2020

Equal Protection Under Algorithms: A New Statistical and Legal Framework

Crystal S. Yang
Harvard Law School

Will Dobbie
Harvard Kennedy School

EQUAL PROTECTION UNDER ALGORITHMS: A NEW STATISTICAL AND LEGAL FRAMEWORK

Crystal S. Yang* & Will Dobbie**

In this Article, we provide a new statistical and legal framework to understand the legality and fairness of predictive algorithms under the Equal Protection Clause. We begin by reviewing the main legal concerns regarding the use of protected characteristics such as race and the correlates of protected characteristics such as criminal history. The use of race and nonrace correlates in predictive algorithms generates direct and proxy effects of race, respectively, that can lead to racial disparities that many view as unwarranted and discriminatory. These effects have led to the mainstream legal consensus that the use of race and nonrace correlates in predictive algorithms is both problematic and potentially unconstitutional under the Equal Protection Clause. This mainstream position is also reflected in practice, with all commonly used predictive algorithms excluding race and many excluding nonrace correlates such as employment and education.

Next, we challenge the mainstream legal position that the use of a protected characteristic always violates the Equal Protection Clause. We develop a statistical framework that formalizes exactly how the direct and proxy effects of race can lead to algorithmic predictions that disadvantage minorities relative to nonminorities. While an overly formalistic solution requires exclusion of race and all potential nonrace correlates, we show that this type of algorithm is unlikely to work in practice because nearly all algorithmic inputs are correlated with race. We then show that there are two simple statistical solutions that can eliminate the direct and proxy effects of race, and which are implementable even when all inputs are correlated with race. We argue that our proposed algorithms uphold the principles of the equal protection doctrine because they ensure that individuals are not treated differently on the basis of membership in a protected class, in stark contrast to commonly used algorithms that unfairly disadvantage minorities despite the exclusion of race.

We conclude by empirically testing our proposed algorithms in the context of the New York City pretrial system. We show that nearly all commonly used algorithms violate certain principles underlying the Equal Protection Clause by including variables that are correlated with race, generating substantial proxy effects that unfairly disadvantage Black individuals relative to white individuals. Both of our proposed algorithms substantially reduce the number of Black defendants detained compared to commonly used algorithms by eliminating these proxy effects. These findings suggest a fundamental rethinking of the equal protection doctrine as it applies to predictive algorithms and the folly of relying on commonly used algorithms.

TABLE OF CONTENTS

<div align="center">INTRODUCTION</div>

There has been a dramatic increase in the use of predictive algorithms in recent years. Predictive algorithms typically use individual characteristics to predict future outcomes, guiding important decisions in nearly every facet of life. In the credit market, for example, these algorithms use characteristics such as an individual's credit and payment history to predict the risk of default, often summarized as a single "credit score."[1] These credit scores are used in almost all consumer-lending decisions, including both approval and pricing decisions for credit cards, private student loans, auto loans, and home mortgages.[2] Credit scores are also widely used in nonlending decisions, such as rental decisions for apartments.[3] In the labor market, predictive algorithms use characteristics such as an individual's past work experience and education to predict productivity or tenure, with employers using these predictions to make hiring, retention, and promotion decisions.[4] In the criminal justice system—the focus of our Article—predictive algorithms use characteristics such as an individual's criminal history and age to predict the risk of future criminal behavior, with these "risk

---

1.   Rob Berger, *A Rare Glimpse Inside the FICO Credit Score Formula*, DOUGHROLLER (Aug. 20, 2020), https://www.doughroller.net/credit/a-rare-glimpse-inside-the-fico-credit-score-formula [https://perma.cc/2A4B-PHWD].

2.   *What Is a Credit Score?*, MYFICO, https://www.myfico.com/credit-education/credit-scores [https://perma.cc/48AW-EFEW].

3.   Jim Rendon, *You Say You're a Dream Renter? Prove It.*, N.Y. TIMES (July 15, 2011), https://www.nytimes.com/2011/07/17/realestate/prospective-renters-have-much-to-prove-to-landlords.html [https://perma.cc/R2E7-RZ82].

4.   *See, e.g.*, George Anders, *Who Should You Hire? LinkedIn Says: Try Our Algorithm*, FORBES (Apr. 10, 2013, 4:31 PM), https://www.forbes.com/sites/georgeanders/2013/04/10/who-should-you-hire-linkedin-says-try-our-algorithm/#175f96f7be66 [https://perma.cc/4KY3-5AG7]; Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), https://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html [https://perma.cc/T2XR-LYYW]; Claire Cain Miller, *Can an Algorithm Hire Better than a Human?*, N.Y. TIMES: THE UPSHOT (June 25, 2015), https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html [https://perma.cc/J27N-VM8L].

assessments" used to inform pretrial-release conditions, sentencing decisions, and the dispatch of police patrols.[5]

The increasing use of these algorithms has contributed to an active debate on whether commonly used predictive algorithms intentionally or unintentionally discriminate against certain groups, in particular racial minorities and other protected classes. In theory, predictive algorithms have the potential to reduce discrimination by relying on statistically "fair" associations between algorithmic inputs and the outcome of interest.[6] Yet, critics argue that the algorithmic inputs are themselves biased, resulting in violations of the equal protection doctrine and antidiscrimination law.[7] For example, many scholars have raised questions about the growing use of predictive algorithms in making hiring and retention decisions, often arguing that Title VII of the Civil Rights Act of 1964, the primary law prohibiting employment discrimination on the basis of protected characteristics such as race, sex, religion, and national origin, proscribes the use of any such characteristics.[8] In addition, scholars have argued that using even seemingly neutral traits in these algorithms can end up "indirectly determin[ing] individuals' membership in protected classes" and subsequently harm class members if these traits are correlated with protected characteristics.[9] Reflecting these concerns, recent policy proposals regarding algorithms have sought

---

5. *See, e.g.*, Jeff Asher & Rob Arthur, *Inside the Algorithm that Tries to Predict Gun Violence in Chicago*, N.Y. TIMES: THE UPSHOT (June 13, 2017), https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html [https://perma.cc/KZG6-HNRA]; Ellora Thadaney Israni, Opinion, *When an Algorithm Helps Send You to Prison*, N.Y. TIMES (Oct. 26, 2017), https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html [https://perma.cc/ZG3C-BYH3].

6. *E.g.*, Israni, *supra* note 5.

7. In the area of credit and lending, laws like the Equal Credit Opportunity Act (ECOA) of 1974 prohibit discrimination on the basis of protected characteristics and have been interpreted to prohibit practices like "redlining," or geographic discrimination using zip codes as proxies for the racial composition of neighborhoods. 15 U.S.C. § 1691(a)(1); *cf.* Conn. Fair Hous. Ctr. v. Corelogic Rental Prop. Sols., LLC, 369 F. Supp. 3d 362, 371 (D. Conn. 2019) (interpreting the Fair Housing Act to ban use of criminal history as it could be a proxy for race). Regulation B of the ECOA also lists many factors that cannot be used in empirically derived credit-scoring systems, including public-assistance status, marital status, race, color, religion, national origin, and sex, 12 C.F.R. § 202.5 (2020), leading some to claim that "the law requires that lenders make decisions about mortgage loans as if they had no information about the applicant's race, regardless of whether race is or is not a good proxy for risk factors not easily observed by the lender." Helen F. Ladd, *Evidence on Discrimination in Mortgage Lending*, J. ECON. PERSPS., Spring 1998, at 41, 43.

8. Solon Barocas and Andrew Selbst, for example, have argued that, in the employment context, "considering membership in a protected class as a potential proxy is a legal classificatory harm in itself" and that "[u]nder formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds." Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 695, 719 (2016).

