Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action

(forthcoming 110 Calif. L. Rev. --- (2022))

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 blacks, have they then engaged in discrimination against whites? Some scholars assume so, and seek to justify those practices as valid forms of affirmative action. This Article argues that they have started the analysis in the wrong place. Rather than assuming that disparate treatment has occurred, we should first ask whether 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 disparate treatment against whites. Only if a particular strategy is found to be disparate treatment is it necessary to consider whether it is justifiable under affirmative action doctrine. This difference in approach matters, because affirmative action programs bear a heavy legal burden of justification. In addition, treating all race-aware algorithms as a form of disparate treatment 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.

* 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, Eugene Vorobeychik, Andrew Estronell, Sanmay Das, Patrick Fowler, C.J. Ho, Brendan Juba, Manish Raghavan, Solon Barocas, Andrew Selbst, Deborah Widiss, Deborah Hellman, and participants at the 2020 Colloquium on Scholarship in Employment and Labor Law, the 2021 Privacy Law Scholars Conference and Washington University School of Law faculty workshops for many helpful conversations about the issues discussed in this article.

Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action

I. Introduction

It is now widely recognized that algorithms can discriminate. As reliance on these tools to make decisions about people increases, concerns are growing 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. While legal scholars have focused on whether or how existing laws apply to these new tools,[3] data scientists and machine learning experts are working to devise technical solutions to prevent discrimination.[4] A

---

[1] *See, e.g.,* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Kristin Johnson et al., *Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation*, 88 FORDHAM L. REV. 499 (2019); Rashida Richardson et al., *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); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017); Crystal S. Yang & Will Dobbie, *Equal Protection under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291 (2020).

[2] Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMMUNICATIONS OF THE ACM 44 (2013); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH 1 (2018); Muhammad Ali et al., *Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes*, 3 PROCEEDINGS OF THE ACM ON HUMAN-COMPUTER INTERACTION 1 (2019); Amit Datta et al., *Discrimination in Online Advertising A Multidisciplinary Inquiry*, 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH 1 (2018); Julia Angwin et al., *Machine Bias*, PROPUBLICA (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=Gg58888u2U5db3W3CsuKrD0LD_VQJReQ.

[3] *See, e.g.,* Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions Essay*, 89 WASH. L. REV. 1 (2014); Barocas & Selbst, *supra* note 1; Kim, *supra* note 1; James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination Comment*, 7 CALIF. L. REV. CIRCUIT 164 (2016); Michael Selmi, *Algorithms, Discrimination and the Law* (forthcoming 2021); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395 (2018).

[4] For a small sampling of work in this area, *see, e.g.*, Irene Chen et al., *Why Is My Classifier Discriminatory?*, ARXIV:1805.12002 [CS, STAT] (Dec. 2018); Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV:1808.00023 [CS] (Aug. 2018); Cynthia Dwork et al., *Fairness through Awareness*, Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12 214 (ACM Press 2012); Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, ARXIV:1610.02413 [CS] (Oct. 2016); Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification without Discrimination*, 33 KNOWLEDGE AND INFORMATION SYSTEMS 1 (Oct. 2012); Toshihiro Kamishima et al., *Fairness-Aware Learning through Regularization Approach*, 2011 IEEE 11th International Conference on Data Mining Workshops 643 (Dec. 2011); Michael Kearns et al., *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*, ARXIV:1711.05144 [CS] (Dec. 2018); Zachary C. Lipton et al.,

robust literature proposes competing methods for 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.[5] Not only can biased outcomes still occur, but discarding demographic information makes bias harder to detect,[6] and, in some cases, could make it worse.[7]

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 anti-discrimination law.[8] The concern is that by taking race into account, these efforts will themselves be considered a form of intentional discrimination forbidden by law.[9] To put it concretely, if model-builders take race into account to prevent racial bias against blacks, have they then engaged in discrimination against whites?[10] What strategies can they employ to reduce discriminatory impacts without running afoul of the law?

Some researchers have assumed that the law prohibits any use of race. If true, many of the de-biasing strategies developed by computer scientists would be doomed to practical irrelevance. More recently, several scholars have sought to defend race-aware algorithms as valid forms of affirmative action. Jason Bent, for example, concludes that such strategies constitute disparate treatment, but may nevertheless be permissible if the legal requirements for justifying an affirmative

---

Does Mitigating ML's Impact Disparity Require Treatment Disparity?, ARXIV:1711.07076 [CS, STAT] (Jan. 2019); Joshua R. Loftus et al., *Causal Reasoning for Algorithmic Fairness*, ARXIV:1805.05859 [CS] (May 2018); Jialu Wang et al., *Fair Classification with Group-Dependent Label Noise*, PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 526 (Mar. 2021); Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment*, Proceedings of the 26th International Conference on World Wide Web 1171 (International World Wide Web Conferences Steering Committee Apr. 2017).

[5] *See, e.g.,* Corbett-Davies & Goel, *supra* note 4; Dwork et al., *supra* note 4; Hardt et al., *supra* note 4; Kamishima et al., *supra* note 4; Loftus et al., *supra* note 4; Yang & Dobbie, *supra* note 1; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

[6] *See, e.g.,* Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS AND PROCEEDINGS 22 (May 2018).

[7] Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019).

[8] *See, e.g.,* Corbett-Davies & Goel, *supra* note 4; Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389 (2019).

[9] *See* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1131 (2019); Sandra G. Mayson, *Bias in, Bias Out*, 128 YALE L. J. 2218, 2230, 2262–63 (2019).

[10] Throughout this article, I use hypotheticals involving measures taken to reduce bias against blacks 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 or race or ethnic bias.

action plan are satisfied.[11] Similarly, Daniel Ho and Alice Xiang argue that models that consider race trigger strict scrutiny under the Equal Protection Clause, but in some instances may pass constitutional muster if they are narrowly tailored to meet a compelling interest.[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.[13]

This Article argues that these scholars have started the analysis in the wrong place. Rather than assuming that any race-aware algorithm is a form of affirmative action that requires special justification, 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 in discussions about affirmative action which sometimes presume that colorblindness is legally required. In fact, as numerous scholars have pointed out, the law does not categorically prohibit race-consciousness.[14] Both private and government decision-makers routinely use information about race in ways that trigger no particular legal concern. Practices such as the collection of demographic information or the use of racial characteristics in suspect profiles are so commonplace that they are rarely remarked upon, let alone subject to legal challenge.[15] And courts have found some race-conscious actions, like an employer's efforts to diversify the applicant pool by expanding recruitment, do not constitute discrimination and are therefore legally permissible.[16] What triggers the special scrutiny imposed on affirmative action plans is not mere race awareness, but the specific ways race is used in the decision process.

The question of when race-conscious action requires justification is framed somewhat differently depending upon the source of law. In the statutory context, affirmative action plans require legal justification when they result in disparate

---

[11] Jason R Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 51 (2020).

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

[13] *See, e.g.,* Anupam Chander, *The Racist Algorithm*, 115 MICH. L. REV. 1023, 1041–41 (2017) (using "affirmative action" in its broadest sense to include any proactive practices to correct deficiencies in equality of opportunity); Mayson, *supra* note 9 (using the term algorithmic affirmative action to describe and asses the normative desirability of different strategies without considering thier legality).

[14] *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 (2016); Justin Driver, *Recognizing Race*, 112 COLUM. L. REV. 404 (2012); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811 (2020) ("the law'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 (2003).

[15] *See* Part IV.B., *infra*.

[16] *See* Part IV.A., *infra*.

treatment. White plaintiffs challenging such plans allege that by considering racial equity, the decision-maker took an adverse action against because of their race. In constitutional law, the focus is on racial classifications. Government decisions that entail racial classifications are unlawful unless they meet the requirements of strict scrutiny.

Efforts to redress racial inequities, however, do not always 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 rather than forms of affirmative action.[17]

This point is likely obscured by the tendency to assume that algorithms are fixed in nature, rather than recognizing them as contingent. In popular and legal discourse, the algorithm is often 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 at 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, the multi-step, iterative process of model building means that there are multiple points and multiple ways in which race might be taken into account in an effort to make a model less biased. And second, 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.

---

[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 11, at 823–24.

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

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

The first point highlights the myriad of possible methods for de-biasing algorithms. These strategies differ widely in approach, and when and how each takes race into account is critically important for judging its legality. This Article argues that some strategies—for example, those that address bias caused by data problems—do not amount to disparate treatment or involve racial classifications at all. For these strategies, arguments about whether they meet the requirements for a valid affirmative action plan are beside the point, because they do not constitute discrimination in the first place.

