *See* Frank v. Xerox Corp., 347 F.3d 130, 137 (5th Cir. 2003); Bass v. Bd. Cnty. Comm'rs, 256 F.3d 1095, 1107 (11th Cir. 2001).

[90] *See* Coppinger v. Wal-Mart Stores, Inc., No. 3:07cv458/MCR/MD, 2009 U.S. Dist. LEXIS 91120 at *26 (N.D. Fla. Sept. 30, 2009), Jones v. Bernanke, 493 F. Supp. 2d 18, 29 (D.D.C. 2007) ("[A]n employer's statement that it is committed to diversity 'if expressed in terms of creating opportunities for employees of different races and both genders . . . is not proof of discriminatory motive . . . .'"); Keating v. Paulson, No. 96 C 3817, 2007 WL 3231437 *8 (N.D. Ill. Oct. 25, 2007); Martin v. City of Atlanta, 579 Fed. Appx. 819, 824-25 (11th Cir. 2014); Plumb v. Potter, 212 Fed. Appx. 472, 472 (6th Cir. 2007) (holding that a USPS supervisor's statement that a specific facility needed more diversity was not evidence of sex discrimination). These types of employer plans are sometimes insufficient to meet even the minimal requirements of establishing a prima facie case under the *McDonnell Douglas* framework. Stacy Hawkins, *What the Supreme Court's Diversity Doctrine Means for Workplace Diversity Efforts*, 33 A.B.A. J. LAB. & EMP. L. 139, 155 (2018) ("In cases where plaintiffs point only to employer commitments to workplace diversity generally, without offering discrete evidence that race or ethnicity was considered in making the challenged employment decision, courts have found this insufficient to satisfy even the minimal burden of establishing a prima facie case of discrimination . . . .").

specific decision rejecting the plaintiff, an employer's affirmative action plan requires no special justification, nor is it examined for validity under the *Weber*/*Johnson* framework. Courts simply conclude that the lack of a causal connection to the adverse outcome means that no disparate treatment has occurred.[91]

Thus, employers are permitted to engage in some types of race-conscious efforts to diversify their workplaces without having to justify them under the *Weber* framework. In *Duffy v. Wolle*,[92] the Eighth Circuit found that

> An employer's affirmative efforts to recruit minority and female applicants does not constitute discrimination. An inclusive recruitment effort enables employers to generate the largest pool of qualified applicants and helps to ensure that minorities and women are not discriminatorily excluded from employment.[93]

The plaintiff in that case complained that a woman was hired for a position he sought after the employer chose to advertise the position nationally in order to have an "open, nationwide, diverse pool of qualified applicants."[94] Even though this effort likely reduced the plaintiff's chances of receiving the promotion by expanding the pool, there was no evidence that the promotion decision itself was based on anything other than the applicants' qualifications. The court noted that "[t]he only harm to white males is that they must compete against a larger pool of qualified applicants," but that increased competition "does not state a cognizable harm."[95]

Duffy and the cases that follow it[96] indicate that race-conscious actions taken to remove unfair policies or diversify the workforce do not constitute disparate treatment against white workers. When an employer expands recruitment efforts to create a broader applicant pool, but does not make actual hiring decisions based on race, it has not engaged in discrimination. More generally, employers may adopt changes to make their processes fairer and more inclusive, so long as they do not make individual employment decisions because of race. The changes may alter a white applicant's chances of success, but that

[91] *See, e.g.*, Mlynczak v. Bodman, 442 F.3d 1050, 1058 (7th Cir. 2006) ("The simple fact that such a policy exists does not prove that intentional discrimination is the reason why a particular individual was not hired or promoted.").

[92] Duffy v. Wolle, 123 F.3d 1026, 1038–39 (8th Cir. 1997) (internal citations omitted).

[93] *Id.*

[94] *Id.* at 1030.

[95] *Id.* at 1039. In *Rogers v. Haley*, 421 F. Supp. 2d 1361 (M.D. Ala. 2006), the court reached a similar result in a case brought under the Constitution. The plaintiff, a white correctional officer employed by the state, complained that his employer's efforts to widely advertise job openings harmed him because it resulted in an "influx of [B]lacks" competing with him for the position he sought. *Id.* at 1365. The court rejected his claim, because there was no evidence that the expanded recruitment program excluded or restricted white applicants, or that the plaintiff had been denied a promotion because of his race. *Id.* at 1367–68.

[96] *See* Rudin v. Lincoln Land Cmty. Coll., 420 F.3d 712, 722 (7th Cir. 2005); Mlynczak v. Bodman, 442 F.3d 1050, 1050 (7th Cir. 2006).

fact alone does not create a cognizable harm. The change in procedures may have been motivated by racial equity considerations; however, if the decision in the plaintiff's case was not made because of race, then the requisite causal connection is missing. No disparate treatment has occurred, and the *Weber* requirements never come into play.

This conclusion accords with the Supreme Court's repeated emphasis on the importance of voluntary employer efforts to remove discriminatory practices. If employers were subjected to potential suit and stringent requirements whenever they sought to address racially inequitable practices, they would be discouraged from meeting their obligation to remove "artificial, arbitrary, and unnecessary barriers" to the employment of racial minorities.[97]

This means that employers are free to make *prospective* changes to practices they discover are biased or discriminatory and to take race into account when doing so. Because future applicants have no fixed entitlement to an employer's past hiring or promotion criteria, changes to these practices do not disrupt legitimate, settled expectations. The fact that changes are motivated by a desire to make the process less racially biased does not make them a form of disparate treatment.

Some commenters have suggested to the contrary, believing that the Supreme Court's decision in *Ricci v. DeStefano*[98] limits employers from prospectively changing their practices to remove disparate impact or to promote diversity goals.[99] As I have argued elsewhere, this conclusion rests on a misreading of the case.[100]

In *Ricci*, the City of New Haven discarded a promotional examination for firefighters because it would have produced a nearly all-white promotional class, and the City feared a disparate impact suit by minority firefighters. A majority of the Supreme Court found that the City's decision to discard the results constituted disparate treatment[101] and would only be permissible if there was "a strong basis in evidence" that the test violated disparate impact law, a showing the City could not make.[102]

The Court's application of the "strong basis in evidence" test was premised on its finding that the City had engaged in disparate treatment against the successful test takers. The injury, according to the Court, arose from "the high,

[97] Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971) ("What is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification.").

[98] Ricci v. DeStefano, 557 U.S. 557, 557 (2009).

[99] *See e.g.*, Barocas & Selbst, *supra* note 1; Kroll et al., *supra* note 4.

[100] *See* Kim, *supra* note 1; Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 191 (2017).

[101] *Ricci*, 557 U.S. at 563. The dissent disagreed that the City's actions constituted disparate treatment, arguing that the plaintiffs had no vested right to promotion and substantial evidence existed that the test was seriously flawed and so the results should not be relied on. *Id.* at 608–09, 619 (Ginsburg, J., dissenting).

[102] *Id.* at 563 (majority opinion).

and justified, expectations of the candidates who had participated in the testing process," some of them investing considerable time and expense to do so. Thus, the case is best understood as protecting the interests of specific individual firefighters who had relied on the City's announced plan to make promotion decisions based on the exam.[103] When an employer adapts a *new* test or procedure going forward, it does not disrupt settled expectations and does not constitute disparate treatment, even if it is motivated by racial equity concerns.[104]

The Court has recognized that an employer may need to take race into account to create fairer processes. In *Ricci*, it noted that an employer is permitted to "consider[], 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."[105] The Court also appeared to view favorably race-conscious strategies used by the City to avoid bias, namely, oversampling minority firefighters when designing the written test and ensuring that each of the panels assessing candidates on the oral part of the exam contained minority members.[106] Thus, while the Court found the City's actions under the circumstances in *Ricci* to be disparate treatment, its decision does not prohibit an employer from considering race when trying to design fair procedures.

* * *

In sum, Title VII doctrine does not categorically prohibit employers from taking race into account when seeking to design fair personnel policies. The Court has repeatedly stated that the best way to achieve the purposes of equal employment opportunity that animate Title VII is to encourage employers to examine their own practices and to voluntarily remove arbitrary barriers to equal opportunity regardless of race. In order to do so effectively, employers will often have to pay attention to race and the ways in which traditional practices and procedures may systematically disadvantage racially subordinated groups. White plaintiffs who challenge these employer efforts must show that they suffered adverse actions *causally related* to the consideration of race. When the employer imposes a racial quota, as in the *Weber* case, the policy constitutes disparate treatment, but the employer may defend it as a valid form of affirmative action. If, however, an employer merely takes account of race in order to design fairer procedures, no disparate treatment has occurred and therefore, no special justification is required. The employer's consideration of race is too remote in

[103] Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1373 (2010).

[104] For example, in *Maraschiello v. City of Buffalo Police Department*, 709 F.3d 87 (2d Cir. 2013), a white firefighter, relying on *Ricci*, alleged that he was discriminated against because he was passed over for promotion after the fire department chose to revise its promotional exam. The Second Circuit found that even if the City's decision to adopt a new test was "motivated in part by its desire to achieve more racially balanced results," the plaintiff could not demonstrate that the changes were the type of "race-based adverse action" at issue in *Ricci*. *Id.* at 95–96. *See also* Carroll v. Mount Vernon, 707 F. Supp. 2d 449, 455 (S.D.N.Y. 2010).

[105] *Ricci*, 557 U.S. at 585.

[106] *Id.* at 565, 593.

time and effect to be causally connected to a specific personnel decision down the road.

B. The Equal Protection Clause

The Equal Protection Clause of the Constitution also forbids discrimination, although it differs from statutory prohibitions in a number of ways. The antidiscrimination statutes target particular types of decisions—in employment, housing, education, etc.—but they generally reach both public and private actors. By contrast, the Constitution restricts only government actors, but applies to a broad range of activities. And unlike Title VII, equal protection doctrine does not permit disparate impact claims.

Despite these differences, the basic frameworks for analyzing race-conscious actions are roughly analogous under the statutory and constitutional frameworks. As discussed above, the initial inquiry under Title VII is whether consideration of race caused disparate treatment; if so, it is unlawful unless justified as a valid affirmative action plan under *Weber*. Under the Equal Protection Clause, the analogous first step is showing that a government action relied on a racial classification. Doing so is prohibited unless the requirements of strict scrutiny are met.

1. The Court's Affirmative Action Jurisprudence

In the mid-twentieth century, spurred by the civil rights movement and growing attention to significant racial gaps in opportunities and measures of well-being, government actors took steps to redress racial inequities in areas like education and public contracting. White plaintiffs who alleged that they were harmed because the government used racial classifications to make decisions challenged these efforts, which became characterized as "affirmative action." Over a series of cases, the Supreme Court settled on several principles relevant to these challenges.

First, the level of scrutiny applied to race-based classifications "is not dependent on the race of those burdened or benefited by a particular classification."[107] In the Court's view, it does not matter if the classification is intended to achieve a benign purpose,[108] such as compensating for existing disadvantages based on race. Remedying "societal discrimination" is not a sufficient justification,[109] although the Court has approved race-based remedies for a government actor's own past discrimination.[110] Second, the appropriate

[107] City of Richmond v. J.A. Croson Co., 488 U.S. 469, 494 (1989).