9. *Id.* at 692.

to prohibit the use of protected characteristics, either directly or through proxies. For example, in 2019, the Department of Housing and Urban Development issued a proposal that allows landlords to use a predictive algorithm to screen tenants but prohibits the use of inputs that are deemed to be "substitutes or close proxies" for protected characteristics.[10]

The debate about whether commonly used predictive algorithms discriminate against minorities has been particularly heated in the criminal justice system, where risk-assessment tools are increasingly utilized.[11] Critics of algorithmic risk assessments have argued that use of demographic characteristics such as race or gender in predictive algorithms "amounts to overt discrimination based on demographics and socioeconomic status" and note that use of these characteristics "can be expected to contribute to the concentration of the criminal justice system's punitive impact among those

---

10.   Andrew D. Selbst, *A New HUD Rule Would Effectively Encourage Discrimination by Algorithm*, SLATE: FUTURE TENSE (Aug. 19, 2019, 10:51 AM), https://slate.com/technology /2019/08/hud-disparate-impact-discrimination-algorithm.html   [https://perma.cc/29AP-NS3L].

11.   The American Bar Association, for example, has urged states to adopt risk-assessment tools in order to protect public safety, with a goal of reducing incarceration and recidivism among low-risk offenders. *See* CRIM. JUST. SECTION, AM. BAR ASS'N, STATE POLICY IMPLEMENTATION PROJECT 18. The National Center for State Courts' Conference of Chief Justices and Conference of State Court Administrators similarly recommends that "offender risk and needs assessment information be available to inform judicial decisions regarding effective management and reduction of the risk of offender recidivism." NAT'L CTR. FOR STATE CTS., CONF. OF CHIEF JUSTS. & CONF. OF STATE CT. ADM'RS, RESOLUTION 7: IN SUPPORT OF THE GUIDING PRINCIPLES ON USING RISK AND NEEDS ASSESSMENT INFORMATION IN THE SENTENCING PROCESS (2011), https://www.ncsc.org/~/media/Microsites/Files/CSI/Resolution-7.ashx [https://perma.cc/DHT3-SFP2]. Several states have also passed legislation in recent years requiring that judges be provided with risk assessments at sentencing. *See, e.g.*, KY. REV. STAT. ANN. § 532.007(3)(a) (LexisNexis 2016) ("Sentencing judges shall consider . . . the results of a defendant's risk and needs assessment included in the presentence investigation."); OHIO REV. CODE ANN. § 5120.114(A)(1)–(3) (LexisNexis 2014) (the Ohio department of rehabilitation and correction "shall select a single validated risk assessment tool for adult offenders" that shall be used for purposes including sentencing); 42 PA. STAT. AND CONS. STAT. ANN. § 2154.7(a) (West Supp. 2019) (in Pennsylvania, a risk-assessment instrument shall be adopted to help determine appropriate sentences); *see also* ARIZ. CODE OF JUDICIAL ADMIN. § 6–201.01(J)(3) (2016) ("For all probation eligible cases, presentence reports shall . . . contain case information related to criminogenic risk and needs as documented by the standardized risk assessment and other file and collateral information."); OKLA. STAT. ANN. tit. 22, § 988.18(A) (West 2016) (an assessment and evaluation instrument designed to predict risk to recidivate is required to determine eligibility for any community punishment). Many other states permit the use of such algorithmic tools. *See, e.g.*, IDAHO CODE § 19-2517 (Supp. 2016) (if an Idaho court orders a presentence investigation, the investigation report for all offenders sentenced directly to a term of imprisonment and for certain offenders placed on probation must include current recidivism rates differentiated based on offender risk levels of low, moderate, and high); LA. STAT. ANN. § 15:326(A) (2016) (some Louisiana courts may use a single presentence investigation validated risk- and needs-assessment tool prior to sentencing an adult offender eligible for assessment); WASH. REV. CODE § 9.94A.500(1) (2016) (requiring a court to consider risk-assessment reports at sentencing if available).

who already disproportionately bear its brunt, including people of color."[12] There are also concerns that seemingly neutral algorithmic inputs such as employment and education may nonetheless result in unwarranted racial disparities because they may serve as proxies for race.[13]

These concerns are echoed in statements made by prominent public officials, including former Attorney General Eric Holder, who argue that "[b]y basing sentencing decisions on static factors and immutable characteristics—like the defendant's education level, socioeconomic background, or neighborhood—they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."[14] Even commonly used algorithmic inputs such as current charge and prior criminal history, which many argue are both relevant and legally permissible,[15] may generate unwarranted disparities. For example, an individual's prior criminal history can be driven, at least in part, by racial biases in policing, not just past criminal behavior. In this scenario, using prior arrests as an algorithmic input can result in past discrimination being "baked in" to the algorithm.[16]

---

12.   Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 806 (2014).

13.   *See, e.g.*, Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 681 (2016) ("[C]ontroversy has begun to swirl around the use of risk assessment in sentencing. The principal concern is that benefits in crime control will be offset by costs in social justice—that is, a disparate and adverse effect on racial minorities and the poor. Although race is omitted from these instruments, critics assert that risk factors that are sometimes included (e.g., marital history and employment status) are 'proxies' for minority race and poverty.").

14.   Eric Holder, Att'y Gen., U.S. Dep't of Just., Remarks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference (Aug. 1, 2014), https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th   [https://perma.cc/VX7R-N87B]. Larry Krasner, the current district attorney in Philadelphia, has similarly argued that "there is a real danger that the components going into the risk assessment are proxies for race and for socioeconomic status." Anna Orso, *Can Philly's New Technology Predict Recidivism Without Being Racist?*, BILLYPENN (Sept. 25, 2017, 9:00 AM), https://billypenn.com/2017/09/25/can-phillys-new-technology-predict-recidivism-without-being-racist/   [https://perma.cc/YM69-RXHW].

15.   *See, e.g.*, Holder, *supra* note 14 ("Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct.").

16.   *See, e.g.*, Stephen Goldsmith & Chris Bousquet, *The Right Way to Regulate Algorithms*, CITYLAB (Mar. 20, 2018, 11:47 AM), https://www.citylab.com/equity/2018/03/the-right-way-to-regulate-algorithms/555998/ [https://perma.cc/5EL2-MWQT] ("Data on patterns of past arrest rates, for example, might cause an algorithm to target low-income neighborhoods where officers were historically more likely to pick up black kids for possession."); *see also* Beth Schwartzapfel, *Can Racist Algorithms Be Fixed?*, MARSHALL PROJECT (July 1, 2019, 6:00 AM), https://www.themarshallproject.org/2019/07/01/can-racist-algorithms-be-fixed [https://perma.cc/UGD2-C9GP] ("But a legacy of aggressive law enforcement tactics in black neighborhoods means that real-world policing leads to 'false positives' in real life—arrests of people who turn

In this Article, we provide a new statistical and legal framework to understand the legality and fairness of using protected characteristics in predictive algorithms under the Equal Protection Clause. The framework we develop sheds new light on the main legal and policy debates regarding which individual characteristics should be included in predictive algorithms, particularly those characteristics related to race. The framework is general in nature and applies to any legal setting involving the use of predictive algorithms, but we focus our theoretical and empirical examples on a context where algorithms are increasingly ubiquitous and consequential: the decision of whether defendants awaiting trial should be detained or released back into the community prior to case disposition.