Closer questions arise when developers use race to shape the model's parameters or as a feature to predict outcomes. Race then appears to play a more direct role in determining who is selected. Even with these strategies, however, the analysis is more complicated than initially appears, and much depends on the details. In certain complex, feature-rich models, for example, information about race might influence predictions on the margins, but play a quite different role than in cases where race is used in a mechanical way to systematically favoring one racial group over another.

The causal question is also complicated because no "correct" model exists prior to consideration of race. In the absence of a clear baseline to serve as the counterfactual, the precise causal connection between considering race during the model-building process and the outcome of an individual decision downstream is quite uncertain.

In contrast to scholars who defend race-aware algorithms as lawful forms of affirmative action, I contend that some such strategies do not constitute discrimination in the first place. The difference between these two approaches is not just semantic. Defending a strategy as justifiable affirmative action differs significantly from recognizing it as non-discriminatory, both doctrinally and conceptually. From a doctrinal perspective, casting the question as one of affirmative action creates a heavy burden of justification, making a race-aware model presumptively unlawful unless a demanding legal standard is met. Even if the case can be made, as a practical matter, this additional burden may discourage developers from voluntarily trying to identify and address sources of bias.

On a conceptual level, the difference between nondiscrimination and affirmative action also matters. Characterizing race-aware algorithms as affirmative action activates a set of assumptions and surrounding rhetoric that are unhelpful and misleading. It reinforces the false notion that any steps taken to reduce bias or level the playing field for disadvantaged groups inherently harms whites and therefore requires special justification. It also plays into common misconceptions, suggesting that algorithms are neutral and objective tools that precisely measure merit or desert, rather than properly understanding them as 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. Affirmative action normally indicates efforts undertaken to increase access by blacks and other disadvantaged groups to resources and opportunities they have been excluded from in the past. In the criminal enforcement context, however, blacks have not been excluded, but instead disproportionately targeted by police and prosecutors. They are *over*-represented, not under-represented in a system that entails, not opportunity, but the risk of punitive sanctions and a cascading set of damaging collateral consequences. It makes little sense to judge efforts to address racially biased algorithms in this context as if they were "affirmative action" plans.

Before developing these arguments in detail, a few 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, the law in the U.S. is deeply and particularly suspicious of the use of race in decision-making.[20] 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 a bit of a misnomer. Computers do not have awareness or consciousness the way humans do, nor do they act with intentionality in any sense relevant to anti-discrimination law.[21] 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.

Finally, the term "bias" encompasses distinct, but related, meanings. A model or a process can be biased in the sense that it produces a skewed outcome, as when blacks are disproportionately screened out of an opportunity relative to whites. "Bias" in this sense describes any process that produces a disparate impact regardless of the cause. A predictive algorithm might produce a skewed outcome

[20] For example, under the Equal Protection Clause, racial classifications are subject to strict scrutiny, while sex classifications face a less demanding intermediate level of scrutiny. Classifications based on age and disability are not subject to any heightened level of review.

[21] Huq, *supra* note 9, at 1089.

for different reasons, and depending upon the cause, the model might or might not be considered normatively unfair or legally impermissible.

One cause is biased estimates, which occur when problems in the data or the construction of the model skews predictions in a way that makes the results inaccurate.[22] When biased estimates coincide with systematic racial bias, the effects are not morally or legally defensible. Whether other sources of observed bias are problematic is more contested. For example, a model may predict a higher risk of loan default for blacks because they in fact earn less money due to discrimination in the labor market. One might debate whether it is fair for the bank to rely on predictions which are accurate, but for reasons shaped by others' discriminatory practices. In many other cases, the reason a predictive model produces biased outcomes is uncertain, making it even more difficult to judge whether the results are normatively or legally defensible or not.

In this article, I do not address normative questions regarding which features are acceptable to rely on even if they have a racial impact.[23] 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 source. 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. On occasion, I refer to "discriminatory bias" to indicate a type of bias that (I believe) is uncontroversially unfair—such as statistical biases that result from unrepresentative or inaccurate data, or that reflect human prejudices.

This Article proceeds as follows. Part II briefly canvasses the evidence of algorithmic bias and the technical responses that have developed in response. Part III 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 IV, I analyze existing anti-discrimination law, focusing first on Title VII as an example of statutory prohibitions and then on constitutional doctrine developed under the Equal Protection Clause. This 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 V applies these insights to a handful of examples involving algorithmic de-biasing strategies, arguing that some such efforts do not constitute disparate treatment at all, while others appear to be legally impermissible. In

[22] *See, e.g,* Kim, *supra* note 1, at 886-88; Deborah Hellman, *Big Data and Compounding Injustice* (forthcoming J. Moral Phil. 2021).

[23] Answering that question is important for determining when an actor should be liable under a disparate impact theory, but that is not the focus of this article.

between lies a gray area of legal uncertainty, but where strong arguments can be made that at least some uses of race—even at prediction time—do not constitute disparate treatment or racial classifications. In Part VI, 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 "affirmative action" which is 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 analysis herein.

II. Algorithmic Bias and Technical Responses

A growing literature documents how predictive algorithms that are used to make socially consequential decisions can systematically disadvantage subordinated groups. Studies have shown that recommender systems deliver job ads to online users in ways that coincide with racial and gender stereotypes,[24] or suggest that people with African-American associated names have criminal records when they do not.[25] A recruitment algorithm systematically downgraded women candidates for computer programming positions because it was trained using a dataset composed primarily of men.[26] A selection algorithm disfavored women and racial minorities for medical school admission based on past discriminatory practices.[27] Facial recognition systems made far more mistakes in identifying people with darker skin.[28] A tool allocating health care directed greater resources to white patients than black patients with the same level of need.[29] An algorithm used to inform bail decisions over-predicted recidivism risks for blacks as compared with whites charged with crime.[30]

Documented instances of algorithmic bias are troubling, but disparate outcomes across groups can occur for many reasons. They might reflect actual differences between groups that are relevant to the decision at hand in ways that are legitimate to consider. In other circumstances, they may result from implicit biases on the part of developers, or reflect problems with the data used for training. The definition of the target variable may involve value choices that implicitly favor one group over another, as, for example, selecting for aggressiveness as a measure of leadership

---

[24] Piotr Sapiezynski et al., *Algorithms that "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, ARXIV:1912.07579 [CS] (2019); 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.

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

[26] 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.

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

[28] Buolamwini & Gebru, *supra* note 2.

[29] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

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

rather than collaboration. Sometimes the data capture human biases by relying on subjective human judgments to code the attributes used for prediction.[31] The dataset used to train an algorithm may also be unrepresentative of the relevant population, resulting in skewed outcomes or inaccurate predictions for the underrepresented subgroup.[32] Additional data problems such as poor quality or missing information can also cause biased predictions. And even highly accurate data can cause biased predictions if they are simply reproducing historical patterns of disadvantage and segregation.[33]

The 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 examined whether or how existing anti-discrimination laws apply to automated decision tools,[34] computer scientists have focused on developing methods to ensure that predictive models are fair.[35] They have proposed a wide range of strategies, from compensating for data problems to designing models that comply with a specified notion of fairness. 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,[36] but these definitions are often incompatible, such that it is not possible to simultaneously satisfy them all.[37]

There is, however, one point on which there is consensus. Merely blinding an algorithm will not prevent bias.[38] 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, due to patterns of residential segregation, zipcode 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 zipcode to make predictions. As a result, some have argued that both race and all proxies for it should be eliminated from predictive models.[39] 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

---

[31] *See, e.g.*, Barocas & Selbst, *supra* note 1, at 680.

[32] *See, e.g.,* Kroll et al., *supra* note 5.

[33] *See, e.g.,* Mayson, *supra* note 9.

[34] *See, supra*, note 3.

[35] *See, supra,* note 4.

[36] *See, e.g.*, Huq, *supra* note 9, at 1115 (referencing 21 different definitions of fairness); Narayan (2018) (cataloging definitions of fairness); Mayson, *supra* note 9.

[37] *See* Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIOLOGICAL METHODS & RESEARCH 3 (Feb. 2021); Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV:1609.05807 [CS, STAT] (Nov. 2016).

[38] *See, e.g.,* Corbett-Davies & Goel, *supra* note 4; Dwork et al., *supra* note 4; Hardt et al., *supra* note 4; Kamishima et al., *supra* note 4; Loftus et al., *supra* note 4; Yang & Dobbie, *supra* note 1; Kroll et al., *supra* note 5.

[39] *See, e.g.,* Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014).

situations to remove its correlates and still have a meaningful model.[40] 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.