[108] Fisher v. Univ. Tex. Austin, 570 U.S. 297, 307 (2013).

[109] Wygant v. Jackson Bd. Educ., 476 U.S. 267, 274 (1986).

[110] Parents Involved Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 715 (2007) (recognizing a prior desegregation decree as valid).

level of scrutiny for examining racial classifications is "strict scrutiny."[111] Strict scrutiny, the Court has instructed, requires that the racial classification "further[s] [a] compelling governmental interest[]" and that the means chosen are "narrowly tailored" to meet that interest.[112] A racial classification that does not meet that exacting standard is unconstitutional.

While the Supreme Court's affirmative action cases impose a high barrier to the use of race by government in its efforts to redress racial inequality, those decisions should not be over-read. Commentators sometimes characterize the jurisprudence as mandating colorblindness, but as numerous scholars have pointed out, that reading is overly simplistic because the prohibition on race-conscious decision-making "is not categorical."[113]

One obvious exception is that strict scrutiny is not inevitably fatal. In *Grutter v. Bollinger*,[114] the Supreme Court approved the University of Michigan Law School's admissions policies which relied upon race as one factor in a holistic review of an applicant's profile. The Court held that the goal of obtaining a diverse student body was "a compelling state interest that can justify the use of race" as a factor, and that the law school's policies were narrowly tailored to meet that compelling interest.[115] The Court in *Fisher v. University of Texas at Austin*[116] similarly approved that University's admissions policies, which took race into consideration as one factor among many in selecting its student body.[117]

While *Grutter* and *Fisher* show that strict scrutiny is not always fatal, a more fundamental but often overlooked point is that not every consideration of race by a government actor triggers strict scrutiny. By extracting certain broad statements from the Court's affirmative action opinions, some commentators have concluded that race consciousness always raises constitutional concerns.

[111] Adarand Constructors v. Pena, 515 U.S. 200, 227 (1995) ("[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.").

[112] *Id.* at 220, 227.

[113] *See* Hellman, *supra* note 13, at 819. *See also* Bagenstos, *supra* note 13; Driver, *supra* note 13. *Cf.* Kim Forde-Mazrui, *The Canary-Blind Constitution: Must Government Ignore Racial Inequality?*, 79 LAW & CONTEMP. PROBS. 53, 58 (2016) (arguing that a plausible interpretation of current equal protection doctrine permits government "to act in response to racial disparities without discriminating by race, provided that the racial motivation is limited to investigating the causes of the disparities").

[114] Grutter v. Bollinger, 539 U.S. 306, 306 (2003).

[115] *Id.* at 325. Key to the Court's conclusion was the fact that the law school did not impose a numerical quota that automatically insulated members of minority groups from comparison with other applicants. Instead, race was treated merely as a "plus" factor in the context of a "highly individualized, holistic review," and not as "the defining feature" of an applicant's file. *Id.* at 336–67.

[116] Fisher v. Univ. Tex. Austin, 570 U.S. 297, 297 (2016).

[117] The University admitted a large proportion of its student body under the Top Ten Percent plan, which guaranteed admission to students graduating from Texas high schools in the top ten percent of their class. *Id.* at 305. For the remaining seats, the University considered an Academic Index and a Personal Achievement Index (PAI). *Id.* at 304. The PAI took a number of factors into account, including not only race, but also leadership, experience, activities, background factors like language, etc. *Id.* Race was thus "a factor of a factor of a factor." *Id.* at 336.

However, the Supreme Court has repeatedly emphasized that it decides concrete cases, not abstract propositions of law. Close attention to the specific factual contexts in which these cases were decided suggests that it is particular uses of race, not mere race consciousness, that triggers strict scrutiny.

In the first major affirmative action case, *Regents of the University of California v. Bakke*,[118] the Court considered a constitutional challenge to a state university's admissions policy which set aside sixteen out of one hundred places in a medical school class for members of disadvantaged minority groups. In *Croson*,[119] the Court evaluated a city's minority set-aside plan that required prime contractors to award a fixed percentage of their subcontracts to entities owned and controlled by minority group members. *Adarand*[120] involved a challenge to a similar plan at the federal level that presumptively advantaged minority-owned businesses by providing them with a fixed financial boost. *Wygant*[121] challenged a school board policy that the percentage of minority teachers laid off could not exceed the percentage employed by the district. Because minority teachers generally had less seniority, the school district laid off white teachers with greater seniority pursuant to the policy. And *Parents Involved*[122] considered school district policies that made school assignments by race to ensure that the racial balance at individual schools fell within a specified range.

These cases, through which the Court developed its affirmative action doctrine, involved government decision-makers using race in a particular way. More specifically, the challenged government decisions all involved applying racial classifications to individuals in a rigidly mechanical way and doing so in order to systematically favor one racial group over another.

2. *Permissible Race Consciousness*

In a variety of situations outside of the affirmative action cases, the government acts in race-aware ways without triggering strict scrutiny.[123] Some practices are so familiar and so widely accepted that they go almost unnoticed. For example, every ten years, the federal government conducts the Census, collecting detailed information, including race, about the U.S. population. In addition to the Census, governments at all levels—local, state, and federal—routinely collect and analyze racial data. This information is essential to understanding where and to what extent racial disparities exist in matters like health care, education, and employment, and to assessing the impact and effectiveness of government policies.

[118] Regents Univ. Cal. v. Bakke, 438 U.S. 265, 266 (1978).

[119] *See* City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493 (1989).

[120] *See* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 204 (1995).

[121] *See* Wygant v. Jackson Bd. Educ., 476 U.S. 267, 270–71 (1986).

[122] *See* Parents Involved Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 709–10 (2007).

[123] *See, e.g.*, Primus, *supra* note 13, at 505.

These practices rarely provoke legal questions, let alone successful constitutional challenges.[124] In one case, plaintiffs sued to bar the collection of racial information in the Census, arguing that the questionnaire involved a racial classification and was subject to strict scrutiny.[125] The district court rejected the claim, noting that there is a "distinction between collecting demographic data so that the government may have the information it believes . . . it needs in order to govern, and governmental use of suspect classifications without a compelling interest."[126] Because the Census involved only the collection of information, it did not even trigger heightened scrutiny. As the Court explained, the concerns plaintiffs raised about the type of information sought on the Census form was "one properly addressed by Congress, not by the courts."[127]

Information about race is highly relevant to addressing public health concerns. Many states have enacted legislation that specifically requires the analysis of racial disparities in health outcomes and sets goals for the reduction of those disparities.[128] Most recently, efforts to address the pandemic have included consideration of the racial disparities in the risks posed by COVID and the obstacles to achieving adequate vaccination levels in communities of color. Evidence of these racial disparities has informed decisions relating to outreach and educational efforts, as well as the location of vaccine clinics. So long as they do not use racial classifications to distribute or withhold benefits to individuals, these activities should not raise constitutional concerns.

Government actors also routinely act with an awareness of race when law enforcement uses suspect profiles.[129] When witnesses to a crime describe a perpetrator, police focus their investigative attention on individuals who match the characteristics provided, including race. Only on occasion are these practices legally challenged, and so far, courts do not appear to agree that they raise constitutional concerns.[130] If the Equal Protection Clause embodied a strict colorblindness theory, race-based subject descriptions should arguably trigger strict scrutiny,[131] but apparently they do not.[132]

There are other examples of race-aware government activity that do not appear to trigger constitutional concerns. For example, when placing children for

[124] *Cf.* Dept. Com. v. New York, 139 S. Ct. 2551, 2561 (2019) (recognizing that demographic questions, including questions about race, have long been included in the Census to inform government policies).

[125] Morales v. Daley, 116 F. Supp. 2d 801, 801–02 (S.D. Tex. 2000).

[126] *Id.* at 814.

[127] *Id.* at 815.

[128] Govind Persad, *Allocating Medicine Fairly in an Unfair Pandemic*, 2021 U. ILL. L. REV. 1085, 1129 n.264 (2021).

[129] *See, e.g.*, Hellman, *supra* note 13, at 859.

[130] *See, e.g.*, Brown v. City of Oneonta, 221 F.3d 329, 333 (2d Cir. 2000); Monroe v. City of Charlottesville, 579 F.3d 380, 382 (4th Cir. 2009).

[131] R. Richard Banks, *Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse*, 48 UCLA L. REV. 1075, 1077–78 (2001).

[132] *See* Huq, *supra* note 8, at 1096.

adoption, agencies sometimes take the preferences of adoptive or biological parents, including racial preferences, into account. And while strict racial matching would likely trigger constitutional concerns, considerations related to racial identity may inform assessments of the child's best interests when making placement decisions.[133]

The Supreme Court's voting rights jurisprudence also makes a distinction between racial classifications and race consciousness. If race is the *predominant* factor motivating a state's redistricting decisions, its decisions are subject to strict scrutiny.[134] On the other hand, if the state pursues other goals, the fact that it relied on race-based information, such as the knowledge that the most loyal Democratic voters are Black voters, does not trigger equal protection concerns.[135] Once again, it appears that government action that is premised on information about racial disparities does not *per se* trigger strict scrutiny. What matters is *how* race is used in the decision-making process.

3. Race Consciousness without Racial Classifications

The above examples fall outside the Court's affirmative action jurisprudence, illustrating that not all race-conscious decision-making is constitutionally suspect. Scholars have characterized the line between permissible race consciousness and uses of race that trigger strict scrutiny in different ways. Justin Driver draws a conceptual distinction between principles of anti-classification and colorblindness,[136] arguing that the anti-classification principle forbids government "from racially categorizing *individuals*," while colorblindness would preclude "taking account of racial considerations . . . within *society* as a whole."[137] This distinction is important, in his view, because it allows courts to take racial realities into account when relevant, as when deciding criminal procedure cases, but without resorting to racially classifying individuals.

Deborah Hellman suggests two principles for identifying permissible race-conscious activities.[138] First, she argues for distinguishing between collection and use of racial information. The former "does not constitute disparate treatment and thus does not give rise to strict scrutiny" because it does not

[133] *See* R. Richard Banks, *The Color of Desire: Fulfilling Adoptive Parents' Racial Preferences Through Discriminatory State Action*, 107 YALE L.J. 875, 879–80 (1998).

[134] *See* Miller v. Johnson, 515 U.S. 900, 916 (1995).

[135] *See* Hunt v. Cromartie, 526 U.S. 541, 551–52 (1999).

[136] *See* Driver, *supra* note 13, at 451.

[137] *Id.*

[138] Hellman argues that there are racial classifications that do not trigger strict scrutiny, citing the examples of racial data collected as part of the Census and the inclusion of racial characteristics in criminal suspect descriptions. *See* Hellman, *supra* note 13, at 855–62. While I share her conclusion that these examples show that some uses of race are legally permissible, I would not characterize them as racial classifications. Instead, I would characterize them as examples of permissible race-conscious government action.

produce the sort of direct, real-world effects that raise constitutional concerns.[139] Although the collection of racial data may reveal disparities and thereby shape government policies to address them, these "downstream consequences" of collecting the information are too remote to trigger strict scrutiny.[140] Second, she asserts that strict scrutiny applies when government makes generalizations *about* racial groups, but not generalizations that refer to race.[141] Suspect profiles do not rely on generalizations about a racial group,[142] and thus, even though they refer to racial characteristics, they are not suspect racial classifications triggering strict scrutiny.