The Article proceeds in six parts. In Part I, we provide an overview of the legal and policy concerns surrounding the use of protected characteristics to make predictions about individuals in the criminal justice system. Protected characteristics are defined as those that can trigger heightened scrutiny under the Equal Protection Clause, with our focus being the use of race. Our review of the legal landscape shows that there are two main concerns related to the use of race in predictive algorithms. First, many have argued that using race directly as an algorithmic input is problematic and likely unconstitutional under the anticlassification principle of the Equal Protection Clause. The general consensus is that the direct use of race will generate unwarranted racial disparities. Second, some have argued that even if race itself is excluded as an algorithmic input, the use of seemingly neutral inputs can still result in unwarranted disparities if those inputs act as racial proxies. For example, zip code is highly correlated with race in the real world, likely due in part to residential segregation. This correlation leads some to argue that using zip code as an algorithmic input is therefore equivalent to using race directly. As a result, numerous legal scholars and policymakers have urged jurisdictions using predictive algorithms to exclude race and factors correlated with race as inputs.[17] As noted by some scholars, the "traditional approach to anti-discrimiantion law" was to "merely . . . deprive[] the AI of information on individuals' membership in legally suspect classes or obvious proxies for such group membership."[18]

We then review the most common predictive algorithms in the criminal justice system and their inputs in Part II. Surveying the field, we find that all commonly used predictive algorithms exclude race as an input. The

---

out to be innocent of any crime—as well as convictions that wouldn't have occurred in white neighborhoods. And because risk assessments rely so heavily on prior arrests and convictions, they will inevitably flag black people as risky who are not.").

17.   *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L.J. 2218, 2224 (2019) ("Among racial-justice advocates engaged in the debate, a few common themes have emerged. The first is a demand that race, and factors that correlate heavily with race, be excluded as input variables for prediction.").

18.   Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 Iowa L. Rev. 1257, 1276 (2020).

universal exclusion of race as an algorithmic input is unsurprising given the mainstream legal view that the direct use of race as an input would be unconstitutional. There is less uniformity in the use of nonracial algorithmic inputs that may be correlated with race. At least some commonly used predictive algorithms purposely exclude nonrace inputs such as education and socioeconomic status out of a concern that they are proxies for race. On the other hand, other commonly used algorithms include many nonrace inputs that are likely to be racial proxies.

In Part III, we develop a statistical framework that formalizes the mainstream legal position that the use of both race and nonrace correlates is problematic on fairness grounds or potentially unconstitutional under the Equal Protection Clause. Building on this mainstream position, we define a predictive algorithm as fair (and "race neutral") if and only if it does not use information stemming from membership in a racial group to form predictions, either directly through the use of race itself or indirectly through the use of nonrace correlates. We illustrate these direct and proxy effects through the use of simple examples, showing exactly how both direct use of race and indirect use of nonrace correlates can generate unwarranted racial disparities.

Building on this statistical framework, we discuss three potential solutions in Part IV that can eliminate the direct and proxy effects of race in predictive algorithms. The first formalistic solution, the "excluding-inputs" algorithm, reflects what we believe to be the general legal mainstream position. This algorithm yields race neutrality by explicitly excluding both race *and* all race-correlated inputs from algorithms, thereby mechanically eliminating both direct and proxy effects of race. While such an algorithm exists in theory, we question its feasibility in practice given the empirical reality that almost every algorithmic input is likely correlated with race due to the influence of race in nearly every aspect of American life today. We argue that, because of this fact, none of the commonly used predictive algorithms in the criminal justice system, even those that explicitly exclude some race-correlated inputs, are able to achieve full race neutrality. Even if there remain some inputs that are uncorrelated with race, the set of permissible inputs under this formalistic solution is likely so small that the accuracy of the algorithm will be substantially degraded.

We then introduce our two proposed solutions, the "colorblinding-inputs" and "minorities-as-whites" statistical models. These two statistical solutions improve upon current practice by purging all predictions of both direct and proxy effects of race. As we demonstrate in Part IV, these statistical solutions are implemented in a two-step procedure where race is used in the first estimation step in order to eliminate proxy effects. In the second prediction step, however, no individual-level race data is utilized. Our first recommended solution purges all algorithmic inputs of the proxy effects of race in the estimation step of the predictive algorithm, and then uses these "colorblind" inputs to predict outcomes in the prediction step. Our second recommended solution instead uses only white individuals in the estimation step of the predictive algorithm, and then uses these

"colorblind" estimates to predict outcomes for both white and Black individuals in the prediction step. Our two recommended solutions allow us to address direct and proxy effects of race without jettisoning all race-correlated inputs. Both our proposed algorithms achieve race neutrality by considering or using race in the first estimation step of the algorithm, but prohibit race from being used in ultimate decisionmaking in the second prediction step. This important concept, however, may run counter to the intuitive but statistically incorrect and overly formalistic anticlassification principle that the use of race in any form would violate the Equal Protection Clause.[19] While not used in practice today, likely because of the perceived unconstitutionality of using race in any form, we argue that our two proposed solutions uphold the primary principles underlying the equal protection doctrine. Our algorithms are consistent with the anticlassification principle, as they ensure that individuals are *not* treated differently because of membership in a particular racial group, eliminating unwarranted racial disparities. Our proposed algorithms are also consistent with the antisubordination principle, as they are designed to avoid inflicting harm on disadvantaged groups.[20]

In Part V, we empirically test our two proposed solutions in the context of the New York City pretrial system. We find that all commonly used algorithmic inputs are correlated with race in the New York City data, including current charge and prior criminal history, thereby generating proxy effects even when race itself is explicitly excluded from a predictive algorithm. These results confirm that commonly used predictive algorithms violate certain principles underlying the Equal Protection Clause by including algorithmic inputs that are correlated with race and thus fail to achieve race neutrality. Our empirical findings also show that the overly formalistic exclusion of race actually generates unwarranted racial

---

19.   *See, e.g.*, Starr, *supra* note 12, at 870 ("The inclusion of demographic and socioeconomic variables in risk prediction instruments . . . is normatively troubling and, at least with respect to gender and socioeconomic variables, very likely unconstitutional."); *see also* Mayson, *supra* note 17, at 2240 ("[C]olorblindness . . . would [simply] prohibit the use of race as an input variable for prediction [and] the intentional use of race proxies[]." (emphasis omitted)).

20.   This antisubordination principle is most closely linked to Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFFS. 107 (1976). As summarized by David Strauss,

> [t]his principle holds that the evil of discrimination does not lie in the use of a racial (or other similar) criterion for distinguishing among people. Rather the evil of discrimination is the particular kind of harm that it inflicts on the disadvantaged group—in varying formulations, it subordinates them, or stigmatizes them, or brands them with a badge of caste. According to the anti-subordination principle, where that particular kind of harm is absent, there is no unlawful discrimination, even if a racial classification is used. Affirmative action is (according to its supporters) an example of the non-subordinating use of a racial classification.

David A. Strauss, *"Group Rights" and the Problem of Statistical Discrimination*, ISSUES IN LEGAL SCHOLARSHIP, 2003, at 1, 1.

disparities, undermining the objective of equal treatment.[21] We then illustrate the value of our two proposed algorithms in predicting pretrial risk. We find that New York City could substantially reduce the number of Black defendants detained if it used our proposed statistical models instead of the more commonly used predictive algorithms.

Finally, in Part VI, we discuss extensions of our proposals. We illustrate how our methods can readily allow for many protected characterisics, not just race. Our statistical approaches can also allow for nonlinearities in the statistical model, more complex interactions between inputs, and extensions for machine-learning algorithms. We end by describing the relevance of our approaches to other contexts such as lending and employment.

Our Article links two important literatures: a legal literature on the constitutionality of predictive algorithms under antidiscrimination law[22] and a social science literature on algorithmic fairness.[23] In our reading, the legal literature has adopted an overly formalistic interpretation of the principles of equal treatment, leading to the misguided conclusion that the use of

---

21.   This view has been noted by only a few legal scholars in recent years. For example, Pauline T. Kim notes in the context of employment discrimination and Title VII that

> because of the problem of omitted variable bias, forbidding the use of protected class variables could exacerbate discriminatory effects under certain circumstances. Thus, a blanket prohibition on the explicit use of race or other prohibited characteristics does not avoid, and may even worsen, the discriminatory impact of relying on a data model.

Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm. & Mary L. Rev. 857, 904 (2017). Similarly, Aziz Z. Huq notes that

> [r]ace is commonly thought to be already highly correlated with socioeconomic characteristics related to criminogenic and victimization distributions. It might hence be reasonably anticipated that many algorithmic tools designed to be predictive of criminality will, even absent any race feature in the training data, generate a function that either mimics, or is a good approximation of, racial distributions in the population.

Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L.J. 1043, 1100 (2019).

22.   *See, e.g.*, Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. Rev. 671, 694–95 (2015); Starr, *supra* note 12, at 821–41; *see also* Barocas & Selbst, *supra* note 8, at 698; Kim, *supra* note 21, at 904.

23.   *See, e.g.*, Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* Discrimination and Privacy in the Information Society 43, 50 (Bart Custers, Toon Calders, Bart Schermer & Tal Zarsky eds., 2013); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ashesh Rambachan, *Advances in Big Data Research in Economics: Algorithmic Fairness*, 108 Am. Econ. Ass'n Papers & Proc. 22, 26 (2018) ("Our central argument is that across a wide range of estimation approaches, objective functions, and definitions of fairness, the strategy of blinding the algorithm to race inadvertently detracts from fairness."); Devin G. Pope & Justin R. Sydnor, *Implementing Antidiscrimination Policies in Statistical Profiling Models*, Am. Econ. J., Aug. 2011, at 206, 218; Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, 2012 Innovations Theoretical Comput. Sci. Conf. 214; Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, NeurIPS (2016), https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf [https://perma.cc/5Y4L-QVMF].

protected characteristics is always unconstitutional. In contrast, the computer science and economics literature has long recognized the value of using protected characteristics in predictive algorithms[24] but has largely ignored the implications of such use under the law.[25] We seek to provide a bridge between these literatures by (1) identifying the key challenges that predictive algorithms pose to existing legal understandings of fairness (as opposed to social science conceptions of fairness) and (2) suggesting statistical solutions that we believe can address these notions of fairness. In doing so, we note that we are not offering a wholehearted endorsement of the use of algorithms in all aspects of life. Instead, we seek to provide a synthesizing framework that tackles prominent legal concerns related to the use of protected characteristics in predictive algorithms.

The main contribution of our Article is to challenge the mainstream legal position that the use of a protected characteristic always violates the Equal Protection Clause, a position that we argue can actually undermine the goals of equal protection, while providing concrete solutions to eliminating unwarranted racial disparities in predictive algorithms. Our findings require a fundamental rethinking of the equal protection doctrine as applied to predictive algorithms. The doctrine should embrace the statistical reality that virtually all algorithmic inputs are correlated with race, and, as a result, that blinding algorithms to race through exclusion does not best serve the goal of equal treatment under the law.

I.   PREDICTIVE ALGORITHMS AND THE EQUAL PROTECTION CLAUSE

In this Part, we review the main legal concerns surrounding the use of protected characteristics such as race and the correlates of those protected characteristics, such as criminal history, in predictive algorithms. We first describe the view that protected characteristics should not be used directly in forming predictions, regardless of whether the use of the characteristic would benefit or harm the protected group, a legal position that arises from an interpretation of the Equal Protection Clause. We then discuss the view that even if protected characteristics are not used directly, the use of other nonprotected characteristics can essentially "proxy" for these protected characteristics because of their correlation with those characteristics. We

---

24.   *See, e.g.*, Hardt et al., *supra* note 23, at 1 ("A naive approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of 'fairness through unawareness' is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features."); *see also* Indrė Žliobaitė, Faisal Kamiran & Toon Calders, *Handling Conditional Discrimination*, 2011 IEEE INT'L CONF. ON DATA MINING 992, 992 ("[D]iscrimination may occur even if the sensitive information is not directly used in the model . . . .").

25.   One exception is Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019) (providing an analysis of the gap between the literature on algorithmic fairness and antidiscrimination law in the context of lending).

conclude by discussing an alternative view of algorithms that prioritizes algorithmic accuracy. Throughout, we define protected characteristics as those that trigger heightened scrutiny (either strict or intermediate) under the Equal Protection Clause, including both suspect and quasi-suspect classes. While we largely focus on race, other examples of these classes include national origin, religion, and gender.

A. *Direct Effects of Protected Characteristics*

The first legal concern surrounding the use of protected characteristics is that their use would directly harm or benefit an individual based solely on membership in a protected class. This "direct effect" of using protected characteristics is a common concern in the context of the criminal justice system because of the robust statistical relationship between protected characteristics and most outcomes of interest. For example, in the context of pretrial-release decisions, Black defendants are often more likely to not appear in court or be rearrested before case disposition compared to otherwise similar white defendants.[26] This positive correlation between race and pretrial misconduct means that predictive algorithms will assign a higher risk score to Black defendants compared to otherwise similar white defendants if race is used as an algorithmic input. The fact that women are statistically less likely to not appear in court or be rearrested before case disposition similarly means that predictive algorithms will assign women a lower risk score compared to otherwise similar men if gender is used as an input.

This concern has led many to argue against the direct use of protected characteristics in algorithms. These claims are usually constitutional in nature and center around the prohibition against classification under the equal protection doctrine. Under the Equal Protection Clause, the use of protected characteristics such as race or national origin is a form of suspect classification. Generally speaking, government laws or policies that contain explicit racial classifications and treat individuals differently on the basis of those classifications, whether to burden or benefit such groups, violate the Constitution's "immunity from inequality of legal protection."[27] While not a blanket ban on the use of racial classifications, the Equal Protection Clause does subject such classifications to strict scrutiny.[28] Under strict scrutiny, a policy with a racial classification must serve a compelling government

---

26.   *See infra* Part V.

27.   Strauder v. West Virginia, 100 U.S. 303, 310 (1879) (invalidating the conviction of a Black defendant tried under a state that limited jury service to "white male persons . . . twenty-one years of age").

28.   *See* Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (using strict scrutiny when "the government distributes burdens or benefits on the basis of individual racial classifications").

interest and must be narrowly tailored to achieve that interest.[29] The Court applies strict scrutiny to all racial classifications "to 'smoke out' illegitimate uses of race by assuring that [the government] is pursuing a goal important enough to warrant use of a highly suspect tool."[30] While many racial classifications are struck down under strict scrutiny, not all are invalidated, including most recently the use of race as a "plus factor" in university admissions.[31]

Classifications along other lines may also pose constitutional issues, despite not being subject to strict scrutiny. In the context of gender, for example, parties who seek to defend gender-based government action must demonstrate an "exceedingly persuasive justification,"[32] grounded in the principle "that neither federal nor state government acts compatibly with the equal protection principle when a law or official policy denies to women, simply because they are women, full citizenship stature—equal opportunity to aspire, achieve, participate in and contribute to society based on their

---

29.   *E.g.*, Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 235 (1995) ("Federal racial classifications, like those of a State, must serve a compelling governmental interest, and must be narrowly tailored to further that interest.").

30.   City of Richmond v. J. A. Croson Co., 488 U.S. 469, 493 (1989) (plurality opinion).

31.   Strict scrutiny is not "strict in theory, but fatal in fact." *Adarand*, 515 U.S. at 237 (quoting Fullilove v. Klutznick, 448 U.S. 448, 519 (1980)). For example, one of the earliest examples of a racial classification that was upheld by the Supreme Court was a federal curfew applicable only to persons of Japanese ancestry. Hirabayashi v. United States, 320 U.S. 81 (1943). While the Supreme Court noted that "racial discriminations are in most circumstances irrelevant and therefore prohibited," it nonetheless upheld the curfew on due process grounds because "circumstances within the knowledge of those charged with the responsibility for maintaining the national defense afforded a rational basis for the decision which they made." *Id.* at 100, 102. Under similar arguments, the Court also upheld Executive Order 9066, which ordered Japanese Americans regardless of citizenship to internment camps under the grounds of "military necessity." Korematsu v. United States, 323 U.S. 214, 218 (1944) (holding that although "exclusion from the area in which one's home is located is a far greater deprivation than constant confinement to the home from 8 p.m. to 6 a.m.," the racially discriminatory order was nonetheless within the federal government's power).