III. Model Building

In *Employing AI*, Charles Sullivan asks the reader to engage in a thought experiment.[41] "Imagine," he writes, that "a company . . . effectively delegates all its hiring decisions to a computer. It gives the computer only one instruction: 'Pick good employees.'"[42] 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 "goes rogue,"[43] 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. In the popular imagination an algorithm is a well-defined tool such that once a goal has been identified—for example, "hire good employees"—there is a single authoritative model that solves the 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."[44] 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."[45] In fact, however, the algorithm or model results from a process involving multiple steps, each entailing choices about how to build the model. Importantly, there is no inevitable destination, no uniquely definitive model that represents the "correct" solution to

---

[40] Yang & Dobbie, *supra* note 1 (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").

[41] Sullivan, *supra* note 3.

[42] *Id.* at 395.

[43] *Id.* at 402.

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

[45] *Id.* at 668.

the problem. Instead, each choice along the way involves weighing tradeoffs and exercising judgment, and each is consequential for the ultimate design of the model.

Lehr and Ohm catalog the multiple steps involved in building machine learning models. First, there is the question of problem definition. An employer cannot simply tell the computer to "pick good employees." Someone—some human—must decide what it means to be a "good" employee. The designer must choose whether to define a "good employee" as someone who is highly productive, or will stay on the job for a long time, or is creative, or has strong interpersonal skills. And this choice will affect what the model looks like and which applicants it picks as good prospects.

These types of questions arise in other contexts as well.[46] 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 target of prediction cannot be directly measured, and someone must decide what observable metric best approximates it.

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 population to be predicted. 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 that contains highly granular information, but includes few observations of under-represented groups.

After selecting a dataset, more 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 that may provide valuable information, or may represent exceptional cases that will distort predictions if they are included in the analysis. Once these decisions have been made, the designer must select a subsample of the data to "train" the model. From exposure to these 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 still decide whether to split the data 50/50 or in some other proportion, a decision that turns on matters 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

[46] *See, e.g.,* Eaglin, *supra* note 1.

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—logistic regression, random forest, neural networks, etc—that use different technical strategies for optimizing the prediction problem. Which model is 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 the model and then just set it loose in the world, 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."[47] 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."[48]

Even after deployment, a model 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 is potentially consequential in determining the final version of the model and will influence the actual predictions it makes when deployed. In sum, there is no single solution to a prediction problem, but a multitude of possible models. The decision *which* model to adopt must be made by humans because the tradeoffs entail value choices and discretionary judgments.

---

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

[48] *Id.* at 669.

From these observations follow two important implications relevant to the question 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, Estornell et al. show that the 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.[49] Similarly, Black and Fredrikson demonstrate that the inclusion or removal of a *single* person in the model's training data can change the outcomes for some other individuals under the resulting model, and this effect occurs with "surprising frequency."[50] If these seemly minor changes—the random draw of a training dataset or the failure to include one observation—can alter the outcome for some individuals, then other choices, such as the structure of the model or how data are labeled, are likely to have even more significant impacts on individual outcomes. More fundamental decisions, like defining the target of prediction, may fundamentally 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 consideration of race and therefore has some settled expectation that was disrupted. Part V considers the legal relevance of these observations in greater depth.

The second implication that follows from understanding the model building process is the recognition that, with so many decision points, there are many different ways in which unfair bias can creep into a model. Conversely, there are also multiple points at which a designer might make choices to try to remove or reduce racial bias. Depending upon the strategies pursued, these decisions will have different impacts on the final model and the outcomes it predicts. The legality of race-conscious de-biasing efforts should depend upon the type of intervention chosen.

---

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

[50] Emily Black & Matt Fredrikson, *Leave-One-out Unfairness* 285 (ACM Mar. 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.* at 285.

IV. The Lawfulness of Race Conscious Decision-Making

Because strategies for building fair algorithms require explicit consideration of race, some researchers question whether they are legal under anti-discrimination law.[51] 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 account in order to prevent an algorithm from being biased against blacks, have they then engaged in discrimination against whites? 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.

This Part explains existing anti-discrimination doctrine before considering how it applies to algorithmic tools. Although the Supreme Court's race jurisprudence has been subject to extensive criticism,[52] my purpose here is not to argue with its past decisions. Similarly, I put aside until Part ---, consideration of how the changed composition of the Court may affect doctrine going forward. Instead, this Part analyzes the law as it currently exists, taking established doctrine at face value and the Justices at their word when they explain their reasoning. This examination suggests that ample room exists for certain types of race-conscious efforts to de-bias algorithms.

Different statutes prohibit discrimination when lending money,[53] hiring workers,[54] selling or renting a home,[55] entering into a contract,[56] or providing educational opportunities.[57] The Constitution also forbids race discrimination, but its prohibitions only apply to state actors. Exploring the nuances of each potentially relevant law is not possible here. Instead, section A. below analyzes in-depth one anti-discrimination statute, Title VII of the Civil Rights Act of 1964. Title VII prohibits discrimination in employment and is a useful example because the case

---

[51] Bent, *supra* note 11; Corbett-Davies & Goel, *supra* note 4; Cofone, *supra* note 8; Ho & Xiang, *supra* note 12.

[52] Critical race scholars argue that an insistence on colorblindness overlooks, and therefore enables and reinforces, the many ways in which societal institutions systematically impose disadvantage on the basis of race. *See, e.g.,* Devon W. Carbado, *Footnote 43: Recovering Justice Powell's Anti-Preference Framing of Affirmative Action*, 53 U.C. DAVIS L. REV. 1117 (2019); Kimberlé W. Crenshaw, *"Framing Affirmative Action"*, 105 Mich. L. Rev. First Impressions 123 (2006); Cheryl I. Harris, *Whiteness as Property*, 106 Harv. L. Rev. 1707 (1993). Others, such as Aziz Huq, content that the Court's existing race jurisprudence is particularly unsuited to problems of discrimination in algorithmic decision-making. Huq, *supra* note 9, at 1101 (arguing that current equal protection doctrine is a poor fit because it poses questions not relevant to algorithmic decision-making).

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

[54] 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)).

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

[56] 42 U.S.C. § 1981.

[57] 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)).

law interpreting it is particularly well-developed. Section B. then examines the prohibition on race discrimination under the Equal Protection Clause of the Constitution.[58]

A. Statutory Law

1. The Title VII Framework

Title VII prohibits discrimination in employment on the basis of race, sex, and other protected characteristics.[59] 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. Disparate impact theory does not require proof of intent, but instead targets facially neutral practices that have the effect of disproportionately harming members of historically disadvantaged groups.

In order 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.[60] Critical to proving disparate treatment is establishing causation, and there are two routes for doing so: showing that the protected characteristics was the "motivating factor" for the adverse decision, and demonstrating that it was a "but-for cause."[61] 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.[62] The second route requires a discrimination plaintiff to show "but-for" causation, a traditional standard of proof imported from

---

[58] U.S. Const. amend. XIV, § 1.

[59] 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).

[60] 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* (forthcoming VIRGINIA LAW REVIEW 2021); Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discrimination Intent*, 109 COLUM. L. REV. 1357 (2009).

[61] Bostock v. Clayton County, Georgia,140 S.Ct. 1731, 1739-40 (2020).

[62] 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, *Southwestern Medical Center v. Nassar*, 570 U.S. 338 (2013), or age discrimination claims, *Gross v. FBL Financial Services, Inc.*, 557 U.S. 167 (2009), which must be proven under the but-for causation standard.

tort law.[63] This standard requires a plaintiff to show that the protected characteristic actually made a difference in the outcome. It asks whether an adverse outcome for a worker would have come out differently if the protected characteristic had not been taken into account.

Disparate impact has a different focus and different standards of proof. It is not concerned with employer intent or motive, but instead focuses on the discriminatory effects of facially neutral practices.[64] 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 blacks than whites from a particular job.[65] 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."[66]

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,[67] 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

At the time Title VII was enacted, many employers had racially segregated workforces and confronted significant risks of legal liability. 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. In this environment, employers had strong incentives to scrutinize their own practices for discrimination

---

[63] Bostock, 140 S.Ct. at 1740. The "but for" causation standard is generally considered more demanding than the motivating factor test, although scholars disagree on the implications of requiring but-for causation in employment discrimination cases. *Compare, e.g.,* Sandra F Sperino, *Let's Pretend Discrimination Is a Tort*, 75 OHIO ST. L. J. 1107 (2014) (criticizing the use of but-for causation in discrimination cases) *with* Eyer, *supra* note 60, (arguing that an expansive understanding of but-for causation is "potentially radical in its legal effects").