Samuel Bagenstos argues that the Court's equal protection cases are best understood as requiring strict scrutiny of all racial classifications, but not necessarily all forms of race consciousness. As he puts it: "[the] Court has never held that all government actions motivated by an effort to achieve racially defined ends trigger strict scrutiny. Rather, the Court has held that all racial *classifications* trigger strict scrutiny."[143] Thus, "state actions that do not classify individuals based on their race are not constitutionally suspect simply because they are motivated by the purpose of integrating the races."[144]

I agree with Bagenstos that the best way to make sense of the Court's equal protection jurisprudence and the broad array of situations in which government action uncontroversially takes account of race is to distinguish between racial classifications and race consciousness. Government practices that rely on racial classifications to make decisions about individuals are presumptively prohibited unless they satisfy strict scrutiny. By contrast, race consciousness, in the sense of taking into account racial realities to shape legitimate policy goals like reducing health disparities or promoting integration in schools, does not trigger heightened constitutional concern.

Although the Court has never clearly delineated what constitutes a racial classification, the reasoning in its affirmative action cases acknowledges the distinction between racial classifications and race consciousness. In his concurring opinion in *Parents Involved*, Justice Kennedy made this distinction explicit:

> School boards may pursue the goal of bringing together students of

[139] *Id.* at 858.

[140] *Id.* at 862.

[141] *See id.* at 858–59.

[142] Hellman explains that when the police investigate persons of a particular race that match a witness's description, they are not relying on a generalization about the members of that racial group. *See id.* at 858–60. Instead, their actions follow from a different type of generalization: that eye-witness descriptions are generally helpful in identifying perpetrators. *See id.* Of course, reliance on a witness description could turn into or be a cover for race-based profiling. When, for example, an investigation indiscriminately sweeps up individuals who share a suspect's race or nationality without any other indicia of connection to the crime, it arguably does constitute a race-based, and therefore suspect, generalization. *See* Shirin Sinnar, *The Lost Story of* Iqbal, 105 GEO. L.J. 379, 419–21 (2017).

[143] Bagenstos, *supra* note 13, at 1119.

[144] *Id.* at 1117.

> diverse backgrounds and races . . . [by] 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. *These mechanisms are race-conscious but 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* . . . .[145]

Writing for the Court majority a few years later in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*,[146] Justice Kennedy reiterated this point. Although the case addressed a question of statutory interpretation, the Court's discussion of remedies spoke to the constitutional permissibility of race-conscious action.[147] The Court first held that the disparate impact theory of liability was available under the Fair Housing Act, then discussed the appropriate remedies for a violation. It concluded that courts should strive to design remedies that "eliminate racial disparities through race-neutral means."[148] It further noted that "race may be considered in certain circumstances and in a proper fashion," and that "mere awareness of race in attempting to solve [problems of racial inequity and isolation] does not doom that endeavor from the outset."[149]

Similarly, in *Fisher* the Court appeared to have no concerns with the University of Texas's "Top Ten Percent Law," which granted automatic admission to any student in the top ten percent of a high school class in Texas.[150] After the Fifth Circuit prohibited any consideration of race in admissions in *Hopwood*, the legislature adopted the plan in order to create a more racially diverse student body than would result if admissions were based solely on test scores.[151] Because many schools and neighborhoods in Texas are racially segregated, admitting the top ten percent from each high school helped to

[145] Parents Involved Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 789 (2007) (Kennedy, J., concurring in part and concurring in the judgment).

[146] Tex. Dep't Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519 (2015).

[147] *See* Bagenstos, *supra* note 13, at 1127–30.

[148] *Tex. Dep't Hous. & Cmty. Affs.*, 576 U.S. at 544–45.

[149] *Id.* at 545. If courts find disparate impact liability, the resulting remedial orders must be consistent with the Constitution. The Court wrote: "If additional measures are adopted, courts should strive to design them to eliminate racial disparities through race-neutral means . . . . While the automatic or pervasive injection of race into public and private transactions covered by the FHA has special dangers, *it is also true that race may be considered in certain circumstances and in a proper fashion*." *Id.* at 544–45 (emphasis added). It further explained that it "does not impugn housing authorities' race-neutral efforts to encourage revitalization of communities that have long suffered the harsh consequences of segregated housing patterns. *Id.* at 545. When setting their larger goals, local housing authorities may choose to foster diversity and combat racial isolation with race-neutral tools, and mere awareness of race in attempting to solve the problems facing inner cities does not doom that endeavor at the outset." *Id.*

[150] Fisher v. Univ. Tex. Austin, 570 U.S. 297, 305 (2013).

[151] *See id.*

increase the number of Black and Hispanic students enrolled at the University.[152] As Justice Ginsburg pointed out, "race consciousness, not blindness" drove the University's Top Ten Percent plan.[153] The majority's apparent unconcern about the plan suggests that constitutional concerns are not triggered by mere race awareness where there is no reliance on racial classifications to make individual negative decisions.

In government contracting cases, the Justices have also acknowledged that states might legitimately seek to increase opportunities for disadvantaged racial groups through appropriate means.[154] As Justice Scalia wrote in his concurrence in *Croson*:

> A State can, of course, act "to undo the effects of past discrimination" in many permissible ways that do not involve classification by race. In the field of state contracting, for example, it may adopt a preference for small businesses, or even for new businesses—which would make it easier for those previously excluded by discrimination to enter the field. Such programs may well have racially disproportionate impact, but they are not based on race.[155]

What appears to trigger strict scrutiny, then, is not the government's mere consideration of race or racial disparities, but its application of racial *classifications* to individuals.

The D.C. Circuit recently confirmed this reading in a case involving a challenge to the Small Business Administration's business development program, which offers participants technical assistance and other advantages in competing for certain federal contracts.[156] The enabling statute made the program available to businesses owned by "socially disadvantaged individuals"—defined as those "who have been subjected to racial or ethnic prejudice or cultural bias because of their identity as a member of a group"—without presuming that any particular individuals could or could not show that they were eligible.[157] The D.C. Circuit wrote:

> [T]he reality that Congress enacted [the statute] with a consciousness of racial discrimination in particular as a source of the kind of disadvantages it sought to counteract does not expose the statute to strict scrutiny. . . . Policymakers may act with an awareness of race—unaccompanied by a facial racial classification or a discriminatory purpose—without thereby subjecting the resultant policies to the rigors of strict constitutional scrutiny.[158]

[152] *See id.* at 335 (Ginsburg, J., dissenting).

[153] *Id.*

[154] *See, e.g.*, City of Richmond v. J.A. Croson Co., 488 U.S. 469, 507 (1989).

[155] *Id.* at 526.

[156] Rothe Dev., Inc. v. United States Dep't Def., 836 F.3d 57, 73 (D.D.C. 2016).

[157] *Id.* at 64.

[158] *Id.* at 71–72.

Because the statute used race-neutral criteria and individuals were not automatically eligible because they belonged to particular racial groups, the D.C. Circuit concluded that the program did not involve racial classifications and was therefore subject to only rational basis review.

Distinguishing between a prohibition on racial classifications and a requirement of "colorblindness" is also necessary to make sense of the legal landscape writ large. Literal application of a colorblindness principle would throw into question enormous swaths of existing law.[159] The entire edifice of civil rights law rests on government actions that were taken with an awareness of racial inequities and the consequences of racial discrimination in our society. The Civil Rights Act of 1964, which encompasses Title VII, and the Voting Rights Act of 1965 were enacted in response to pressing concerns about racial segregation and the exclusion of Black citizens from the mainstream of American economic, social, and political life. Every state and a multitude of local governments have also recognized the harms caused by racial discrimination and passed laws making it unlawful. And every time a court considers a claim of racial discrimination under one of those laws and provides a remedy to victims of discrimination, it is acting in a race-conscious way. If the Constitution categorically forbade government from taking race into account in its decision-making, all this statutory and decisional law would be suspect. The irony, of course, is that the basis for questioning these civil rights laws would be the Equal Protection Clause, which was enacted in the wake of the Civil War to secure basic rights and freedoms for Black people who were newly freed from slavery.

To avoid such incoherence, it is necessary to distinguish *race consciousness*, which does not per se trigger special constitutional scrutiny, and *racial classifications*, which are presumptively prohibited. Although the Supreme Court has never clearly defined what constitutes a racial classification,[160] the Court's affirmative action cases suggest some critical factors. The programs that the Court has disapproved under strict scrutiny have taken a certain form—namely, they have applied racial criteria to individuals in a mechanical way that consistently favors one racial group over another. The Court is particularly concerned that racial classifications that benefit disadvantaged groups may operate as quotas, reserving a fixed number of slots for minority groups or aiming for a permanent racial balance.[161] Individual

[159] *See, e.g.*, Bagenstos, *supra* note 13, at 1143.

[160] *See id.* at 1119, 1142.

[161] *See, e.g.*, Regents Univ. Cal. v. Bakke, 438 U.S. 265, 289 (1978) (striking down a racial quota in medical school admissions); City of Richmond v. J.A. Croson Co., 488 U.S. 469, 499 (1989) (disapproving "the use of an unyielding racial quota"); Grutter v. Bollinger, 539 U.S. 306, 334 (2003) (explaining that a narrowly tailored program "cannot use a quota system"); United Steelworkers v. Weber, 443 U.S. 193, 208 (1979) (approving an affirmative action plan because it was not intended to maintain racial balance).

Justices have also expressed concerns that racial classifications are demeaning to individuals and will perpetuate hostilities and racial divisiveness.[162]

However, when government acts to address racial inequities in its policies and practices without relying on racial classifications, the concerns expressed by the Justices in the affirmative action cases do not apply. An awareness of racial realities may lead policy-makers to take actions that remove arbitrary barriers or level the playing field but do not impose quotas or determine individual outcomes on a racial basis. For example, an awareness of racial disparities in access to higher education might lead a university to increase spending to increase applications from racially marginalized communities—race-conscious action that does not entail the use of racial classifications. Policies that do not deploy racial categories in a determinative way can continue to treat individuals as persons, and thereby avoid inflicting dignitary harm or exacerbating racial tensions.

The Court's equal protection doctrine, then, targets racial classifications that operate in a mechanical way to systematically favor one racial group over another. At the same time, mere race consciousness by a government actor in developing policies and practices aimed at ameliorating inequities does not trigger strict scrutiny. Although uncertainty remains about exactly what constitutes a racial classification, the critical point is that the Equal Protection Clause does not forbid all race consciousness. Despite popular rhetoric about "colorblindness," government is not categorically prohibited from taking the realities of racial disparities into account.

* * *

Although the nuances differ, the statutory and constitutional prohibitions on race discrimination share a common structure. Courts have placed limits on how race can be used to advance racial equity goals, but not all race-conscious practices are presumptively unlawful. The special legal scrutiny imposed on affirmative action plans only kicks in after a plaintiff first shows that discrimination has occurred. In the statutory context, this requires the white plaintiff to establish disparate treatment—namely, that the affirmative action plan caused the plaintiff to suffer an adverse action causally connected to race. In the constitutional context, the use of racial classifications triggers scrutiny. The often-overlooked point, however, is that forms of race consciousness that do not amount to disparate treatment or racial classifications are permissible and do not require any special justification.