In recent years, the application of strict scrutiny has not invalidated the use of race in certain admissions policies. For example, in *Grutter v. Bollinger*, the Supreme Court upheld the use of race as one factor in the University of Michigan Law School's admissions program, a consideration designed to "achieve that diversity which has the potential to enrich everyone's education and thus make a law school class stronger than the sum of its parts." 539 U.S. 306, 315, 343 (2003). Applying strict scrutiny, the Court held that the Law School had a "compelling interest in attaining a diverse student body" and that the admissions policy was narrowly tailored because race, a "plus factor," was used in a "flexible, nonmechanical way" that allowed for a "truly individualized consideration." *Grutter*, 539 U.S. at 328, 334. Similarly, in *Fisher v. University of Texas at Austin*, the Court upheld a race-conscious admissions program at the University of Texas, where race was one factor considered in each applicant's "Personal Achievement Score" (PAS). 136 S. Ct. 2198, 2205–07 (2016).

32.   United States v. Virginia, 518 U.S. 515, 519, 531 (1996) (holding that the exclusively male admissions policy of the Virginia Military Institute (VMI) at the time violated the Equal Protection Clause).

individual talents and capacities."[33] The Supreme Court has stated a demanding standard for gender-based classifications, requiring the state to show "at least that the [challenged] classification serves 'important governmental objectives and that the discriminatory means employed' are 'substantially related to the achievement of those objectives.'"[34] While there are numerous examples of gender-based classifications that have been invalidated, some have been upheld.[35]

To date, there is no legal precedent on how these anticlassification principles are applied to predictive algorithms. The mainstream view on this issue is best exemplified in a widely cited article by Sonja Starr, who decries the use of demographic (race and gender) and socioeconomic traits in risk assessment.[36] Focusing on risk-assessment tools used at sentencing, Starr argues that risk-assessment instruments using characteristics such as race and gender "amount[] to overt discrimination based on demographics and socioeconomic status."[37] Starr specifically argues that using demographic and socioeconomic characteristics to generate predictions of future criminality violates the Equal Protection Clause and that using such traits "can be expected to contribute to the concentration of the criminal justice system's punitive impact among those who already disproportionately bear its brunt, including people of color."[38] One of Starr's main concerns is

---

33. *Id.* at 532 (citing Kirchberg v. Feenstra, 450 U.S. 455, 462–63 (1981), and Stanton v. Stanton, 421 U.S. 7 (1975)).

34. *Id.* at 533 (alteration in original) (citations omitted). For other cases that applied intermediate scrutiny to gender classifications, see *Mississippi University for Women v. Hogan*, 458 U.S. 718 (1982), and *Craig v. Boren*, 429 U.S. 190 (1976).

35. For example, in *Nguyen v. INS*, the Supreme Court upheld a federal statute that imposed different requirements for a child's acquisition of citizenship depending on whether the citizen parent is the mother or father. 533 U.S. 53, 70 (2001) ("It is almost axiomatic that a policy which seeks to foster the opportunity for meaningful parent-child bonds to develop has a close and substantial bearing on the governmental interest in the actual formation of that bond."). Similarly, in *Califano v. Webster*, the Supreme Court upheld a federal statute that favored the calculation of old-age insurance benefits for female wage earners relative to otherwise similarly situated male wage earners. 430 U.S. 313, 316–17 (1977) (per curiam) ("Reduction of the disparity in economic condition between men and women caused by the long history of discrimination against women has been recognized as such an important governmental objective.").

36. Starr, *supra* note 12, at 806.

37. *Id.* Aziz Huq calls this assertion a "dubious proposition" and "not . . . an accurate statement of current law." Huq, *supra* note 21, at 1058. Richard Primus has also noted, "[M]any practices that do involve government actors' identifying people by race are not always subject to strict scrutiny." Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 505 (2003) (citing to examples like the collection of demographic data by the Census Bureau, state legislatures' race-based redistricting practices, and social service agencies' race-conscious adoption placements).

38. Starr, *supra* note 12, at 806, 819; *see also* Sonja B. Starr, *The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong*, 27 FED. SENT'G REP. 229, 230 (2015) ("When the government instructs judges to consider risk scores based on factors

therefore that the use of protected characteristics will exacerbate unwarranted disparities in the criminal justice system, particularly along racial lines.

    Race as an Algorithmic Input: The strongest arguments against the use of protected characteristics as algorithmic inputs concern race and ethnicity.[39]

---

like these, it is explicitly endorsing sentencing discrimination based on factors the defendant cannot control. It is embracing a system that is bound to worsen the intersectional racial, class, and gender disparities that already pervade our criminal justice system.").

    39.   In contrast to the general consensus that race is prohibited from algorithms, the use of gender and socioeconomic factors as algorithmic inputs is far less settled. For example, the Model Penal Code on Sentencing, while expressly disapproving of using race in predicting risk, has argued that "consideration of gender for the narrow purpose of risk and needs assessments is expressly permitted." MODEL PENAL CODE: SENTENCING § 6B.09 reporter's note (AM. L. INST., Tentative Draft No. 2 2011). Similarly, in a recent article, Christopher Slobogin argues that "race should never be a risk factor. Other noncriminal risk factors should be included in an RAI only if they appreciably improve predictive validity. This limitation would probably still permit reliance on variables such as age and gender, since they appear to improve accuracy significantly." Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583, 592 (2018). For example, John Monahan has argued, with respect to gender, that the fact "[t]hat women commit violent acts at a much lower rate than men is a staple in criminology and has been known for as long as official records have been kept." John Monahan, *A Jurisprudence of Risk Assessment: Forecasting Harm Among Prisoners, Predators, and Patients*, 92 VA. L. REV. 391, 416 (2006). Based on this fact, Monahan unequivocally states that

> classifying by gender for the purpose of violence risk assessment should have little diffi-
> culty surviving an equal protection challenge: The government's police power objective
> in preventing violence in society is surely "important," and including gender as a risk
> factor on an actuarial prediction instrument is "substantially related" to the accuracy
> with which such an instrument can forecast violence—and therefore assist in its preven-
> tion.

Id. at 431. However, other scholars like Starr have argued that equal protection principles forbid the use of gender and poverty in risk-assessment tools. With respect to gender, for example, Starr claims that Supreme Court cases pertaining to drinking, juries, and workforce participation have prohibited actors from making decisions that differ by gender simply because there is a statistical difference between groups. *See* Starr, *supra* note 12, at 823–29. Starr specifically questions the notion that "actuarial fairness," or relatedly statistical discrimination, is permissible under the Constitution. *Id.* at 825–26 (citing cases like *Craig v. Boren*, 429 U.S. 190, 191–92 (1976)). She concludes that "the Supreme Court has squarely rejected statistical discrimination—use of group tendencies as a proxy for individual characteristics—as a permissible justification for otherwise constitutionally forbidden discrimination." *Id.* at 827. She therefore argues that the use of gender in risk-assessment tools would be constitutionally impermissible as well, even though consideration of gender would typically lead to lower predicted risk for women. *See id.* at 825*; see also* Sidhu, *supra* note 22, at 700 (arguing that sex-based classifications would also fail intermediate scrutiny). With respect to poverty and socioeconomic status, some argue that these inputs would be constitutionally permissible in predictive algorithms. *See, e.g.*, *id.* at 700–01 ("Whereas classifications based on race, national origin, religion, and sex are presumptively unconstitutional, different treatment premised on socioeconomic status enjoys a presumption of constitutionality. . . . Accordingly, socio-economic status does not seem to offend the constitutional guarantee of Equal Protection . . . ."); *see also* Harris v. McRae, 448 U.S. 297, 323 (1980) ("[T]his Court has held repeatedly that poverty, standing alone, is not a suspect classification."). Others argue that that use of socioeconomic inputs in