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

[65] 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. 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 anti-discrimination law.

[66] 42 U.S.C. § 2000e-2(k).

[67] *See* note 3, *supra*.

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 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.[68] 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*.[69] In light of a history of near-total exclusion of blacks from craftwork positions,[70] the employer, Kaiser Aluminum & Chemical Corp., and the steelworkers union created a program to train its unskilled workers for skilled craft positions. Applicants were accepted in the program based on seniority, with the caveat that at least 50% 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 blacks in the local labor force (39%).[71] 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, it appeared that Kaiser had engaged in disparate treatment, unless its actions were justified. 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,"[72] 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 and catalyst to cause 'employers and unions to self-examine and to self-evaluate their employment practices and to endeavor to eliminate" the last vestiges of the country's history of racial segregation.[73]

---

[68] *See, e.g.,* Firefighters v. City of Cleveland, 478 U.S. 501 (1986); United States v. Paradise, 480 U.S. 149, 166 (1987).

[69] 443 U.S. 193 (1979).

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

[71] *Id.* at 198.

[72] *Id.* at 202.

[73] *Id.* at 204.

Although some uses of race to promote equality might violate anti-discrimination law, the Court concluded that Kaiser's affirmative action plan "falls on the permissible side of the line,"[74] pointing to several relevant considerations. First, its purpose mirrored that of Title VII—"to break down old patterns of racial segregation and hierarchy."[75] In addition, the plan "does not unnecessarily trammel the interests of white employees."[76] 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.[77] Finally, the plan was temporary, and not intended to maintain a permanent racial balance in the workforce.[78]

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*,[79] 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.[80]Applying the framework it had established in *Weber*, the court rejected Johnson's claim and concluded that the affirmative action plan was permissible.[81]

Weber and *Johnson* provide a legal framework for assessing voluntary affirmative action plans; however, not everything an employer does that might be labeled "affirmative action" triggers this analysis. The term is not well-defined and has been applied to a broad range of activities aimed at redressing racial inequality 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.[82] 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."

---

[74] *Id.* at 208.

[75] *Id.*

[76] *Id.*

[77] *Id.*

[78] *Id.*

[79] 480 U.S. 616 (1987).

[80] *Id.* at 621. Both Johnson and Joyce were among the top-ranked applicants, and the Agency Director considered numerous factors, including the affirmative action plan, before deciding to promote Joyce.

[81] 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. 480 U.S. 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.

[82] 443 U.S. at 203.

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. Brian Weber alleged he was discriminated against because the training program required that half the workers admitted be black. To defend against his claim of disparate treatment, the employer had to demonstrate the validity of its affirmative action plan. Unlike in *Weber*, the plan in *Johnson* 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.[83] In both cases, then, the Court scrutinized the employers' affirmative action plans on the premise that those plans had caused the employers to take adverse actions motivated by race or sex.[84]

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.[85] 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.[86] 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.[87] In the

[83] One might question this conclusion given the facts. Although the panel that initially interviewed the candidates had rated Johnson a 75 and Joyce a 73, the employer was not required to promote the person with the highest score—it was permitted to select any of the 7 applicants rated qualified—nor was it clear that the difference between a 75 and 73 was meaningful. The Director testified that he considered numerous factors, including the severe underrepresentation of women in the relevant job category, in reaching a decision, and the affirmative action plan did not require any particular number of female hires.

[84] After the Civil Rights Act of 1991 was passed, some suggested that the provision it added making unlawful any adverse employment decision "motivated by" a protected characteristic, 703(m), 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 Act of 1991, Pub. L. No. 102-166, § 116, 105 Stat. 1071 (1991) (codified at 42 U.S.C. § 1981 note). In any case, the "motivating factor" provision does not appear to have affected how courts treat affirmative action or diversity plans.

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

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

[87] Coppinger v. Wal-Mart Stores, Inc., 2009 U.S. Dist. LEXIS 91120 (N.D. Fla. Sept. 30, 2009); Jones v. Bernanke, 493 F. Supp. 2d 18 (D.D.C. 2007); Keating v. Paulson, 2007 U.S. Dist. LEXIS 80516 (N.D. Ill. Oct. 25, 2007); Martin v. City of Atlanta, 579 Fed. Appx. 819 (11th Cir. 2014); Plumb v. Potter, 212 Fed. Appx. 472 (6th Cir. 2007). 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 ABA J. OF LAB. & EMP. LAW 139, 155 (2018)

absence of a clear connection to the 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.[88]

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. For example, in *Duffy v. Wolle*,[89] 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.[90]

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."[91] 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' qualification. 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 claim."[92]

Duffy and the cases that follow it[93] indicate that race-conscious actions taken to remove unfair policies or diversify the workforce do not necessarily 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

---

("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.").

[88] *See, e.g.,*Mlynczak v. Bodman, 442 F.3d 1050 (7th Cir. 2006).

[89] 123 F.3d 1026 (8th Cir. 1997).

[90] *Id.* at 1038–39.

[91] *Id.* at 1030.

[92] *Id.* 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 advertise job openings widely harmed him because it resulted in an "influx of blacks" 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.

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

adopt changes in order 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 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.[94] *Ricci v. DeStefano*[95] is not to the contrary. In *Ricci*, the Supreme Court majority held that the City of New Haven engaged in disparate treatment when it discarded a promotional examination for firefighters because it would have produce a nearly all-white promotional class.[96] The City defended its action on the grounds that it feared a disparate impact suit by minority firefighters, but a majority of the Supreme Court held that discarding the results would only be permissible if the City had "a strong basis in evidence" that the test violated disparate impact law—a showing that the City could not make.[97]

Some commenters have suggested that *Ricci* bars employers from proactively changing their practices to remove disparate impact or promoting diversity goals unless they can meet the "strong basis in evidence" test.[98] As I have argued elsewhere, this conclusion rests on a misreading of *Ricci*.[99] The Court's announcement 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, 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.[100]

---

[94] Griggs v. Duke Power Co., 401 U.S. 424, 430–31 (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.").

[95] 557 U.S. 557 (2009).

[96] *Id.* at 563. The dissent disagreed that the City's actions constituted disparate treatment, arguing that the plaintiffs had no vested right to promotion and that substantial evidence existed that the test was seriously flawed and so the results should not be relied on. *Id*. at 608-09, 619.

[97] *Id.* at 563.

[98] *See* Barocas & Selbst, *supra* note 1; Kroll et al., *supra* note 5.

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

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

Employers, however, are free to make *prospective* changes to practices they discover are biased or discriminatory and to take race into account when doing so.[101] Future applicants have no fixed entitlement to an employer's past hiring or promotion criteria, and thus, 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. And in seeking to create fairer processes, an employer may need to take race into account. The *Ricci* court acknowledged 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 race,"[102] and it 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 minorities sat on every panel assessing the oral part of the exam.[103] Thus, while the Court found the City's decision to discard the exam results to be disparate treatment under the circumstances in *Ricci*, that 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 in designing its 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 time and effect to be causally connected to a specific personnel decision down the road.

B. The Equal Protection Clause

---

[101] In *Maraschiello v. City of Buffalo Police Dep't*, 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. City of Mt. Vernon, 707 F. Supp. 2d 449 (S.D.N.Y. 2010).

[102] Ricci, 557 U.S. at 585.

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

The Equal Protection Clause of the Constitution also forbids discrimination, although it differs from statutory prohibitions in a number of ways. The anti-discrimination 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 state actors, but applies to a broad range of government 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 in the last section, the initial inquiry under Title VII is whether a particular employer policy or action constitutes disparate treatment; if so, it must be justified as a valid affirmative action plan under *Weber*. Under the Equal Protection Clause, government action that relies on racial classifications must be justified under the standard of strict scrutiny. The key first step in the analysis is showing that a racial . . . .

In the mid-twentieth century, spurred by the civil rights movement and growing attention to significant racial gaps in access to opportunities and measures of well-being, government actors took steps to redress racial inequities in areas like education and public contracting. Characterized as "affirmative action," these efforts were challenged by white plaintiffs who alleged that they were harmed because the government used racial classifications to make decisions. 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."[104] In the Court's view, it does not matter if the classification is intended to achieve a benign purpose,[105] such as compensating for existing disadvantages based on race. Remedying "societal discrimination" is not a sufficient justification,[106] although the Court has approved race-based remedies for a government actor's own past discrimination.[107] Second, the appropriate level of scrutiny for examining racial classifications is "strict scrutiny."[108] Strict scrutiny, the Court has instructed, requires that the racial classification "furthers a compelling government interest" and that the means chosen are "narrowly tailored" to meet that interest.[109] A racial classification that does not meet that exacting standard is unconstitutional.