[162] *See, e.g.*, *Grutter*, 539 U.S. at 394 (Kennedy, J., dissenting) (expressing concern that programs that are tantamount to quotas will perpetuate hostilities).

IV.
THE LEGALITY OF RACE-AWARE ALGORITHMS

This Section takes the legal framework laid out in Part III and considers how it applies to race-conscious efforts to de-bias predictive algorithms. The legality of considering race in the model-building process is more complicated than previously recognized because race consciousness is not categorically forbidden by antidiscrimination law. Under both statutory and constitutional law, racial realities may be taken into account to create fair processes without triggering special legal scrutiny so long as doing so does not entail disparate treatment or reliance on racial classifications. Although there is not yet case law on this exact point, existing precedent leaves room for designers to explicitly consider the racial impact of predictive algorithms and to explore strategies for reducing or removing bias.

Unfortunately, some scholars have assumed that any use of race is a form of discrimination that requires special legal justification. For example, Daniel Ho and Alice Xiang assumed that equal protection doctrine prohibits the use of algorithmic fairness techniques.[163] Jason Bent similarly concluded that deploying an algorithm that includes a race-aware fairness constraint constitutes disparate treatment under Title VII.[164] These scholars then focus on whether such strategies can be justified under existing affirmative action doctrine.

I believe these scholars start their analysis in the wrong place. *Before* asking whether race-conscious model-building strategies pass muster under the Court's affirmative action cases, it is important to first ask: *is this particular race-conscious strategy a form of discrimination at all*? Given the complexity of the model-building process, there is no simple answer to that question. Rather, how the law views these strategies should depend upon when and how a particular approach takes account of race.[165] Only when a strategy constitutes discrimination in the first place is it necessary to ask whether it satisfies the requirements of affirmative action doctrine.

Part V discusses in greater detail why posing the questions in this order matters so much both doctrinally and conceptually. Some race-conscious approaches are not discriminatory, but rather entail removing unfairness. Recognizing this will lower the stakes both legally and rhetorically for designers interested in exploring options for reducing algorithmic bias.

This Part focuses on the first question—namely, "when do race-conscious strategies constitute discrimination?" Because model-building is a multi-step, iterative process, race may play many different roles in shaping the final model

[163] Ho & Xiang, *supra* note 11, at 134 (arguing that algorithmic fairness techniques "pos[e] serious legal risks of violating equal protection").

[164] Bent, *supra* note 10, at 825.

[165] *See* Hellman, *supra* note 13, at 848 (arguing that the conclusion that using race or other protected characteristics in algorithms is legally prohibited is "overstated;" doing so is likely impermissible in some cases and permissible in others).

and these differences should matter legally. The computer science literature now encompasses a vast array of proposed strategies for de-biasing algorithms[166] and it is not possible to analyze them all. Instead, Part V.A below applies the legal framework to a handful of examples to highlight the relevant considerations and to begin mapping out what space exists under current law for exploring bias-mitigating strategies. Part V.B takes a deeper dive into the causal question of when taking race into account causes an individual to experience an adverse outcome based on race. As explained there, this determination is complicated by the fact that no single "correct" model exists to serve as a baseline for determining the impact of racial fairness considerations.

A preliminary caveat is necessary here. By suggesting that a particular strategy is legal, I am not necessarily arguing that it constitutes a desirable policy or best practice. The best choice among competing models depends heavily on the setting, including the use for which the algorithm is deployed, the structure of the underlying data, the consequences of different types of errors, and other highly context-specific factors. My purpose is not to engage the debates about how best to define fairness or which techniques are preferable. These debates pose important policy questions rather than legal ones. The focus here is to explore when existing law permits race-aware strategies to achieve fairness ends.

A. De-Biasing Strategies

This section discusses some examples of algorithmic de-biasing strategies and analyzes whether they constitute disparate treatment or rely on racial classifications requiring special justification. It begins with a handful of examples that seem rather clear-cut—strategies that are easily categorized as permissible or impermissible under current law. It then analyzes some closer cases where the legal outcome is less certain, but good arguments exist that race-aware de-biasing strategies should not trigger any special legal scrutiny.

1. Dealing with Data Problems

One way bias can creep into a predictive model is due to problems with the training data. Depending upon the source or the manner in which it is collected, data may reflect systemic inequalities or human biases, or have other limitations in terms of accuracy or completeness that affect a model's predictive output.

Richardson et al. described how several jurisdictions developed predictive policing tools during periods in which corrupt or racially discriminatory policing practices were documented.[167] If the data used to build the models reflect those troubling practices, the predictive outputs would reproduce and further reinforce those harms. Similarly, Kristen Altenburger and Daniel Ho found that algorithms used to target food safety inspections disproportionately burden Asian

[166] *See supra* note 3 and sources cited therein.

[167] Richardson et al., *supra* note 1, at 40–41.

establishments when they rely on consumer complaints or online reviews because those data reflect anti-Asian stereotypes about lack of cleanliness.[168] Other studies have documented that consumer data tend to have more errors in records of marginalized populations and that disadvantaged groups are often less well-represented in large datasets. Models built on datasets with these limitations are likely to be less accurate for those groups, which risks deepening the disadvantages they face.

Developers might take a number of steps to address these limitations. Lehr and Ohm argued that the "playing with the data" stages of model-building offer numerous opportunities for reducing bias.[169] Developers could analyze the dataset for implicit biases before relying on it[170] or oversample from an under-represented group.[171] They could collect additional data from certain groups[172] or remove features for which there is little reliable data from marginalized groups. Or they might reject a specific dataset altogether. In the validation phase, they might engage additional techniques to identify bias in the training data and then take steps to mitigate the effect of that bias.

Each of these strategies would be race-conscious in the sense that they require an awareness of racial disparities. And acting on that awareness to prevent these issues from distorting the output of the model might entail race-conscious actions, such as collecting more information from an underrepresented racial group. Nevertheless, strategies like these, which aim to address problems or limitations of the data, should not raise legal concerns.

Consider again an employment selection algorithm. Suppose the designers discovered that supervisor evaluations included in the training data consistently downgraded Black employees relative to others even though they demonstrated the same level of productivity. The decision to remove that feature when training the algorithm is race-conscious but does not discriminate against white employees. Although they might have a better chance of promotion if the biased data is included, they are not entitled to evaluation by a model that gives them an unfair advantage. Similarly, a strategy such as oversampling a racial minority group is race-conscious, but does not create a suspect racial classification. Even if the employer does not hire the white candidate, it neither relied on race to make that particular decision nor put a mechanical thumb on the scale intended to favor

[168] Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. INSTITUTIONAL & THEORETICAL ECON. 98 (2018).

[169] Lehr & Ohm, *supra* note 18, at 657.

[170] Altenburger and Ho, for example, used scores from routine, scheduled food safety inspections to test whether consumer restaurant reviews suggesting food safety problems are accurate or display racial bias. *See* Altenburger & Ho, *supra* note 168, at 102–09.

[171] *See, e.g.*, Kamiran & Calders, *supra* note 3.

[172] Chen et al. propose a method for estimating the effect of poor data quality on the level of discrimination, arguing that additional data collection may be preferable to imposing fairness constraints. Chen et al., *supra* note 3, at 5.

only certain groups. Instead, these types of strategies are more accurately understood as removing bias from processes that would otherwise be unfair.

2. Problem Formulation

As discussed in Part II, one of the key decisions involved in building a predictive algorithm is how to operationalize a problem. Very often, the goals of prediction are abstract, high-level objectives (e.g., "find the best employees") that must be translated into an easily measurable target variable. The choice of the target variable can be highly consequential, affecting both predictions about specific individuals and the overall distribution of outcomes across populations. As Samir Passi and Solon Barocas put it, designers "should be paying far greater attention to the choice of the target variable, both because it can be a source of unfairness and a mechanism to avoid unfairness."[173] Thus, paying attention to how a problem is formulated is an important tool for avoiding unnecessary racial inequities. Although choosing the target variable to avoid inequity involves race-conscious decision-making, it clearly falls on the legally permissible side of the spectrum because it does not involve making decisions about individuals based on race.

Ziad Obermeyer et al. offered a good example of the critical role of problem formulation in avoiding racial bias.[174] Their study analyzed a health care algorithm used to predict which patients are high-risk and should be targeted to receive additional medical resources to improve outcomes. The researchers found that, among those given the same score by the algorithm, Black patients had more severe health conditions than white patients receiving the same score as measured by biological markers. The result was that white patients with fewer health conditions were targeted for the additional resources as compared with Black patients assigned the same risk score. The racial disparities in prediction arose because the designers had used medical expenditures as the proxy for health risk and, for a variety of economic, structural, and cultural reasons, Black patients consume less health care than white patients at the same level of health need. The researchers further demonstrated the impact of changing the problem definition to predict chronic health conditions rather than cost. Using this alternative target variable, the resulting algorithm was similarly highly predictive, but the racial disparity was substantially reduced.

3. Proportional Outcomes

At the other end of the spectrum are strategies aimed at ensuring proportional outcomes—what computer scientists refer to as "demographic parity." These strategies seek to equalize the probability of a positive outcome

[173] Passi & Barocas, *supra* note 48.

[174] Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 447 (2019).

across demographic groups. Put differently, they ensure that demographic groups receive positive outcomes in proportion to their actual representation. For example, if Black people are twenty percent of the relevant population, they should be positively classified twenty percent of the time, or within some specified range of that proportion (e.g., seventeen to twenty-three percent). These types of strategies are typically motivated by a desire to prevent a model from having a disparate impact.

One strategy to achieve demographic parity is to rank people according to the predicted target (e.g., success on the job, repayment of a loan), and then select a fixed percentage of the top scorers within each racial group to ensure that the benefit is distributed equally across groups. Another strategy would transform individuals' scores based on their racial group, so that the distribution of positive predictions is proportional across different subsets of the population. These types of strategies use information about race to achieve a proportional distribution of positive outcomes but likely violate antidiscrimination law.

In the hiring context, for example, these strategies might be considered race-norming—a practice of adjusting scores or using different cutoff scores on employment tests based on race that is specifically prohibited by Title VII.[175] A predictive algorithm might not be considered an "employment test" covered by the statute if it relies on historical data (e.g., the type of information found on a resume) rather than measuring responses on assigned tasks. And Title VII's prohibition on adjusting test scores does not apply outside of the hiring and promotion context. Nevertheless, the use of race to ensure a fixed distribution of outcomes would activate one of the Supreme Court's central concerns, namely, that race will be used to impose quotas or as a means of pursuing racial balancing. Thus, strategies that aim to achieve a particular numerical distribution of outcomes for their own sake will likely trigger close legal scrutiny.

4. Disparate Learning Processes (DLPs)

Disparate learning processes (DLPs) are strategies that use racial information during training, but do not allow the model to access race when making predictions.[176] Zach Harned and Hanna Wallach argued that DLPs offer the "just right" Goldilocks solution because they take race into account to de-bias algorithms but do not run afoul of antidiscrimination law by using race to

[175] 28 U.S.C. § 2000e-2(a)(l). The prohibition applies generally to any such adjustments or alterations of test scores taken on the basis of race, color, religion, sex, or national origin.