For example, Starr claims that there "appears to be a general consensus that using race would be unconstitutional."[40] Starr therefore takes the position that it is relatively settled in the law that race is an impermissible input into risk-assessment instruments. A more recent paper by Dawinder Sidhu echoes many of these claims, stating that the Supreme Court's anticlassification cases "should put to rest any suggestion that [race and religion] are constitutionally appropriate in risk-assessment[]."[41] Sharing these views, Christopher Slogobin raises similar equal protection issues with risk assessment in the juvenile context. As he notes, "use of race[and] ethnicity . . . as risk factors should require a compelling justification"[42] because they are highly suspect classifications. But he argues that such a justification is "unlikely, given the less-than-robust correlation between these characteristics and risk, as well as the large number of other risk factors available to the government."[43] Ultimately, Slogobin states that "most courts have accepted the proposition that race may not be considered in determining dangerousness."[44] Significantly, because the Equal Protection Clause has been viewed as prohibiting classifications based on protected characteristics, regardless of whether the classification would harm *or* benefit the protected group, it does not matter if race would in some instances benefit individuals in the protected class.

The view that race is impermissible as an algorithmic input is perhaps not surprising, and even intuitive, given that courts have typically struck down sentencing decisions made by human decisionmakers on the basis of race.[45] Numerous courts and sentencing commissions have, for example,

---

risk-assessment tools is unconstitutional because it is equivalent to "punishing a person for his poverty." Starr, *supra* note 12, at 831, 834; Bearden v. Georgia, 461 U.S. 660, 671 (1983).

40.   Starr, *supra* note 12, at 812.

41.   Sidhu, *supra* note 22, at 699.

42.   Christopher Slobogin, *Risk Assessment and Risk Management in Juvenile Justice*, CRIM. JUST., Winter 2013, at 10, 13–14.

43.   *Id.* at 14.

44.   *Id.* at 13–14.

45.   *See generally* Carissa Byrne Hessick, *Race and Gender as Explicit Sentencing Factors*, 14 J. GENDER, RACE & JUST. 127 (2010) (providing an in-depth history of the use of race and gender in sentencing). For example, in *United States v. Kaba*, 480 F.3d 152 (2d Cir. 2007), the Second Circuit vacated and remanded the defendant's case, finding that the district court impermissibly based its sentence on the defendant's national origin. While the district court justified the sentence on deterrence grounds, the Second Circuit stated that "[a]lthough deterrence is undoubtedly a proper consideration in imposing sentence, we reject the view that a defendant's ethnicity or nationality may legitimately be taken into account in selecting a particular sentence to achieve the general goal of deterrence." *Id.* at 156 (quoting United States v. Leung, 40 F.3d 577, 586 (2d Cir. 1994)). In another case, *United States v. Borrero-Isaza*, the Ninth Circuit vacated and remanded the defendant's case, finding that the district court judge impermissibly considered the defendant's Colombian nationality when setting his sentence. 887 F.2d 1349, 1355 (9th Cir. 1989) ("The conclusion is unavoidable: Borrero was penalized because of

proclaimed that "[a] defendant's race or nationality may play no adverse role in the administration of justice, including at sentencing."[46]

Two examples are particularly notable. The first is the Sentencing Reform Act (SRA) of 1984, which directed the United States Sentencing Commission to "assure that the guidelines and policy statements are entirely neutral as to the race . . . of offenders."[47] This provision embodies the Judiciary Committee's position that it is inappropriate "to afford preferential treatment to defendants of a particular race."[48] However, this provision was made with respect to decisions made by human judgment alone and is related to concerns about unwarranted sentencing disparities,[49] not decisions made with the aid of risk assessments, which may generate statistically valid differences across groups. Thus, the extension of the SRA to risk-assessment tools is unclear, although some scholars have claimed that the SRA shows that "Congress declared race . . . off-limits in risk-assessment instruments in the federal system."[50]

The second noteworthy example is the American Law Institute's Draft of the Model Penal Code (MPC), a highly influential law-reform project that takes the position that race is impermissible in risk assessments. In general, the MPC has expressly endorsed the use of risk-assessment instruments:

> Responsible actors in every sentencing system—from prosecutors to judges to parole officials—make daily judgments about . . . the risks of recidivism posed by offenders. These judgments, pervasive as they are, are notoriously imperfect. They often derive from the intuitions and abilities of individual decisionmakers, who typically lack professional training in the sciences of human behavior.
>
> . . . .
>
> . . . Actuarial—or statistical—predictions of risk, derived from objective criteria, have been found superior to clinical predictions built on the

---

his national origin, and not because he trafficked in drugs that emanated from a source country.").

46.   *See, e.g.*, United States v. Leung, 40 F.3d 577, 586 (2d Cir. 1994).

47.   *See* 28 U.S.C. § 994(d).

48.   S. REP. NO. 98-225, at 171 (1983).

49.   *See id.* at 38 ("[E]very day Federal judges mete out an unjustifiably wide range of sentences to offenders with similar histories, convicted of similar crimes, committed under similar circumstances. . . . These disparities, whether they occur at the time of the initial sentencing or at the parole stage, can be traced directly to the unfettered discretion the law confers on those judges and parole authorities responsible for imposing and implementing the sentence."); *id.* at 49 ("[T]he present practices of the Federal courts and of the Parole Commission clearly indicate that sentencing in the Federal courts is characterized by unwarranted disparity and by uncertainty about the length of time offenders will serve in prison.").

50.   *See, e.g.*, Sidhu, *supra* note 22, at 694.

professional training, experience, and judgment of the persons making predictions.[51]

However, according to the reporter's note in the March 2011 draft of the Model Penal Code on Sentencing, "[t]he consideration of race and ethnicity is disapproved . . . and raises serious constitutional concerns."[52]

Case law suggests that a court would still likely apply heightened scrutiny in assessing the permissibility of using protected traits in algorithms even if statistical differences in risk between Black and white individuals are "actuarially fair."[53] Relying on statistically fair differences in risk is akin to the economics concept of statistical discrimination, or the use of observable group traits, such as race, to form accurate beliefs about the unobservable characteristics of defendants, such as risk.[54] While the Supreme Court has never explicitly addressed the constitutionality of statistical discrimination on the basis of race,[55] it has suggested that strict scrutiny would likely apply to most policies that rely on this type of rationale.[56] This is due to the fact

---

51.   MODEL PENAL CODE: SENTENCING § 6B.09 cmt. a (AM. L. INST., Tentative Draft No. 2, 2011).

52.   *Id.* § 6B.09 reporter's note.

53.   One of the cases that addresses the idea of statistical discrimination, although not framed in those terms, is *Palmore v. Sidoti*, 466 U.S. 429 (1984). In that case, a local judge granted custody of a child to the father rather than the white mother, who had remarried a Black man since being initially granted custody. The judge reasoned that this decision was in the best interests of the child because "it is inevitable that [the child] will, if allowed to remain in her present situation and attains school age and thus more vulnerable to peer pressures, suffer from the social stigmatization that is sure to come." *Id.* at 431. Despite finding that "[t]he goal of granting custody based on the best interests of the child is indisputably a substantial governmental interest for purposes of the Equal Protection Clause," and acknowledging that a child living with a stepparent of a different race may face social pressures, the Supreme Court unanimously reversed the decision, holding that "[t]he effects of racial prejudice, however real, cannot justify a racial classification removing an infant child from the custody of its natural mother found to be an appropriate person to have such custody." *Id.* at 433–34. Thus, *Palmore* suggests that statistical discrimination may be impermissible, although the Court has often described the danger of such predictions as being driven by no more "than personal speculations or vague disquietudes," Watson v. City of Memphis, 373 U.S. 526, 536 (1963), suggesting that statistical evidence showing a true relationship between race and risk may yield a different conclusion. However, most recently, in *Buck v. Davis*, an ineffective assistance of counsel case where the defense attorney introduced statistical evidence that the defendant was more likely to act violently because he is Black, the Court stated that "[i]t would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race." 137 S. Ct. 759, 775 (2017).