---

[104] Richmond v. J.A. Croson Co., 488 U.S. 469, 493–94 (1989).

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

[106] Wygant v. Jackson Bd. of Education, 476 U.S. 267, 272 (1986).

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

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

[109] *Id.* at 220.

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."[110]

One obvious exception is that strict scrutiny is not inevitably fatal. In *Grutter v. Bollinger*,[111] 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.[112] The Court in *Fisher v. University of Texas at Austin*[113] similarly approved that University's admissions policies, which took race into consideration as one factor among many in selecting its student body.[114]

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. 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, *Bakke*,[115] the Court considered a constitutional challenge to state university's admissions policy which set aside 16 out of 100 places in a medical school class for members of disadvantaged minority

---

[110] *See,* Hellman, *supra* note 14, at 819. *See also* Bagenstos, *supra* note 14; Driver, *supra* note 14; *Cf.* Kim Forde-Mazrui, *The Canary-Blind Constitution: Must Government Ignore Racial Inequality Race and Reform in Twenty-First Century America*, 79 LAW & CONTEMP. PROBS. 53 (2016).

[111] 539 U.S. 306 (2003).

[112] Key to its 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 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 337.

[113] 126 S. Ct. 2198 (2016).

[114] 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 2206. For the remaining seats, the University considered an Academic Index and a Personal Achievement Index (PAI). 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 2207.

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

groups. In *Croson*,[116] 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*[117] 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*[118] challenged a school board policy that the percentage of minority teachers laid off could not exceed their percentage employed by the district. Because minority teachers generally had less seniority, white teachers with greater seniority were laid off pursuant to the policy. And *Parents Involved*[119] considered school district policies that made school assignments by race in order 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.

In a variety of other situations, however, government acts in race-aware ways, apparently without triggering strict scrutiny.[120] 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.

These practices rarely provoke legal questions, let alone successful constitutional challenges.[121] In one case, plaintiffs sued to bar the collection of racial information in the U.S. Census, arguing that the questionnaire involved a racial classification and was subject to strict scrutiny.[122] 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."[123] Because the Census involved only the collection of information, it did not even trigger heightened scrutiny. As the court explained, the concerns plaintiffs

---

[116] Richmond v. J.A. Croson Co., 488 U.S. 469 (1989).

[117] Adarand Constructors v. Pena, 515 U.S. 200 (1995).

[118] Wygant v. Jackson Bd. of Education, 476 U.S. 267 (1986).

[119] Parents Involved in Cmty. School. v. Seattle Sch. Dist. No. 1, 551 U.S. 701 (2007).

[120] *See, e.g.,* Primus, *supra* note 14, at 505.

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

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

[123] *Id*. at 814.

raised about the type of information sought on the Census form was "one properly addressed by Congress, not by the courts."[124]

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.[125] 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, but so long as they do not use racial classifications to distribute or withhold benefits to individuals, they should not raise constitutional concerns.

Government actors also routinely act with an awareness of race when law enforcement uses suspect profiles.[126] 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.[127] If the Equal Protection Clause embodied a strict colorblindness theory, race-based subject descriptions should arguably trigger strict scrutiny,[128] but apparently they do not.[129]

There are other examples of race-aware government activity that do not appear to trigger constitutional concerns. For example, when placing children for 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 best interests of the child when making placement decisions.[130]

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.[131] On the other hand, if the state pursues other goals, the fact that it relied

---

[124] *Id*. at 815.

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

[126] *See, e.g.,* Hellman, *supra* note 14, at 859.

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

[128] Ralph Richard Banks, *Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse*, 48 UCLA L. REV. 1075 (2001).

[129] Huq, *supra* note 9, at 1096.

[130] R. Richard Banks, *The Color of Desire: Fulfilling Adoptive Parents' Racial Preferences through Discriminatory State Action*, 107 YALE L.J. 875 (1998).

[131] Miller v. Johnson, 515 U.S. 900, 916 (1995).

on race-based information, such as the knowledge that the most loyal Democratic voters are black voters, does not trigger equal protection concerns.[132] 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.

These examples, which fall outside the Court's affirmative action jurisprudence, illustrate 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,[133] 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."[134] 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.[135] 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 produce the sort of direct, real-world effects that raise constitutional concerns.[136] 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.[137] Second, she asserts that strict scrutiny applies when government makes generalizations *about* racial groups, but not generalizations that refer to race.[138] Suspect profiles do not rely on generalizations about a racial group,[139] and thus, even though they refer to racial characteristics, they are not suspect racial classifications triggering strict scrutiny.

---

[132] Hunt v. Cromartie, 526 U.S. 541 (1999).

[133] Driver, *supra* note 14.

[134] *Id.*

[135] 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. 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, but as examples of permissible race-conscious government action.

[136] Hellman, *supra* note 14, at 858.

[137] *Id.* at 862.

[138] *Id.* at 859.

[139] 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. Instead, their actions follow from a different type of generalization: that eye-witness descriptions are generally helpful in identifying perpetrators. 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

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."[140] 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."[141]

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 acknowledge 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 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 . . .*"[142]

Writing for the Court majority a few years later in *Texas Department of Housing and Community Affairs v. Inclusive Communities*,[143] 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.[144] The Court first held that the disparate impact theory of liability

---

therefore suspect, generalization. *See* Shirin Sinnar, *The Lost Story of* Iqbal, 105 GEO. L.J. 379, 419–21 (2017).

[140] Bagenstos, *supra* note 14, at 1119.

[141] *Id.* at 1117.

[142] *Id*. at 788-89.

[143] 576 U.S. 519 (2015).

[144] *See* Bagenstos, *supra* note 14, at 1127–30.

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."[145] 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."[146]

Similarly, in *Fisher* the Court appeared to have no concerns with the University of Texas's "Top Ten Percent" plan, which granted automatic admission to any student in the top 10% of a high school class in Texas.[147] After the Fifth Circuit's decision in *Hopwood* prohibited any consideration of race in admissions, 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.[148] The plan was effective in meeting that objective because many schools and neighborhoods in Texas are racially segregated.[149] As Justice Ginsburg pointed out, "race consciousness, not blindness" drove the University's Top Ten Percent plan.[150] The majority's apparent unconcern about the plan suggests that constitutional concerns are triggered not by mere race-awareness apart from a reliance on racial classifications.

In government contracting cases, the Justices have also acknowledged the legitimacy of state efforts to increase opportunities for disadvantaged racial groups.[151] 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.[152]

---

[145] *Id.* at 545.

[146] *Id.* 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 ensure revitalization of communities that have long suffered the harsh consequences of segregated housing patterns. 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.* at 545.

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

[148] *Id.*

[149] *Id.* at 335.

[150] *Id.*

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

[152] *Id*. at 526.

What appears to trigger strict scrutiny, then, is not the mere consideration of race or racial disparities by government, but the 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 opportunities to bid on federal contracts in a "sheltered market."[153] The enabling statute made the program available to businesses owned by "socially disadvantaged individuals" who are 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" but without presuming that any particular individuals could or could not show that they were eligible.[154] 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.[155]

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. A literal application of a colorblindness principle would throw into question enormous swaths of existing law.[156] The entire edifice of civil rights laws 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, of which Title VII is a part, and the Voting Rights Act of 1965 were enacted in response to pressing concerns about racial segregation and the exclusion of blacks from the mainstream of American economic, social and political life. Every state and a multitude of local governments have also passed laws recognizing the harms caused by racial discrimination and 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 in fact categorically forbade government from taking race into account in its decision-making, all of this statutory and decisional law would be suspect.

---

[153] Rothe Dev., Inc. v. United States DOD, 836 F.3d 57 (D.D.C. 2016).
[154] *Id.* at 64.
[155] *Id*. at 72.
[156] *See, e.g.,* Bagenstos, *supra* note 14.

The irony, of course, would be 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 newly freed blacks.

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,[157] examining the affirmative action cases suggests some critical factors. The programs subjected to strict scrutiny take a certain form—namely, they apply racial criteria to individuals in a rigidly mechanical way that consistently favors one racial group over another. The Court is particularly concerned that racial classifications that benefit disadvantaged groups operate as quotas, reserving a fixed number of slots for minority groups or aiming for a permanent racial balance.[158] Individual Justices have also expressed concerns that racial classifications are demeaning to individuals and will perpetuate hostilities and racial divisiveness.[159]

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 to policies designed to remove arbitrary barriers or to level the playing field without imposing quotas, or requiring particular outcomes. 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.