[176] *See* Lipton et al., *supra* note 3, at 2 ("DLPs operate according to the following principle: The protected characteristic may be used during training, but is not available to the model at prediction time.") (emphasis removed). For examples of DLPs, see Kamishima et al., *supra* note 3; Kamiran & Calders, *supra* note 3; Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez & Krishna P. Gummadi, *Fairness Constraints: A Flexible Approach for Fair Classification*, 20 J. MACH. LEARNING RSCH. 1 (2019); Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Discrimination Aware Decision Tree Learning*, 2010 IEEE INT'L CONF. ON DATA MINING 869 (2010).

predict outcomes.[177] Ignacio Cofone similarly argued that data pre-processing techniques that do not give an algorithm access to sensitive information at prediction time do not violate disparate treatment law.[178] A number of vendors who build applicant screening tools advertise that their model development process follows such a strategy,[179] likely in an effort to signal their compliance with antidiscrimination law.

DLPs have come under criticism as a solution to algorithmic bias. Some researchers have argued that they have a limited ability to remove bias because proxies for race are available in many datasets.[180] Others have argued that relying on DLPs is costly because they may reduce the accuracy of models[181] and can harm some members of the protected group.[182] These criticisms are relevant to the broader policy debate over which strategies for addressing algorithmic bias are preferable but do not speak to whether DLPs violate antidiscrimination law.

Drawing the line between race-awareness at training time versus prediction time has intuitive appeal because it maps onto the formalist notion that disparate treatment equates to intentional discrimination. At the same time, it offers a route for designers to take account of race at the model-building phase in order to de-bias algorithms. Although drawing such a line is a reasonable first cut at the problem, the distinction between using race at training time and at prediction time should not necessarily decide the legal question. Some uses of race at prediction time arguably should be considered lawful, a possibility I discuss below. And some DLPs, even though they do not rely on race to make predictions, may constitute disparate treatment.

Even though an algorithm does not access racial data at prediction time, it could nevertheless be discriminatory. It is widely understood that feature-rich datasets may contain variables that are highly correlated with race. If a nefarious actor used information about race to identify close proxies for a disfavored racial group and then trained the model to exclude members of that group, it would undoubtedly violate discrimination law even if race was not explicitly used at prediction time. Although technically a disparate learning process, the designer's intent to exclude on the basis of race would be sufficient to constitute disparate treatment.

But what if the intent is not nefarious, and instead the designer seeks to remove an adverse impact? The legal status of such a strategy is not entirely clear

[177] Zach Harned & Hanna Wallach, *Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer*, 47 FLA. ST. U. L. REV. 617, 639–40 (2020).

[178] Cofone, *supra* note 7, at 1429–30 (arguing that addressing discrimination by preprocessing data inputs to an algorithm would not be vulnerable to claims of disparate treatment).

[179] *See* Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, 2020 CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 469 (2020).

[180] *See, e.g.*, Dwork et al., *supra* note 3.

[181] *See, e.g.*, *id.*; Lipton et al., *supra* note 3.

[182] *Id.*

and likely depends on the particular approach taken. One possibility is that the designer uses demographic data during the training phase to assess whether a model has a disparate racial impact, and, if so, to determine which features contribute to producing that impact. The designer might then choose to eliminate some features that correlate highly with a protected characteristic after concluding that their use is not practically or morally justified. For example, if a machine learning algorithm was found to rely on irrelevant high school activities to predict job performance[183] or customer reviews that reflect racial stereotypes,[184] then removing those features because of their racial impact should not be legally problematic. Similarly, a designer building a model to predict recidivism might conclude that it is unfair to rely on factors an individual cannot control, such as having family members with criminal system involvement, particularly if those factors reflect racially discriminatory policing practices. These types of discretionary decisions by the model-builder are similar to permissible choices decision-makers often make when designing fair processes outside the algorithmic context.

It is less clear how to judge strategies that automate the de-biasing process. In the training stage, features that correlate with race may automatically be removed until any disparate impact is reduced to an acceptable level, or model structure might be modified and the results tested iteratively until observed disparate impacts have disappeared.[185] Zachary Lipton et al. raised the concern that redundant encoding may cause powerful DLPs that are intended to reduce disparate impact to effectively constitute a form of "treatment disparity" based on race.[186] Similar techniques might be used not to ensure demographic parity but to achieve some other definition of fairness, such as equal predictive accuracy across groups.

Whether or not these methods constitute disparate treatment is quite uncertain, but existing law suggests a couple of guideposts. The more it appears that a model is intended to produce proportional outcomes along racial lines without regard to other relevant considerations, the more vulnerable it will be to legal challenge. On the other hand, the more that the designers can articulate substantive (fairness) reasons for their choices—e.g., this feature was removed because the data it captured is unreliable or reflects past discriminatory practices—the more defensible the model will be.

[183] One applicant screening model found that having the name "Jared" and playing high school lacrosse correlated positively with job performance. Dave Gershgorn, *Companies Are on the Hook if Their Hiring Algorithms Are Biased*, QUARTZ (Oct. 23, 2018) https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/ [https://perma.cc/384B-92W5].

[184] *See, e.g.*, Altenburger & Ho, *supra* note 168 (finding that consumer ratings were biased against Asian restaurants).

[185] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel & Shayak Sen, *Use Privacy in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs*, 2017 ACM SIGSAC CONF. ON COMPUT. & COMMC'NS SEC. 1193 (2017).

[186] Lipton et al., *supra* note 3, at 2.

5. Using Race at Prediction Time

Some proposed strategies for addressing algorithmic bias require using information about race, not just in training, but also in making predictions. Although, as discussed above, using race to enforce demographic parity would likely trigger legal scrutiny, there are other ways race might be incorporated into a model. A robust debate exists over whether fairness in predicting recidivism risk requires equally accurate predictions across demographic groups, or equal false-positive or false-negative error rates. Regardless of what definition is chosen, strategies to ensure compliance with a fairness metric often require making use of race at prediction time. Race can also be used at prediction time to set different cutoff scores for decisions,[187] or to segment data and create separate classification models for each group.[188] Race might also be included as one feature among many in a model, interacting with other attributes and modifying their impact on the model's outcomes.

One approach would be to treat all these strategies as presumptively unlawful, on the assumption that any use of race at prediction time constitutes disparate treatment. While this is a common conclusion,[189] it is far too simplistic. As explored in Part III, the legality of race-conscious decision-making depends upon how race enters the decision-making process. When a racial classification is applied to an individual to achieve a goal of overall racial balance, a prima facie case of discrimination is established, triggering strict scrutiny. The same is not true when race is taken into account to build fair processes that are applied consistently across all individuals.

Under existing law, then, there are strong arguments that including race as a feature at prediction time does not always constitute disparate treatment or a forbidden racial classification. If a model relies solely on race, or uses race in a mechanical way to achieve numerical goals, it would likely trigger legal scrutiny. However, that information might be included in other ways that do not have the effect of favoring certain individuals because of their race. In a complex, feature-rich model, the effects of each feature can be quite subtle, shifting the weights given to other factors depending upon the statistical interactions between them. A model that takes account of race in this way might be warranted where different factors are relevant to predicting the target outcome for different groups.

[187] *See, e.g.*, Kleinberg et al., *supra* note 5.

[188] *See, e.g.*, Cynthia Dwork, Nicole Immorcila, Adam Tauman Kalai & Max Leiserson, *Decoupled Classifiers for Group-Fair and Efficient Machine Learning*, 81 PROC. MACH. LEARNING RSCH. 1 (2018).

[189] Bent, for example, argues that the human designer's decision to "[inject] a protected characteristic into the computer's programming" amounts to "sufficient intent to trigger disparate treatment protections." Bent, *supra* note 10, at 826. Harned and Wallach, and Cofone, similarly assume that any explicit use of race to mitigate bias would amount to disparate treatment. *See* Harned & Wallach, *supra* note 177, at 635; Cofone, *supra* note 7, at 1429.

For example, Sam Corbett-Davies et al. hypothesized that "housing stability might be less predictive of recidivism for minorities than for whites."[190] If this is the case, then including housing stability as a feature without taking race into account might cause the model to predict increased recidivism risks for all suspects who are housing insecure, even though the factor would be far less relevant for Black defendants. Cynthia Dwork et al. suggested another example: suppose the culture of one subgroup steers the most talented students toward engineering, rather than finance, whereas the culture of another subgroup does the opposite.[191] If a model predicts which applicants are most talented by prioritizing students who focused on finance, it will be systematically unfair to members of the other subgroup.

In situations like these, failing to take into account race when constructing a model will force all factors to have the same impact on the predicted outcome for everyone, even though in reality a factor may influence outcomes for members of different groups differently.[192] And where one group is more numerous than another in the data, the model will necessarily disadvantage the smaller group because its predictions will be less accurate for that group. On the other hand, including race in the model will not necessarily cause any disadvantage to the majority group, while at the same time improving accuracy for the minority. In situations like these, a race-aware model can improve *both* accuracy *and* fairness for all individuals.[193]

Apart from bowing to formalist conventions, it is difficult to see why including the sensitive attribute in these types of circumstances constitutes disparate treatment. Individuals are not being reduced to their racial identities and sorted on that basis. The consideration of race does not drive outcomes toward some fixed numerical proportions. And although the overall distribution of positive outcomes might shift somewhat as a result,[194] the direction and distribution of such changes are not easily predicted. Because no individual has been deprived of an entitlement or barred from an opportunity because of race, incorporating race into a model in this way should be considered lawful.

All of this is not to say that allowing a model to access race at prediction time is always unproblematic. Clearly, there will be instances when doing so is discriminatory. The point here is that the sole fact that the model is "race-aware,"

[190] Corbett-Davies et al., *supra* note 7, at 805.

[191] Dwork et al., *supra* note 3, at 218.

[192] More often, Black people, who likely represent a numerical minority, will be disadvantaged because when one group has much greater representation in the dataset, the model will perform better overall by selecting features that best predict success of the majority group. Predictions will be less accurate for the minority group, potentially disadvantaging them in the long run.

[193] *See* Hellman, *supra* note 13, at 855.

[194] The actual impact on the distribution of outcomes is uncertain and depends upon the structure of the underlying data and the relationships among existing features. *See, e.g.*, Mayson, *supra* note 8, at 2298–300 (explaining how equalizing error rates across racial groups could increase the burden on communities of color); Estornell et al., *supra* note 52.

even at prediction time, should not be decisive of the question whether disparate treatment has occurred. Instead, determining whether a model entails disparate treatment requires a closer inquiry into the role race plays and its impact on the decision-making process.

B. The Causation Question

The issue of causation gained salience in Title VII cases after the Supreme Court's decision in *Bostock v. Clayton County*, which emphasized the relevance of the "but-for" causation standard.[195] Applying that standard, one might argue that if an algorithm takes race into account at prediction time, then race must be playing a causal role in any adverse outcome. It turns out, however, that determining causation is more complicated than it initially appears.