54.   *See, e.g.*, Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659, 659 (1972); Kenneth J. Arrow, *The Theory of Discrimination* 1 (Princeton Univ. Indus. Rels. Section, Working Paper No. 30A, 1971).

55.   The Supreme Court, however, has made it clear that Title VII prohibits statistical discrimination. *See* City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 716–17 (1978); Int'l Union, UAW v. Johnson Controls, Inc., 499 U.S. 187, 210 (1991).

56.   *See* Huq, *supra* note 21, at 1086 ("The Court has not been clear on whether such statistical discrimination triggers constitutional concerns. . . . All that can safely be said is that, at

that using race to predict the behavior of individuals is at odds with a core commitment of the anticlassification approach to equal protection, which is to treat people as individuals.[57] However, the Court has noted that strict scrutiny does not preclude the use of race-based policies narrowly tailored to the government's compelling interest in maintaining safety in the criminal justice system.[58]

Perhaps given the government's important objective of maintaining public safety, not all legal scholars agree that race is impermissible as an algorithmic input under the Equal Protection Clause. For example, J.C. Oleson argues that even under strict scrutiny, a risk assessment that included race would likely survive such analysis because race operates as a "plus factor" analogous to the use of race in affirmative action cases like *Grutter v. Bollinger*.[59] In *Grutter*, the Supreme Court upheld the use of race as one factor in the University of Michigan Law School's admissions program, a consideration designed to "achieve that diversity which has the potential to enrich everyone's education and thus make a law school class stronger than the sum of its parts."[60] Applying strict scrutiny, the Court held that the Law School had "a compelling interest in attaining a diverse student body" and

---

least in some instances, statistical discrimination will be subject to close judicial scrutiny, and sometimes it won't be. The cut-point between those domains remains to be defined."); Strauss, *supra* note 20, at 4 ("It has, I think, been generally understood that, except in extraordinary circumstances, a claim under the Equal Protection Clause or the civil rights laws cannot be defended on the ground that the act of discrimination conformed to an accurate generalization. But there was little explicit consideration of this issue, and the reason for forbidding rational statistical discrimination was never fully worked out by courts or commentators."); *see also* Starr, *supra* note 12, at 827 ("[T]he Supreme Court has squarely rejected statistical discrimination—use of group tendencies as a proxy for individual characteristics—as a permissible justification for otherwise constitutionally forbidden discrimination.").

    57.   *See* Benjamin Eidelson, *Respect, Individualism, and Colorblindness*, 129 YALE L.J. 1600, 1629 (2020) (stating that the Supreme Court's cases on race-based inferences "stand for a fairly straightforward proposition: practices that treat race as predictive of what individual people are likely to think or do show disrespect for the fact that they are individuals, not fungible members of a racial group").

    58.   Johnson v. California, 543 U.S. 499, 514 (2005). In this case, the Court considered an unwritten California prison policy that racially segregated inmates for up to sixty days upon arrival. *Id.* at 502. The asserted rationale for the policy was to prevent violence by racial gangs because an "inmate's race is a proxy for gang membership, and gang membership is a proxy for violence." *Id.* at 517 (Stevens, J., dissenting). However, while the Court held that strict scrutiny would apply to this policy, it noted that "[p]risons are dangerous places, and the special circumstances they present may justify racial classifications in some contexts. Such circumstances can be considered in applying strict scrutiny, which is designed to take relevant differences into account." *Id.* at 515. In doing so, the Court noted that "[s]trict scrutiny does not preclude the ability of prison officials to address the compelling interest in prison safety. Prison administrators, however, will have to demonstrate that any race-based policies are narrowly tailored to that end." *Id.* at 514.

    59.   J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. REV. 1329, 1377, 1385–86 (2011).

    60.   539 U.S. 306, 315 (2003).

that the admissions policy was narrowly tailored because race, a "'plus' factor," was used in a "flexible, nonmechanical way" that allowed for a "truly individualized consideration."[61]

With these cases in mind, Oleson argues that protecting the public from crime is a compelling state interest,[62] and that inclusion of race in predicting risk is narrowly tailored given studies showing that race is highly correlated with recidivism.[63] Finally, he claims that no less restrictive means will achieve the state's public-safety goal given that exclusion of race decreases the predictive accuracy of models, such that using race directly would withstand strict scrutiny.[64] As he argues, "[r]ace and its correlates can be excluded from evidence-based sentencing, but only at the cost of compromising the ability of the government to achieve its compelling interest (preventing crime)."[65] Similarly, Judge Richard Kopf has argued that "a sentencing system based upon a robust actuarial data set consisting of *all* factors [including age, race, and gender] statistically correlated with risk would arguably pass constitutional muster, even under strict scrutiny."[66] Many of these dissenting views therefore stem from the belief that, in order to protect the community from crime, one ought to use the fullest set of input characteristics possible, even protected characteristics such as race.

Summary: Based on our review, we see the mainstream legal view as generally rejecting the direct use of protected characteristics in predictive algorithms, with the strongest consensus on the impermissibility of race. This mainstream legal position views the use of protected characteristics like race as running afoul of the Equal Protection Clause's prohibition on racial classifications. This consensus is summarized well in a recent Berkman Klein report on the use of algorithms in the criminal justice system, where the authors argue that

> [v]irtually everyone agrees that race would be a constitutionally impermissible factor to include, and thus it is not included as an explicit variable in any of these systems. . . . Thus if race was explicitly included as

---

61.	*Grutter*, 539 U.S. at 328, 334.

62.	Oleson, *supra* note 59, at 1385.

63.	*Id.* at 1350, 1385–86 (citing meta-analysis of studies that identify the variables most predictive of re-offending, which include having criminal peers, antisocial personality, criminogenic needs, adult criminal history, and race).

64.	*See id.* at 1337 (citing Joan Petersilia & Susan Turner, *Guideline-Based Justice: Prediction and Racial Minorities*, 9 CRIME & JUST. 151, 173 (1987) (noting that omitting race-correlated factors reduces accuracy of recidivism prediction by five to twelve percentage points)).

65.	*Id.* at 1386.

66.	Richard G. Kopf, *Federal Supervised Release and Actuarial Data (Including Age, Race, and Gender): The Camel's Nose and the Use of Actuarial Data at Sentencing*, 27 FED. SENT'G REP. 207, 213 (2015).

an input . . . , its use in sentencing criminal defendants would almost certainly constitute an Equal Protection violation.[67]

We also view it as highly likely that courts in the near future will have to address the constitutionality of using protected characteristics, in particular race and gender, in risk-assessment instruments. For example, the United States stated in its brief as amicus curiae in *Loomis v. Wisconsin*, a case addressing the constitutionality of risk assessments at sentencing, that "the use of actuarial risk assessments might raise issues of gender or racial bias."[68] Citing the concerns raised by scholars like Starr and Sidhu,[69] the United States flagged this important question for the Supreme Court, claiming that "[i]t is a serious constitutional question, however, the extent to which actuarial assessments considered at sentencing may take account of statistical differences for male and female offenders, such as, for example, in recidivism rates. That question may warrant the Court's attention in the future in an appropriate case."[70]

B. *Proxy Effects of Protected Characteristics*

The second legal concern regarding protected characteristics is that seemingly neutral algorithmic inputs such as criminal history can proxy for suspect classes such as race. In this scenario, the use of these seemingly neutral inputs can also indirectly harm or benefit individuals based on membership in a protected class. Zip code of residence is, for example, highly correlated with race in a variety of contexts, potentially due in part to residential segregation. The correlation between race and zip code, along with the positive correlation between, say, race and pretrial misconduct, means that predictive algorithms will assign a higher risk score to individuals from majority-Black zip codes compared to otherwise similar individuals from majority-white zip codes, even when the zip code of residence has no direct effect on outcomes. As a result, some have argued that using residential zip codes in predictive algorithms is "almost tantamount to using race."[71] For example, Zach Harned and Hanna Wallach have argued, "[a] decision maker who selects applicants on the basis of race and a decision

---

67. DANIELLE KEHL, PRISCILLA GUO & SAMUEL KESSLER, RESPONSIVE CMTYS., ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM: ASSESSING THE USE OF RISK ASSESSMENTS IN SENTENCING 24 (2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf [https://perma.cc/KYP9-8DCY] (citation omitted).