---

[157] *Id.* at 1119, 1142.

[158] *See, e.g.*, Regents of University of California v. Bakke, 438 U.S. 265, 289 (1978) (disapproving of racial quota in medical school admissions); Croson v. City of Richmond, 488 U.S. 469, 499 (1989) (amorphous claim of past discrimination "cannot justify the use of an unyielding racial quota"); Grutter v. Bollinger, 539 U.S. 306, 334 (2003) (a narrowly tailored program "cannot use a quota system"); United Steelworkers v. Weber, 443 U.S. 193, 208 (1979) (approving affirmative action plan because it was not intended to maintain racial balance).

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

* * * * *

Although the nuances of the doctrines 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. However, not all race-conscious practices are presumptively unlawful. The special legal scrutiny imposed on affirmative action plans only kicks in after a plaintiff has first shown discrimination has occurred. In the statutory context, this requires the white plaintiff to establish disparate treatment—that he suffered an adverse action causally connected to race. In the constitutional context, it is the use of racial classifications that triggers scrutiny. Once disparate treatment or a racial classification is established, a practice might be defended as a valid form of affirmative action or by meeting the requirements of strict scrutiny. However, the often overlooked point 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.

V. Race-Aware Algorithms

This Part takes the legal framework laid out in Part IV 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 anti-discrimination law. Under both statutory and constitutional law, racial realities may be taken into account in order 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 directly on point, existing precedent leaves room for explicitly considering the racial impact of predictive algorithms and exploring 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 assume that equal protection doctrine prohibits the use of algorithmic fairness techniques.[160] Jason Bent similarly concludes that deploying an algorithm that includes a race-aware fairness constraint constitutes disparate treatment under Title VII.[161] These scholars then focus on whether such strategies can be justified as valid forms of affirmative actions.

I believe these scholars start their analysis in the wrong place. *Before* asking whether race-conscious model building strategies can be justified under affirmative action doctrine, 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

[160] Ho & Xiang, *supra* note 12, at 134 (arguing that algorithmic fairness techniques "pose serious legal risks of violating equal protection").

[161] Bent, *supra* note 11, at 825.

strategies should depend upon when and how a particular approach takes account of race.[162] Only in those instances in which a strategy constitutes discrimination in the first place is it necessary to ask the further question whether it is justifiable under affirmative action doctrine.

Part VI *infra* discusses in greater detail why posing the questions in this order matters so much both doctrinally and conceptually. As explained there, recognizing that some race-conscious approaches are acceptable, non-discriminatory forms of *removing* unfairness 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 and these differences should matter legally. The computer science literature now encompasses a vast array of proposed strategies for de-biasing algorithms[163] and it is not possible to analyze them all. Instead, Part 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 B. takes a deeper dive into the causal question. As explained there, determining when taking race into account causes an adverse outcome based on race 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 which definition of fairness or what techniques are preferable. Those debates pose important policy questions, but are distinct from 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 a racial classification 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.

---

[162] *See* Hellman, *supra* note 14, at 484.

[163] *See* note 4, *supra*, and sources cited therein.

1. Dealing with Data Problems

One of the ways bias can creep into a predictive model is from 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. describe how several jurisdictions developed predictive policing tools during periods in which corrupt or racially discriminatory policing practices were documented.[164] If the data used to build the models reflect those troubling practices, the predictive outputs would reproduce and further reinforce those harms. Similarly, Altenburger and Ho found that algorithms used to target food safety inspections disproportionately burden Asian establishments when they rely on consumer complaints or online reviews, because those data reflect anti-Asian stereotypes about lack of cleanliness.[165] 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 argues that the "playing with the data" stages of model building offer numerous opportunities for reducing bias.[166] Developers could analyze the dataset for implicit biases before relying on it,[167] or oversample from an under-represented group.[168] They could collect additional data from certain groups,[169] 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.

---

[164] Richardson et al., *supra* note 1.

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

[166] Lehr & Ohm, *supra* note 18, at 657. *See also* Barocas & Selbst, *supra* note 1, at 717-19.

[167] 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.

[168] *See, e.g.,* Kamiran & Calders, *supra* note 4.

[169] 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 4.

Nevertheless, strategies like these that are aimed at addressing 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 whites 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 have no entitlement to be judged 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. If the employer does not hire the white candidate, it is not because it relied on race to make that particular decision, nor did it put a mechanical thumb on the scale intended to favor 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 III, 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 measureable target variable. The choice of the target variable can be highly consequential, both in terms of predictions about specific individuals and the overall distribution of outcomes across populations. As Passi and 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." Thus, paying attention to how a problem is formulated is an important tool for avoiding unnecessary racial inequities. Although it involves race-conscious decision-making, it clearly falls on the legally permissible side of the spectrum because it does not involve making decisions about individual based on race.

Obermeyer, et al. offer a good example of the critical role of problem formulation in avoiding racial bias.[170] 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, blacks consume less health care than

[170] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

whites 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 across demographic groups. Put differently, they ensure that demographic groups receive positive outcomes in proportion to their actual representation For example, if blacks are 20% of the relevant population, they should be positively classified 20% of the time, or within some specified range of that proportion (e.g. 17 – 23%). 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 would be 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 in order to ensure that the benefit is distributed equally across groups. Another strategy would transform the score of each individual based on her racial group, so that the distribution of positive predictions is proportional across different subsets of the population.[171] These types of strategies use information about race to achieve a proportional distribution of positive outcomes, but likely violate anti-discrimination 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.[172] 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 the prohibition 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—the fear that race will be used to impose quotas, or as a means of pursuing racial balancing. Thus, strategies that are aimed at achieving a particular numerical distribution of outcomes for its own sake will likely trigger close legal scrutiny.

4. Disparate Learning Processes (DLPs)

---

[171] For example, Feldman et al. outline a strategy that would separate the data by the protected attribute and then alter the values of other predictive factors for each group separately in order to reduce inter-group differences. Michael Feldman et al., *Certifying and Removing Disparate Impact*, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 259 (ACM Aug. 2015).

[172] 28 U.S.C. § 2000e-2(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.

Disparate learning processes (DLPs) are strategies that use racial information during training, but do not allow the model to access race when making predictions.[173] Harned and Wallach argue that DLPs offer the "just right" Goldilocks solution because they take race into account to de-bias algorithms, but do not run afoul of anti-discrimination law because race is not used to predict outcomes.[174] Cofone similarly argues that data pre-processing techniques that do not give an algorithm access to sensitive information at prediction time do not violate disparate treatment law.[175] A number of vendors who build applicant screening tools advertise that their model development process follows such a strategy,[176] likely in an effort to signal their compliance with anti-discrimination 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 of the availability of proxies for race in many datasets.[177] Others have argued that they are too costly in terms of reduced accuracy,[178] and that they can harm some members of the protected group.[179] 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 anti-discrimination law.

Drawing the line between race-awareness at training time versus prediction time has intuitive appeal because it maps onto formal notions of disparate treatment as involving 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 a reasonable first cut at the problem, the distinction between using race at training time and at prediction time should not necessarily be decisive of 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

---

[173] *See* Lipton et al., *supra* note 4, 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.*"). For examples, of DLPs, see, e.g., Kamishima et al., *supra* note 4; Kamiran & Calders, *supra* note 4; Muhammad Bilal Zafar et al., *Fairness Constraints: A Flexible Approach for Fair Classification*, AISTATS 1 (2017); Faisal Kamiran et al., *Discrimination Aware Decision Tree Learning*, 2010 IEEE International Conference on Data Mining 869 (Dec. 2010).

[174] 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).

[175] Cofone, *supra* note 8, at 1421.

[176] *See* Manish Raghavan et al., *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices*, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 469 (ACM Jan. 2020).

[177] *See, e.g.,* Dwork et al., *supra* note 4.

[178] *See, e.g., id.*; Lipton et al., *supra* note 4.

[179] *Id.*

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 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 are highly correlated with protected status after concluding that their use is not practically or morally justified. For example, if a tool was found to rely on irrelevant high school activities to predict job performance,[180] or customer reviews that reflect racial stereotypes,[181] 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 over which an individual has no control, such as 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 seeking to develop 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.[182] Lipton et al. raise the concern that redundant encoding may cause powerful DLPs intended to reduce disparate impact to effectively constitute a form of "treatment disparity" based on race.[183] 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

---

[180] One applicant screening model found that having the name "Jared" and playing high school lacrosse correlated positively with job performance. https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased/.