A naïve approach might ask whether the outcome would differ if the race variable for a rejected individual was changed, but all other variables were left untouched. This way of framing the question seems to accord with Justice Gorsuch's suggestion in *Bostock*: "change one thing at a time and see if the outcome changes. If it does, we have found a but-for cause."[196] Gorsuch's epigram, however, was developed to identify causation when dealing with human decision-makers. It is the wrong way to think about the question in the context of algorithmic tools.

Algorithms have no agency and therefore it is a mistake to ask how a decision might change by looking inside the machine. Instead, the inquiry should focus on the decisions made by the *humans* who created the machine. In other words, the proper comparison is not the algorithm's output if the rejected applicant's racial identity was different, but the outcome that would have occurred absent the designer's choice to take race into account. The relevant question is whether the decision to incorporate race in the model has the requisite causal relationship to the applicant's rejection.

Answering that question is not as simple as flipping a switch.[197] If the decision was motivated by an intent to exclude the racial group to which the plaintiff belongs, or to ensure some fixed numerical level of representation, then a causal connection seems clear. With more complex algorithms, however, the fact that someone did not receive a positive outcome under a race-aware model

[195] Bostock v. Clayton Cnty., 140 S. Ct. 1731, 1740 (2020).

[196] *Id.*

[197] Issa Kohler-Hausmann offers a broad critique of this counterfactual approach to proving causation in racial discrimination cases. *See generally* Issa Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination*, 113 NW. U. L. REV. 1163 (2019). In contrast, Bent assumes that determining the counterfactual is both simple to do and theoretically correct. Bent, *supra* note 10, at 829 (arguing that "the race-aware fairness constraint could be temporarily removed" and then "the results for any individual candidate could be directly compared with and without the fairness constraint" to see if the outcome changes, thereby proving causation).

does not mean that the designer's choice to take race into account caused the negative outcome.

The difficulty is that, as discussed in Part II, there is no single "correct" model that exists prior to considering race that represents the relevant counterfactual. If racial equality concerns had not been taken into account, the designers would have to choose from numerous different mathematical models, select among possible training datasets, decide upon a sampling strategy, and determine which features to include. Not only is there a wide range of legitimate choices that could be made, but the inherent randomness of some steps in the process means that the outcome for any given individual can be quite unstable, sometimes reflecting chance, as much as relevant considerations.

Recent work in computer science illustrates this instability in model outcomes. Emily Black and Matt Fredrickson documented how the removal of a single individual from the training dataset for machine-learning models can affect whether or not another individual receives a positive outcome.[198] This effect occurs in the absence of any de-biasing efforts and is observed with "surprising frequency."[199] Andrew Estornell et al. similarly showed that, prior to imposing group fairness constraints, significant natural variation occurs in the outcomes for a given individual, depending upon choices made in building a model.[200] For example, the outcome for a given individual varies when using the exact same learning algorithm and the same dataset to train a predictive algorithm, simply due to the random draws of the training subsample. This variation would only increase if the choices among different learning models, different datasets, and different parameters are taken into account as well.

With no clear baseline mode for comparison, it is difficult to know whether a particular individual would or would not have received the benefit absent the inclusion of race in the model. Because each individual faced some risk of rejection across the multitude of possible alternative models, it would be more accurate to characterize the effect of taking race into account as altering the probability of success for the individual applicant.

The rejected applicant might then argue that the required causal requirement is met by showing that her odds of success were reduced by the choice to adopt a race-aware model. Aside from the practical impossibility of calculating those odds for a particular individual across all plausible alternative models, the rejected applicant's argument faces other difficulties. It rests on the implicit assumption that any race-conscious effort to reduce bias against Black applicants will automatically and consistently work to the disadvantage of white applicants. But this need not be the case when it comes to de-biasing algorithms. Depending upon the approach taken, a race-aware strategy will not necessarily have uniform effects within a racial group. It may reduce the odds of a positive

[198] Black & Fredrikson, *supra* note 53.

[199] *Id.* at 285.

[200] *See* Estornell et al., *supra* note 52.

outcome for some white applicants, but not all. Other white applicants might see their chances of success increase.

Consider again the hypothetical model that includes race as a feature because it has been determined that housing stability has a different effect on the recidivism risk for different racial groups. Because the weight given to the feature housing stability will vary for different racial groups, individuals within each group will be affected differently based on information about their housing situation, not solely their race. Race alone, in this type of situation, does not have a determinative effect on outcomes, suggesting that the requisite causal connection is absent.[201]

This conclusion is consistent with cases finding that a change in recruitment procedures is not discriminatory. Even though a race-conscious decision to expand the applicant pool creates more competition, the mere fact that an individual's odds of success were altered does not constitute disparate treatment. Similarly, so long as a race-aware model is neither intended to exclude nor designed to systematically disadvantage one racial group, there is a strong case that no disparate treatment has occurred.

C. Looking Beyond the Code

Once an algorithm has been deployed in the real world, it might behave differently than it did in the testing environment. Good design practices call for on-going monitoring of the performance of algorithms "in the wild" and making adjustments as appropriate to improve accuracy. The same is true for fairness. Entities relying on predictive algorithms should audit their performance to detect unjustified racial disparities.[202] Once again, this process requires a measure of race consciousness, but that fact alone does not trigger legal concerns. If an algorithm turns out to have unjustified racial impacts, the entity is free to modify it or abandon its use, so long as it does not disrupt any legitimate, settled expectations in doing so.

Such expectations generally do not exist apart from highly unusual situations, like in *Ricci*, where the City announced that a particular test would be used for promotion decisions, and where employees invested substantial time and resources to study for that test in reliance on the declared policy.[203] Because predictive algorithms generally rely on observational data, and not on separately administered tests, *Ricci*'s holding has little relevance to most decisions to alter models prospectively or to change or abandon them altogether. Of course, *what*

[201] If using a race-aware algorithm to make decisions has a disproportionate negative effect on disadvantaged groups without justification, it might still be challenged under a disparate impact theory, but the focus of this discussion is whether it constitutes disparate treatment.

[202] *See* Kim, *supra* note 100, at 196 (explaining the importance of auditing algorithms for discrimination once implemented, rather than relying solely on technical tools when building models to prevent bias).

[203] *See supra* Part III.A.

an entity is permitted to do in order to de-bias a model may be restricted by antidiscrimination law, but as discussed above, many race-conscious strategies are likely permissible under existing doctrine.

The discussion in Part IV largely focused on race-conscious model-building strategies when addressing a well-defined optimization problem. However, as discussed earlier, one of the most consequential decisions determining the racial impacts and fairness of algorithms occurs at the outset, when formulating the problem to be solved. Scrutinizing the target variable for its racial impacts and endeavoring to select a target that is not implicitly biased against disadvantaged groups would not run afoul of antidiscrimination law.

Entities that rely on predictive algorithms also make choices about how those tools fit into their overall decision process. For example, an employer might use an algorithm to actually make hiring decisions, or to screen out clearly unqualified candidates, or merely as an estimate of one aspect of future job performance that is weighed alongside other factors in the ultimate hiring decisions.[204] An employer might even seek to use algorithmic processes to counter known human biases, for example, by enacting a technological version of the "Rooney Rule" to ensure that some members of previously disadvantaged groups are included in the group of candidates that is given closer scrutiny.[205] The law likely permits entities to take racial impacts into account when deciding how to incorporate algorithms into their decision processes so that the overall process is equitable.

V.
THE DIFFERENCE BETWEEN NON-DISCRIMINATORY STRATEGIES AND AFFIRMATIVE ACTION CASE LAW

A. Limited Applicability of Affirmative Action Doctrine

As seen in Part IV, it is a mistake to assume that any race consciousness in the model-building process automatically triggers close legal scrutiny under affirmative action doctrine. Many de-biasing strategies do not amount to disparate treatment under statutory law or racial classifications under equal protection doctrine. They are more accurately seen as entirely permissible efforts to *remove* unlawful or unfair biases that would otherwise distort the decision process. Even when a model takes account of race at prediction time, there are strong arguments that it does not amount to discrimination, depending upon how it is incorporated in the model.

Scholars, however, have tended to overlook these important nuances, treating all de-biasing strategies from the outset as forms of disparate treatment

[204] *See* Miranda Bogen & Aaron Rieke, *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*, UPTURN 5–6 (2018).

[205] See Jon Kleinberg & Manish Raghavan, *Selection Problems in the Presence of Implicit Bias*, ARXIV:1801.03533 [CS, STAT] (2018) for a formalization of the tradeoffs involved in such a practice.

or racial classification that must be analyzed under affirmative action case law. Jason Bent, for example, argued that if an employer adds a race-aware fairness constraint in order to avoid producing racially disparities that harm minorities, it has likely engaged in disparate treatment under Title VII.[206] Although he recognized that models can take race into account in a variety of ways,[207] and that differences in the timing and method of doing so "may prove important,"[208] his legal analysis considered only a generic "race-aware fairness constraint" without parsing how different de-biasing strategies might—and arguably *should*—be treated differently for purposes of determining liability. Instead, he asserted that the "most promising" way for the employer to defend the use of a race-aware fairness strategy is to argue that it is a permissible form of voluntary affirmative action under Title VII because it meets the requirements laid out in *Weber*.[209]

Similarly, Daniel Ho and Alice Xiang broadly asserted that algorithmic de-biasing strategies "pose serious legal risks of violating equal protection"[210] without recognizing that different approaches warrant different legal responses. They examined a few examples, such as using different cutoff scores for Black and white applicants when determining loan eligibility, and concluded that algorithmic fairness strategies likely run afoul of the equal protection clause. Their analysis of the specific examples may have been correct; however, they seemed to assume more broadly that any consideration of race in building a model will also trigger strict scrutiny. As a result, they focused their efforts on showing how algorithmic fairness efforts could meet the requirements of strict scrutiny, without considering that some strategies might not entail racial classification, and therefore not be subject to such scrutiny at all.

The ultimate goal of these scholars is to defend the lawfulness of race-conscious strategies and they do so by invoking affirmative action doctrine. As argued above, however, not all race-conscious strategies should be analyzed under affirmative action doctrine because they are not discriminatory in the first place. The difference in framing of the legal issue is consequential. As a practical matter, it will be much harder to legally defend race-conscious de-biasing strategies if it is assumed that they must be justified under affirmative action doctrine. And as a conceptual matter, utilizing the affirmative action legal frame invokes the wrong set of assumptions. Because of misunderstandings about affirmative action, the rhetoric surrounding it erroneously suggests that any

[206] Bent, *supra* note 10, at 824–25.
[207] *Id.* at 816–24.
[208] *Id.* at 824.
[209] *Id.* at 835.
[210] Ho & Xiang, *supra* note 11, at 134.

effort to reduce algorithmic bias inflicts harm on members of previously advantaged groups,[211] even if the prior arrangements were unfair.