[181] *See, e.g.* Altenburger & Ho, *supra* note 165 (finding that consumer ratings were biased against Asian restaurants).

[182] Datta, et al., Use Privacy in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs, CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017).

[183] Lipton et al., *supra* note 4 at 2.

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 it reflects past discriminatory practices—the more defensible the model will be.

5. Using Race at Prediction Time

Some proposed strategies for addressing algorithmic bias require using information about race, not just in training, but also to make predictions. Although, as discussed above, using race to enforce demographic parity would likely trigger legal scrutiny, there are other ways it 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 requires making use of race at prediction time. Race can also be used at prediction time to set different cutoff scores for decisions,[184] or to segment data and create separate classification models for each group.[185] Race might also be included as one feature among many in a model, interacting with other attributes and modifying their impact on the 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,[186] it is far too simplistic. As explored in Part IV, 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

[184] *See, e.g.,* Kleinberg et al., *supra* note 6.

[185] *See, e.g.,* Cynthia Dwork et al., *Decoupled Classifiers for Group-Fair and Efficient Machine Learning*, 81 PROCEEDINGS OF MACHINE LEARNING RESEARCH 1 (2018).

[186] 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 11, at 826. Harned and Wallace, and Cofone, similarly assume that any explicit use of race to mitigate bias would amount to disparate treatment. Harned & Wallach, *supra* note 177, at 635; Cofone, *supra* note 8, at 1429. *See also, e.g.,* Corbett-Davies & Goel, *supra* note 4.

to other factors depending up 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.

For example, Sam Corbett-Davies et al. hypothesize that housing stability might be predictive of recidivism for whites but not blacks.[187] If housing stability is included as a feature, it might disadvantage blacks relative to whites because it will increase the risk scores for both groups, even though it is not in fact relevant to predicting risk for black defendants. Dwork, et al. suggest 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.[188] 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.[189] 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, at the same time that it improves accuracy for the minority. In situations like these, a race-aware model can improve *both* accuracy *and* fairness for all individuals.[190]

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,[191] the direction and distribution of the changes is not easily predicted in advance. 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.

---

[187] Corbett-Davies & Goel, *supra* note 4.

[188] Dwork et al., *supra* note 4.

[189] More often, blacks, 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.

[190] *See* Hellman, *supra* note 14, at 855.

[191] 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 9, at 2298-2300 (explaining how equalizing error rates across racial groups could increase the burden on communities of color); Estornell, et al., *supra* note 48.

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

B. The Causation Question

The issue of causation gained salience in Title VII cases after the Supreme Court's decision in *Bostock*, which emphasized the relevance of the "but for" causation standard. 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."[192] 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.[193] 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

[192] 140 S. Ct. at 1740.

[193] Issa Kohler-Hausmann offers a broad critique of this counterfactual approach to proving causation in racial discrimination cases. Issa Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination*, 113 NW. U. L. REV. 1163 (2018–2019). In contrast, Bent assumes that determining the counterfactual is both simple to do and theoretically correct. Bent, *supra* note 11, 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).

someone did not receive a positive outcome under a race-aware model 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 III, 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. Black and Fredrickson have 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.[194] This effect occurs without reference to any de-biasing efforts, and is observed with "surprising frequency."[195] Estornell, et al. similarly show 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.[196] 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 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, this argument faces other difficulties. It rests on the implicit assumption that any race-conscious effort to reduce bias against blacks will automatically and consistently work to the disadvantage of whites. 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

[194] Black & Fredrikson, *supra* note 50.
[195] *Id.* at 285.
[196] *See*, Estronell, et al., *supra* note 48.

necessarily have uniform effects within a racial group. It may reduce the odds of a positive outcome for some white applicants, but not all. Others 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 recidivism risk for different racial groups. Because the weights given to the other 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.[197]

This conclusion is consistent with the cases that found 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 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 of fairness. Entities relying on predictive algorithms should audit their performance to detect unjustified racial disparities.[198] 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 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 employees invested substantial time and resources to study in reliance on that declared policy.[199] Because predictive algorithms generally rely on observational data, not 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* an entity is permitted to do in order to debias a model may be restricted by anti-discrimination law, but as discussed

[197] If the practice had 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.

[198] Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017).

[199] *See* Part IV.A., *supra*.

above, many race-conscious strategies are likely permissible under existing doctrine.

The discussion in Part V. 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 occur 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 anti-discrimination 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 along with other factors in the ultimate decisions.[200] 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" and ensuring that some members of previously disadvantaged groups are included in the group of candidates that is given closer scrutiny.[201] Taking into account the racial impacts of these different ways of structuring its decision processes in order to reduce bias is also likely legally permissible.

VI. Nondiscrimination vs. Affirmative Action

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

Scholars, however, have tended to overlook the important difference between nondiscrimination and affirmative action when evaluating algorithmic fairness efforts, treating such strategies from the outset as forms of affirmative action that require legal justification. Jason Bent, for example, suggests that algorithms that are built with an awareness of sensitive characteristics like race are "a form of algorithmic affirmative action."[202] Similarly, Daniel Ho and Alice Xiang argue that

---

[200] MIRANDA BOGEN & AARON RIEKE, HELP WANTED: AN EXAMINATION OF HIRING ALGORITHMS, EQUITY, AND BIAS 5–6 (2018).

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

[202] Bent, *supra* note 11, at 816.

algorithmic fairness strategies will "likely be deemed 'algorithmic affirmative action.'"[203] These scholars then go on to argue how "algorithmic affirmative action" can be justified under existing legal doctrine.

Their ultimate goal may be to defend the lawfulness of some race-conscious strategies, however, the difference between viewing a practice as a justifiable form of affirmative action versus not discriminatory in the first place matters quite a bit both practically and conceptually. As a practical matter, it will be much harder to legally defend race-conscious model-building if it is assumed to be a form of affirmative action. And as a conceptual matter, utilizing the affirmative action frame often invokes the wrong set of assumptions. It erroneously suggests that any effort to reduce algorithmic bias somehow inflicts harm on members of previously advantaged groups,[204] even if the prior arrangements were unfair.

Consider the practical perspective first. Suppose an unsuccessful white candidate sues an employer that relied on a race-aware algorithm alleging a Title VII violation. If the employer relied on a race-conscious process in order to remove unjustifiable sources of bias, arguably no disparate treatment has occurred and the employer would bear no further burden of justification. Bent's analysis, however, starts with the assumption that a prima facie case of disparate treatment exists. By assuming that race-aware models are discriminatory, his approach places all such efforts under a legal cloud unless it can be shown that they address a "manifest imbalance" in "traditionally segregated job categories" and do not "unnecessarily trammel[]" the rights of other employees.[205] 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.[206]

Ho and Xiang similarly conclude 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."[207] They then turn to affirmative action cases in the government-contracting context to argue that algorithmic affirmative action is legally justified 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 believe these strategies will pass muster when they

---

[203] Ho & Xiang, *supra* note 12, at 134.

[204] 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 14, at 848, n. 88.

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

[206] *Id*. at 204.

[207] Ho & Xiang, *supra* note 12, at 134.

are calibrated to respond to discrimination by a specific government actor, they urge technologists to "quantify specific forms of historical discrimination."[208]

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.[209] Instead, what is required is proof of a racially discriminatory purpose.[210] 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.[211]

While Ho and Xiang believe they have 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, 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 possibility that poor, or poorly-informed, choices in building models inadvertently 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 steps in the model-building process may inadvertently introduce unfairness through flawed assumptions, biased data and the like. For example, when a government entity learns that an algorithm denies benefits to black claimants at higher rates than whites, it should investigate to understand the source of the disparity. It may be the case that the differential grant rates are not the result of actual differences in eligibility, but reflect artifacts of the model-building process—for example, lack of accurate data about marginalized groups, or cognitive biases on the part of humans responsible for coding key inputs. If so, proving a discriminatory motive should not be a prerequisite to addressing 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

---

[208] *Id.* at 148.

[209] 426 U.S. 229, 238.

[210] *Id.* at 240.

[211] As the Second Circuit put it, requiring government actors to provide evidence of their own past discrimination in order 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. Rodrigez, 863 F.2d 233 (2d Cir. 1988).

race into account may be necessary to ensure fairness and doing so should not depend upon a finding of prior discrimination by the government actor.

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 disparate treatment, from those which impose racial classifications on individuals in ways that trigger strict scrutiny.

On a conceptual level, the distinction between non-discrimination and affirmative action also matters quite a lot. Framing race-conscious efforts to ensure fair models as "affirmative action" invokes a set of assumptions and surrounding rhetoric that are not helpful and even misleading in the context of predictive algorithms.