B. Practical Consequences

Consider the practical perspective first. Suppose an unsuccessful white applicant sues an employer that relied on a race-aware algorithm for discrimination. If the employer acted to remove a source of racial bias, no disparate treatment has occurred and the employer should bear no further burden of justification. Bent's analysis, however, started with the assumption that a prima facie case of disparate treatment exists. By assuming that race-aware models are discriminatory, his approach placed all such efforts under a legal cloud unless it could be shown that they address a "manifest racial imbalance" in "traditionally segregated job categories" and do not "unnecessarily trammel[]" the rights of other employees.[212] This added legal burden would likely discourage some employers from trying to understand whether the algorithms they utilize are implicitly biased and to seek proactively to fix these issues. Such an outcome would be in direct contravention of Title VII's purpose of encouraging employers to engage in self-examination and voluntarily seek to avoid discriminatory practices.[213]

Ho and Xiang similarly concluded that race-aware algorithmic fairness strategies, when used by government actors, likely violate the anti-classification principle of the Equal Protection Clause and therefore face "serious legal risks."[214] They then turned to affirmative action cases in the government-contracting context to argue that algorithmic fairness strategies can be legally defended where a state actor can show that its own past discrimination contributed to current racial disparities and that the means chosen—the method of combatting the algorithmic bias—is narrowly tailored. Because they believed these strategies would pass muster when they are calibrated to respond to discrimination by a specific government actor, they urged technologists to "quantify specific forms of historical discrimination."[215]

Even assuming this is the best approach for satisfying strict scrutiny, it will not often succeed, because establishing historical discrimination by a specific government actor is exceedingly difficult. Part of the problem stems from the law. The Supreme Court held in *Washington v. Davis* that mere statistical disparities in outcomes across racial groups are not evidence of government discrimination.[216] Instead, what is required is proof of a racially discriminatory

[211] Hellman similarly argues that the term "algorithmic affirmative action . . . misleadingly conveys that the explicit use of race within algorithms provides minorities with a benefit when compared with non-minorities." Hellman, *supra* note 13, at 848 n.88.

[212] United Steelworkers v. Weber, 443 U.S. 193, 196 (1979).

[213] *Id.* at 204.

[214] Ho & Xiang, *supra* note 11, at 134.

[215] *Id.* at 148.

[216] Washington v. Davis, 426 U.S. 229, 238 (1976).

purpose.[217] That sort of proof of motive to explain historical disparities is elusive, especially as we move further in time from explicitly discriminatory government policies. And even if this type of evidence were available, government entities will be unwilling to voluntarily assemble evidence of their own past discrimination, which would open them to liability.[218]

While Ho and Xiang believed they had found a path forward for developing and implementing algorithmic fairness strategies, it is an exceedingly narrow one, and unnecessarily so. Relying on proof of past discrimination to justify de-biasing efforts is not only unrealistic, but it also misses entirely one of the crucial reasons why the effort to create fair algorithms is so pressing. As government entities expand their use of algorithmic tools, the risk of bias arises not so much from the evil intent of some bureaucrat or computer programmer, but from the likelihood that poor choices in building models encode or reproduce patterns of inequality, thereby deepening the disadvantages faced by historically marginalized groups. A backward-looking focus on establishing historical discrimination by a specific government actor does nothing to identify and address these concerns.

Algorithmic fairness efforts should instead seek to understand where and how choices in the model-building process may introduce unfairness through flawed assumptions, biased data, and the like. If a government entity learns that an algorithm denies benefits to Black claimants at higher rates than their white counterparts, it should investigate to understand the source of the disparity. The differential grant rates may not reflect actual differences in eligibility, but instead result from artifacts of the model-building process such as a lack of accurate data about marginalized groups or cognitive biases on the part of humans responsible for coding key inputs. Government actors should not be required to demonstrate that they engaged in intentional discrimination in order to correct those problems. In other situations, it may be the case that different factors influence the relevant outcome for different racial groups, or that some data is noisier for certain groups. In these situations as well, taking race into account may be necessary to ensure fairness for all individuals and doing so should not depend upon a finding of prior discrimination by the government actor.

Importantly, the argument that some race-conscious strategies are nondiscriminatory and therefore require no special justification does not preclude defending other practices under affirmative action doctrine when warranted. If a strategy is found to constitute disparate treatment or entail the use of racial classifications, the arguments made by Bent and Ho and Xiang become relevant, and may provide a legal basis for justifying those strategies. To *start* the analysis there, however, concedes—often inaccurately and unnecessarily—

[217] *Id.* at 240.

[218] As the Second Circuit put it, requiring government actors to provide evidence of their own past discrimination to show that they have a compelling interest in taking action to prevent disparate impact puts them “on the horns of a dilemma.” Barhold v. Rodriguez, 863 F.2d 233, 237 (2d Cir. 1988).

that efforts to remove or prevent algorithmic bias against disadvantaged communities of color somehow discriminates against white people.

Once triggered, strict scrutiny is a demanding standard to meet. Although not necessarily fatal, it imposes a high burden, and the Supreme Court has rarely found it to be satisfied. From a practical perspective, then, it is critically important to differentiate at the outset those strategies which are race-aware, but permissible, because they do not involve discrimination, from those which impose racial classifications on individuals in ways that trigger strict scrutiny.

C. Conceptual Differences

On a conceptual level, it also matters quite a lot if de-biasing strategies are characterized as non-discriminatory, as opposed to trying to justify them under affirmative action doctrine. Analyzing race-conscious efforts to ensure fair models under the affirmative action cases invokes a set of assumptions and surrounding rhetoric that are inapt, even misleading, in the context of predictive algorithms.

Much of the debate about algorithmic fairness has focused on the specific example of recidivism prediction software. However, the popular concept of affirmative action is largely irrelevant to the criminal law context. Typically, affirmative action describes race-conscious efforts intended to assist disadvantaged minority groups by increasing their access to scarce resources. Our current system of criminal law enforcement, however, does not offer resources and opportunity, but instead threatens individuals with loss of liberty and other punitive sanctions, as well as devastating collateral consequences in the civil sphere. As Michelle Alexander and others have documented, a criminal record is often the basis for denying employment, housing, educational opportunities, public benefits, and the right to vote.[219] And because of discriminatory policing and prosecutorial practices, these effects are visited disproportionately on Black and other communities of color.[220]

Given these realities, it makes no sense to treat race-conscious efforts to reduce racial bias in criminal law enforcement as "affirmative action" that somehow benefits Black defendants at the expense of white defendants. Unlike jobs or spots in a college class, incarceration is not a scarce resource that different groups are competing for. Reducing arrest and incarceration rates in the Black community need not result in greater enforcement efforts or increased imprisonment of white people. Rather, proposals for change that are motivated by racial justice concerns—from reform to abolition—would benefit white criminal defendants as well. For all these reasons, the affirmative action framing seems inapt in capturing what is at stake in the criminal law context. Invoking that frame has the unfortunate consequence of erroneously suggesting that efforts

[219] ALEXANDER, *supra* note 22; Jain, *supra* note 21.

[220] ALEXANDER, *supra* note 22.

to address unequal impacts of criminal enforcement on Black communities somehow burden white people and therefore should be suspect.

The concept of affirmative action appears more relevant in contexts like education and employment, but can nevertheless be misleading when applied to efforts to de-bias algorithms that are used in those domains. In popular discourse, the term "affirmative action" has come to be associated with race- and sex-based preferences, and is often conceived to involve numerical quotas. That conceptualization in turn activates a set of stock arguments that these policies undermine meritocratic norms and harm innocent white people. These objections, however, do not apply to many algorithmic de-biasing strategies.

Implicit in the arguments against affirmative action are two related premises. One is that these policies are "preferences"—i.e., that they grant benefits to Black people or other marginalized groups that they would not have received without putting a thumb on the scale in their favor. The second is that these policies necessarily impose harms on white people because they are not part of the preferred group. Both premises in turn rest on the assumption that absent consideration of race, there is some fair, neutral baseline for distributing benefits or opportunities that is being disrupted.[221] For example, opponents of affirmative action in education often assume that the fair way to allocate places at a university is based on test scores. Any deviation from this presumed-to-be-fair baseline is deemed a preference that unfairly harms those who would have won under the old rules.

Critical race scholars have long challenged the notion that past practices objectively measure merit or that the prior distribution of resources and opportunities is a fair baseline against which to assess race-conscious measures.[222] They point to the myriad of ways in which private discrimination and implicit biases create systematic disadvantage for marginalized groups. Numerous empirical studies have shown that Black people, regardless of economic class, are often subjected to biased and inaccurate assessments of their abilities,[223] and face systemic disadvantages in employment, housing, and other markets. This type of evidence led Luke Charles Harris and Uma Narayan to argue that affirmative action policies are not "preferential treatment," but should be understood as "attempts to equalize opportunity" in a society "marked by pervasive inequalities."[224] Devon Carbado similarly rejected the framing of

[221] *See, e.g.*, Luke Charles Harris & Uma Narayan, *Affirmative Action and the Myth of Preferential Treatment: A Transformative Critique of the Terms of the Affirmative Action Debate*, 11 HARV. BLACKLETTER L.J. 1, 14 (1994).

[222] *See, e.g.*, *supra* note 113 and sources cited therein.

[223] Carbado, *supra* note 56, at 1118, 1127; Devon W. Carbado, Kate M. Turetsky & Valerie Purdie-Vaughns, *Privileged or Mismatched: The Lose-Lose Position of African Americans in the Affirmative Action Debate*, 64 UCLA L. REV. DISCOURSE 174, 215–18 (2016).

[224] Luke Charles Harris & Uma Narayan, *Affirmative Action as Equalizing Opportunity: Challenging the Myth of "Preferential Treatment*," 16 NAT'L BLACK L.J. 127, 128, 132 (1998).

affirmative action as a preference, characterizing it instead as "a countermeasure that offsets racial disadvantages."[225]

Devon Carbado argued that shifting the frame in this way allows for a more full-throated defense of affirmative action.[226] The difficulty, however, is that the term "affirmative action" is not precisely defined and has been applied to a broad range of efforts to address racial disparities. Some are easily understood as countermeasures—for example, expanding recruitment efforts instead of relying on word-of-mouth hiring that favors ethnic groups currently dominant in a workplace. Other policies, however, are more vulnerable to being characterized as preferences, such as the program challenged in *Bakke* which admitted applicants from racial minority groups through a special admissions program to fill at least sixteen places in the class.

Whether or not such a special admissions program amounts to a preference, most algorithmic de-biasing strategies operate quite differently, and therefore, they do not raise the concerns that drove the Court's decisions in *Bakke* and subsequent affirmative action cases. Unlike under policies that reserve a fixed number of spots for racial minorities, the actual impact of most race-aware strategies on the distribution of outcomes is somewhat uncertain. While these strategies tend to increase positive outcomes for previously disadvantaged groups,[227] they do not necessarily drive results toward proportional outcomes. Nor do most of these strategies impose a racial classification on individuals that is determinative of whether they receive a benefit or an opportunity. Instead, an awareness of race informs choices that go into shaping a model without pre-ordaining outcomes for particular individuals.

The problem with characterizing any race-conscious de-biasing effort as "algorithmic affirmative action" is that it conflates a broad range of strategies and equates them all with the types of policies that have been most vulnerable to legal challenge in the past. Worse still, it validates the false notion that de-biasing efforts involve a departure from some fair, objective method of making decisions. This reinforces the mistaken belief, common among non-technical people, that algorithms are objective and neutral, and that considering racial equity somehow entails a departure from the "true" model. For example,

[225] Devon W. Carbado, *States of Continuity or State of Exception? Race, Law and Politics in the Age of Trump*, 34 CONST. COMMENT. 1, 9 (2019). *See also*, Carbado et al., *Privileged or Mismatched: The Lose-Lose Position of African Americans in the Affirmative Action Debate*, *supra* note 223, at 180 (arguing that affirmative action is a mechanism for leveling the playing field rather than providing a preference).

[226] Carbado also argues that it provides an opening to ask whether strict scrutiny is the appropriate standard when race-conscious government action operates as a countermeasure for disadvantage rather than a racial preference. Carbado, *supra* note 56, at 1128. While this argument has considerable appeal, I do not engage with it here because my more limited goal is to explore the contours of the doctrine as it currently exists.

[227] This is not always the case depending upon the fairness model and the structure of the underlying data. *See, e.g.*, Estornell et al., *supra* note 52; Lipton et al., *supra* note 3; Mayson, *supra* note 8.

discourse often assumes that when a firm uses a predictive model to select employees or to decide who will get a loan, there is a "correct" solution. In fact, as discussed above, there is no single, canonical model that best predicts future outcomes, but instead, there are multiple equally plausible models. The final model that is selected reflects a series of choices, tradeoffs, random effects, and weighing of values, each of which shifts the odds that any particular individual will receive the benefit.[228]

This richer understanding of the model-building process means that the choices made along the way, even ones taken with racial equity goals in mind, are not disrupting some preexisting fair allocation. Where known biases affect the data, or past practices worked to exclude certain groups, members of previously favored groups have no entitlement that the designer's choices retain those advantages. Correcting those biases and inaccuracies are steps towards greater fairness, not preferences that burden one group at the expense of another. As Justice Powell recognized in a largely overlooked footnote[229] in *Bakke*: "To the extent that race and ethnic background were considered only to the extent of curing established inaccuracies in predicting academic performance, it might be argued that there is no 'preference' at all."[230]

* * *

As explained at the outset, my purpose here has been to examine existing law and doctrine to determine what space exists for race-conscious efforts to de-bias algorithms. I have concluded that under current law many algorithmic de-biasing strategies do not appear to entail disparate treatment or the use of racial classifications at all. They should, therefore, be legally permissible without having to survive some form of heightened legal scrutiny.

However, one might question whether this conclusion will hold in light of the changed composition of the Supreme Court. Regarding its race jurisprudence, scholars have pointed to cases like *Parents Involved*[231] and *Shelby County*[232] as illustrations of the Roberts Court's "post-racial" worldview, which ignores persistent patterns of racial injustice and assumes that discrimination is now rare and aberrational.[233] The addition of three Trump appointees to the Court

[228] *See supra* Part II.

[229] *See generally* Carbado, *supra* note 56.

[230] Regents Univ. Cal. v. Bakke, 438 U.S. 265, 306 n.43 (1978).

[231] Parents Involved Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701 (2007) (striking down school assignment plans that used racial classifications to promote racially integrated schools).

[232] Shelby Cnty. v. Holder, 570 U.S. 529, 529 (2013) (holding unconstitutional the coverage formula in the Voting Rights Act that determines which jurisdictions are subject to preclearance requirements on the grounds that conditions regarding access to voting by racial minorities have changed).

[233] *See, e.g.*, Mario L. Barnes, "*The More Things Change . . .": New Moves for Legitimizing Racial Discrimination in a "Post-Race" World*, 100 MINN. L. REV. 2043 (2016); Cedric Merlin Powell, *The Rhetorical Allure of Post-Racial Process Discourse and the Democratic Myth*, 2018 UTAH L. REV. 523 (2018).

has led to predictions that it will double-down on colorblindness and move to end all affirmative action.

These predictions appeared even more plausible when the Court granted certiorari in January 2022 to *Students for Fair Admissions v. Harvard College* and *Students for Fair Admissions v. University of North Carolina*, a pair of cases challenging affirmative action programs in university admissions that are currently pending before the Court. The lower court decisions in both cases fell well within existing doctrine, suggesting that the Justices may be preparing to overrule the Court's precedents in the area. Indeed, the petitions for certiorari in these cases specifically ask the Court to overrule its decision in *Grutter v. Bollinger*,[234] which permits universities to consider race as part of an individualized, holistic application review process in order to create a diverse student body. Although the Justices regularly pledge fealty to precedent, they have in recent years moved aggressively to undermine or overturn cases with which they disagree. The Roberts Court has already faced criticism for its willingness to overturn long-established precedent in other areas of the law.[235]

While the membership of the Court certainly shapes its rulings, it remains important to take the Justices at their word and to engage the explanations they offer for their decisions. A legal realist stance does not render precedent irrelevant. Courts must still act *through* doctrine. Long-established norms demand that they justify their decisions based on precedent and legal reasoning. Because advocates and practitioners also have to work with existing case law, it makes sense to engage with and leverage existing doctrine to the extent possible.[236]

How then will the outcome of the college admissions cases affect the legality of race-aware algorithms? That question is impossible to answer in advance, of course, because it depends not only on the outcome of those cases, but, more significantly, on the reasons the Justices offer for reaching their conclusions. It is possible, although I think unlikely, that the Court could speak in sweeping terms, concluding that any form of race consciousness anywhere in the decision process triggers strict scrutiny. Such broad reasoning would go far beyond the issues raised in the admissions cases. It would not only end affirmative action in college admissions, but also entail a radical jurisprudential

[234] Grutter v. Bollinger, 539 U.S. 306, 343 (2003).

[235] *See, e.g.*, Donald Ayer, *The Supreme Court Has Gone Off the Rails*, N.Y. TIMES (Oct. 4, 2021), https://www.nytimes.com/2021/10/04/opinion/supreme-court-conservatives.html [https://perma.cc/4FGM-A4GE] (arguing that the Supreme Court is disregarding long-standing precedent, citing as examples recent decisions in *Cedar Point Nursery v. Hassid*, *Fulton v. City of Philadelphia*, and *Americans for Prosperity Foundation v. Bonta*); Catherine L. Fisk & Martin H. Malin, *After* Janus, 107 CALIF. L. REV. 1821 (2019) (criticizing the Court's decision in *Janus v. AFSCME* for disrupting long-established law governing public sector unions).

[236] *See, e.g.*, Daniel Harawa, *Lemonade: A Racial Justice Reframing of the Roberts Court's Criminal Jurisprudence*, 110 CALIF. L. REV. 681 (2022) (arguing that the Supreme Court's recognition of racial injustices in several recent criminal law cases could provide a jurisprudential hook for advocates to push for racial justice reforms).

shift, destabilizing broad swaths of existing civil rights and antidiscrimination doctrine. It would also call into question many widely accepted practices, entangling the courts in detailed review and supervision of routine goal-setting and policy decisions by both government and private actors.

What is more likely is that the Court's holding will be specific to higher education, particularly since the questions presented by the petitioners appear limited to that setting. When race is considered in the model-building process in order to de-bias algorithms, not only is the context very different, but the ways in which race is taken into account and the effects of doing so are quite distinct from treating race as a "plus" factor in a college application file. As a result, there will likely remain room for race-conscious efforts to remove bias from algorithms, regardless of the outcomes in the pending lawsuits.

In any case, if the Justices have set their sights on further restricting affirmative action, then it is all the more important to be conceptually clear about how algorithms work and to distinguish nondiscriminatory de-biasing strategies from the type of affirmative action that has triggered the Court's disapproval in the educational setting. Once the complex, multi-step process of model-building is understood, it becomes clear that many available strategies for de-biasing algorithms bear very little resemblance to the policies disapproved of by conservative Justices in past affirmative action cases. In those cases, race was used to ensure fixed numerical outcomes[237] or to tip the scales decisively in favor of one race over another.[238] Strategies like addressing data quality and representativeness or adjusting the target variable to avoid biased measures do not entail using race to determine outcomes in individual cases. They are therefore entirely distinct from the policies that provoked concerns in the Court's prior affirmative action cases.

Models that make use of information about race at prediction time fall into an area of greater uncertainty. Even then, when race is taken into account in a manner that does not systematically favor one racial group over another in making individual decisions, the concerns expressed by some Justices about demeaning individuals or dividing communities do not apply. As a result, there are strong arguments that these strategies do not involve disparate treatment or racial classifications, and therefore warrant no special scrutiny.

[237] *See, e.g.*, Regents Univ. Cal. v. Bakke, 438 U.S. 265 (1978) (scrutinizing admissions policy that reserved sixteen out of one hundred places in medical school class for members of racial minority groups); United Steelworkers v. Weber, 443 U.S. 193 (1979) (examining program that required that Black workers make up half of all those admitted to an on-the-job training program).

[238] *See, e.g.*, Adarand Constructors v. Pena, 515 U.S. 200, 227 (1995) (addressing challenge to federal contracting program that presumed members of racial minority groups are socially and economically disadvantaged and therefore entitled to preferences in contracting); Gratz v. Bollinger, 539 U.S. 244 (2003) (rejecting university admissions policy which automatically awarded twenty points to applicants from underrepresented racial or ethnic minority groups).

CONCLUSION

Scholars and advocates concerned about bias in predictive models have begun calling for "algorithmic affirmative action." The call to pay attention to the risks of algorithmic discrimination and to address the harms they may cause to racially subordinated groups should be heeded. However, characterizing efforts to ensure fair algorithms as a form of "affirmative action" erroneously suggests that the Court's past affirmative action cases are directly applicable. This framing of the issue is unfortunate, because it misapprehends what actually happens when designers engage in race-aware strategies to reduce bias in models and invokes a set of assumptions that are not relevant.

Despite rhetoric about colorblindness, the law does not in fact prohibit all forms of race consciousness in private and governmental decision-making. Entities are permitted to take account of race in order to design fair procedures so long as they do not use racial classifications to determine the outcome of individual decisions. As a result, many, although not all, race-conscious model-building strategies do not amount to disparate treatment in the first place, and therefore do not require special legal justification.

This observation matters because recognizing that many de-biasing strategies are non-discriminatory not only lowers the legal risk for designers exploring these strategies, it lowers the temperature as well. The rhetoric surrounding affirmative action suggests that special justification is needed because these programs harm others. However, when race-conscious strategies work to remove unfair bias from these systems, no one is unfairly harmed. Algorithms reflect the myriad choices of their creators, and not some objective, underlying truth about who deserves what. Understanding the contingency of this process undermines claims that anyone is legally entitled to maintain a pre-existing system of advantage.

To be clear, by arguing that some forms of race-aware model-building should not be considered disparate treatment, I am not endorsing the adoption of any particular strategy. Others have argued that fairness-constrained strategies may exact a cost in terms of accuracy,[239] or end up harming the groups they are intended to protect.[240] Whether or when these strategies should be pursued are difficult questions and answering them requires close attention to things like the structure of the underlying data, the social context, and the consequences of predictions. The point here is this: taking race into account when building a model does not make it presumptively unlawful. Courts should refrain from preemptively stepping in and taking certain options off the table; instead, the choice among competing strategies is more appropriately left to vigorous debate among policy-makers and the public.

[239] *See, e.g.*, Berk et al., *supra* note 39, at 5 (arguing that "challenging tradeoffs are required between different kinds of fairness and between fairness and accuracy").

[240] *See, e.g.*, Lipton et al., *supra* note 3; Mayson, *supra* note 8.