To start, the concept of affirmative action has little relevance in the criminal context. Typically, affirmative action describes race-conscious actions intended to assist disadvantaged minority groups by increasing access to scarce resources. Involvement with the criminal enforcement system, however, does not offer resources and opportunity, but rather threatens punitive sanctions and damaging collateral consequences. And blacks are over-represented, not under-represented, among criminal defendants due to discriminatory policing and prosecutorial practices. Law enforcement actors generally do not take race-based actions in order to advantage blacks, and no prominent cases have alleged that whites have been harmed by race-conscious government action in this sphere. Given these realities, the concept of affirmative action is misplaced in the criminal law context. And it has the unfortunate consequence of activating arguments that efforts to address racial impacts on blacks somehow burden whites, even though that dynamic is largely absent in the criminal administration system.

The concept of "affirmative action" has more relevance in contexts like education, employment and government contracting, but can nevertheless be misleading when applied to efforts to de-bias algorithms. Part of the difficulty is that the term "affirmative action" is not precisely defined and has been used to refer to a broad range of efforts to address racial disparities. Some of those actions are best understood as non-discriminatory methods of leveling the playing field—for example, increasing outreach efforts to marginalized groups. In popular discourse, however, the term "affirmative action" has increasingly come to be associated with race- and sex-based preferences, often involving rigid numerical quotas. And that conceptualization in turn activates a set of stock arguments that these efforts are unfair to whites and males.

Implicit in many of the arguments against affirmative action are two related premises. One is that suspect racial classifications act as "preferences" for minorities—i.e. that they would not have received the benefit without the policy

putting a thumb on the scale in their favor. The second is that these policies necessarily impose harms on whites because they are not part of the preferred group. Both of these 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.

"Algorithmic affirmative action" implies that race-conscious model-building strategies are like the affirmative action policies subject to the legal challenges in the past. However, the strategies employed in the algorithmic context operate quite differently, such that the major premises underlying the Court's affirmative action doctrine do not apply. Unlike 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 they will tend to increase positive outcomes for previously disadvantaged groups,[212] 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, as when race was used to make school assignments. Instead, an awareness of race informs choices that go into shaping a model, usually without pre-ordaining outcomes for particular individuals.

The rhetoric of affirmative action also implies that any such efforts inherently harm members of non-protected groups. Legal challenges to affirmative action come from whites who believe that any effort by private entities or the government to remedy the effects of past racial discrimination harms them. This claim rests on the assumption that these efforts are disrupting a previously fair baseline distribution of benefits. In other words, it assumes that there exists some objectively fair method for deciding who gets what—an argument often framed in terms of "merit." For example, challengers assume that allocating places at a university should depend on test scores, or that government contracts should be given to the cheapest provider. They argue that any deviation from this presumed-to-be-fair baseline has deprived them of their rights.

Critical race scholars have challenged the notion that the prior distribution of resources and opportunities is a fair baseline against which to assess race-conscious measures.[213] They point to the myriad of ways in which private discrimination and implicit biases create systematic disadvantage for marginalized groups. Seeking to remedy that disadvantage should not be thought of as disrupting a fair baseline, but as creating a more level playing field by taking into account the realities of past unequal access to resources and opportunities.

Labeling race-conscious model building as "affirmative action" ignores these insights and validates the idea that de-biasing efforts involve a departure from some

[212] 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 48; Lipton et al., *supra* note 4; Mayson, *supra* note 9.

[213] *See, e.g.,* note 112 and sources cited therein.

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, discourse often assumes that when a firm uses a predictive model to select employees, or a bank uses one to decide who should grant a loan, there is a "correct" solution. In fact, as discussed above, there is no single, canonical model that best predicts future outcomes. Instead, there are a multitude of possible models, each reflecting a series of choices, tradeoffs, uncertainties, and weighing of values, each of which will shift the odds that any particular individual will receive the benefit.[214]

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. Instead, efforts to remove or address barriers to equal opportunity should be understood as steps toward leveling the playing field rather than enacting "preferences" for minorities.

* * * *

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 bias algorithms. In light of the changing composition of the Supreme Court, one might ask whether this analysis is beside the point. The Roberts Court has already faced growing criticism for its willingness to overturn long established precedent.[215] Regarding its race jurisprudence, scholars have pointed to cases like *Parents Involved*[216] and *Shelby County*[217] 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.[218] The addition of Brett Kavanaugh and Amy Coney Barrett to the Court has heightened concerns that the Court will double-down on colorblindness and move to end all affirmative action.

---

[214] *See* Part III., *supra*.

[215] See, e.g., Donald Ayer, *Opinion | The Supreme Court Has Gone Off the Rails - The New York Times*, https://www.nytimes.com/2021/10/04/opinion/supreme-court-conservatives.html (last visited Oct. 22, 2021) (arguing that Supreme Court is disregarding long-standing precedent and citing as examples its 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).

[216] Parents Involved in Community Schools v. Seattle School District, 551 U.S. 701 (2007).

[217] Shelby County, Ala. v. Holder, 570 U.S. 529 (2013).

[218] *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).

Rather than speculate about what the Justices will do, this Article instead focused on demonstrating that under the Court's existing jurisprudence, many efforts to debias algorithms are entirely permissible. Although there are limited institutional constraints to prevent the Court from departing from precedent, it is important to take the Justices at their word and to engage their explanations for their decisions.[219] Taking its past opinions at face value, many algorithmic debiasing strategies do not appear to entail disparate treatment or the use of racial classifications at all. For the Court to conclude that *all* race-consciousness strategies trigger close legal scrutiny would entail a radical shift. Doing so would not only destabilizing the coherence of existing doctrine, it would also call into question many widely-accepted practices and entangle the courts in reviewing and supervising routine goal-setting and policy decisions by both government and private actors.

In any case, if, as some have suggested, the current 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 types of affirmative action policies that have triggered scrutiny in the past. 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 policies disapproved of by conservative Justice in past affirmative action cases. In those cases, race was used to ensure fixed numerical outcomes,[220] or to tip the scales decisively in favor of one race over another.[221] 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, and are therefore entirely distinct from the policies that provoked concerns in the affirmative action cases.

Models that access race at prediction time fall into an area of greater uncertainty, but even there, when they incorporate race in manner that does not systematically favor one racial group over another in making individual decisions, concerns 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, no special scrutiny is warranted.

---

[219] A legal realist stance does not render precedent irrelevant. Courts must still act *through* doctrine and judicial norms demand that they justify their decisions based on precedent and legal reasoning. And because advocates and practitioners have to work with existing case law, it makes sense to engage with and leverage existing doctrine to the extent possible. *See, e.g.,* Daniel Harawa, *Lemonade: A Racial Justice Reframing of the Roberts Court's Criminal Jurisprudence*, (forthcoming CALIF. L. REV. 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).

[220] *See, e.g.,* Regents of Univ. of Cal. v. Bakke, 438 U.S. 265 (1978); United Steelworkers v. Weber, 443 U.S. 193 (1979).

[221] *See, e.g.,* Adarand Constructors v. Pena, 515 U.S. 200, 227 (1995); Gratz v. Bollinger, 539 U.S. 244 (2003).

VII. Conclusion

Scholars and advocates concerned about bias in predictive models have begun calling for "algorithmic affirmative action." That phrase is an unfortunate one, because it entails a set of assumptions and invokes rhetoric that obscures what happens when designers engage in race-aware strategies to reduce bias in models.

Despite rhetoric about colorblindness, the law does not in fact prohibit all forms of race-consciousness in private and government decision-making. These 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 it not only lowers the legal risk for designers exploring these strategies, it also lowers the temperature as well. The rhetoric surrounding affirmative action suggests that specific justification is needed because these programs harm others. Recognizing that some race-conscious strategies are not disparate treatment at all when they work to remove unfair bias from these systems, counters this rhetoric. By highlighting the fact that models reflect the myriad choices of their creators, not some objective, underlying truth about who deserves what, it undermines claims of entitlement to a system of pre-existing 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 some fairness constrained strategies may exact too large a cost in terms of accuracy,[222] or end up harming the groups they are intended to protect.[223] Whether or when they 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 simply that the fact that a strategy takes account of race should not make it presumptively unlawful. Rather than courts preemptively stepping in and taking certain options of the table, the question of which strategies are most appropriate in a given context should be subject to vigorous debate among policy-makers and the public.

[222] *See, e.g.,* Dwork et al., *supra* note 4.
[223] *See, e.g.*, Lipton et al., *supra* note 4; Mayson, *supra* note 9.