68. Brief for the United States as Amici Curiae at 19, Loomis v. Wisconsin, 137 S. Ct. 2290 (2017) (No. 16-6387).

69. *See supra* notes 36–41 and accompanying text.

70. Brief for the United States, *supra* note 68, at 19 (although arguing that the petition for a writ of certiorari should be denied in this case).

71. Cathy O'Neil, *The Ethical Data Scientist*, SLATE: FUTURE TENSE (Feb. 4, 2016, 8:30 AM), https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html [https://perma.cc/23XJ-U3XJ].

maker who selects applicants by inferring their race from their zip code are doing 'exactly the same [thing], only [the latter uses] two steps rather than one. This too is a form of disparate treatment.'"[72]

It is important to note that these proxy effects of protected characteristics are completely distinct from the direct effects discussed above. Even when race itself is directly excluded from an algorithm, the inclusion of correlated algorithmic inputs may generate racial disparities.[73] We show formally in Part IV that these potentially harmful "proxy effects" will emerge whenever there is a correlation between an algorithmic input and the protected characteristic. Our empirical results demonstrate that all commonly used inputs are highly correlated with race, such that all inputs have the potential to generate proxy effects.

As with direct use of protected characteristics, there is no legal precedent regarding the use of proxies in general. Nevertheless, the mainstream view is that these proxy effects are likely problematic from a fairness perspective, and thus inputs such as zip code of residence, education, and employment status should be excluded from predictive algorithms,[74] although whether any particular algorithmic input is actually correlated with race is an empirical question that may differ across contexts.[75]

Racial Proxies as Algorithmic Inputs: The strongest arguments against the use of proxies again center on race. In the context of the criminal justice system, the main concern is that use of algorithmic inputs correlated with race will "exacerbate the unacceptable racial disparities in our criminal justice system."[76] For instance, Larry Krasner, the current district attorney in Philadelphia, has argued that "there is a real danger that the components

---

72. Zach Harned & Hanna Wallach, *Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer*, Fla. St. U. L. Rev. (forthcoming) (manuscript at 25) (quoting James Grimmelmann & Daniel Westreich, Response, *Incomprehensible Discrimination*, 7 Calif. L. Rev. Online 164, 176 (2017)).

73. Excluding protected characteristics from predictive algorithms may be completely pointless if there are other potential inputs such as socioeconomic status or education that are highly correlated with the protected characteristic. *E.g.*, Kim, *supra* note 21, at 904. Computer scientists have also highlighted the importance of proxy effects, labeling this problem "redundant encodings," defined as a situation where membership in a protected class is highly correlated with, and thus already coded, in other characteristics used in the algorithm. *See, e.g.*, Dwork et al., *supra* note 23, at 226. Economists have also noted the potential importance of proxy effects in predictive algorithms, in particular how such proxy effects could generate unwarranted disparities. *E.g.*, Pope & Sydnor, *supra* note 23, at 206.

74. *E.g.*, Starr, *supra* note 12, at 838 ("[S]ocioeconomic and family variables that [the instruments] include are highly correlated with race, as is criminal history, so they are likely to have a racially disparate impact.").

75. For example, repayment history and credit scores may generate proxy effects in the context of lending but not the criminal justice system.

76. Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 Fed. Sent'g Rep. 237, 237 (2015). These racial proxy effects are enormously prevalent as "most data we collect has some proxy power, and we are often unaware of it." O'Neil, *supra* note 71.

going into the risk assessment are proxies for race and for socioeconomic status."[77] These concerns have led to the exclusion of inputs such as education, employment status, zip code, and socioeconomic status from many predictive algorithms in the criminal justice system, as we will explore in further detail below.[78] Despite the fact that current charge and prior criminal history are routinely used,[79] some have also argued that use of these inputs "will unquestionably aggravate the already intolerable racial imbalance in our prison populations" because of their correlation with race.[80] For example, prior arrests may reflect not just actual criminal behavior but also biases in policing, such that use of prior arrests can result in past discrimination being "baked in" to the algorithm.[81] As noted by Richard Frase in the context of sentencing, for instance:

> Even when [racial] disparity results from the application of seemingly appropriate, race-neutral sentencing criteria, it is still seen by many citizens as evidence of societal and criminal justice unfairness; such negative perceptions undermine the legitimacy of criminal laws and institutions of justice, making citizens less likely to obey the law and cooperate with law enforcement.[82]

Some of the arguments against the use of racial proxies are constitutional in nature. However, the Equal Protection Clause is relatively permissive when it comes to the use of racial proxies in predictive algorithms. For instance, if a risk-assessment instrument utilized an algorithmic input such as employment or education but was otherwise facially neutral, the legality of the instrument would likely turn on the motivation for including the characteristic in the first place.[83] This position

---

77.   Orso, *supra* note 14.

78.   *See infra* Section II.B.

79.   *E.g.*, Holder, *supra* note 14 ("Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct.").

80.   Harcourt, *supra* note 76, at 237; *see also* Kelly Hannah-Moffat, *Actuarial Sentencing: An "Unsettled" Proposition*, 30 JUST. Q. 270, 279–84 (2013) (critiquing the use of criminal history variables in risk assessments because criminal history may be influenced by past discrimination).

81.   *E.g.*, Goldsmith & Bousquet, *supra* note 16. ("But many worry that the biases are simply baked into the algorithms themselves. Some opponents have argued that policing algorithms will disproportionately target areas with more people of color and low-income residents because they reinforce old stereotypes: Data on patterns of past arrest rates, for example, might cause an algorithm to target low-income neighborhoods where officers were historically more likely to pick up black kids for possession.").

82.   RICHARD S. FRASE, JUST SENTENCING: PRINCIPLES AND PROCEDURES FOR A WORKABLE SYSTEM 210–11 (2013).

83.   *See* Slobogin, *supra* note 42, at 14 ("A more complicated question is whether risk factors that might serve as a proxy for one of these classifications are legitimate. For instance, employment and education status could be statistical stand-ins for both race and age. Under

reflects current law, which states that, with respect to facially neutral laws, a government policy or law is only constitutionally problematic under the Equal Protection Clause if "motivated by a racially discriminatory purpose."[84] Indeed, the Supreme Court has clarified that "official action will not be held unconstitutional solely because it results in a racially disproportionate impact. . . . Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause."[85]

But there is a growing recognition that racial proxies (even if seemingly neutral) can be normatively troubling and undesirable even if their use is not premised on a racially discriminatory motive.[86] Thus, legal scholars have bemoaned that the equal protection doctrine would likely be a poor basis for any challenge of a facially neutral risk-assessment instrument because it would be difficult to show that the algorithm was specifically designed with a racially discriminatory motive.[87] As explained in a Berkman Klein report, while

> using factors which correlate with race may be troubling, existing constitutional doctrine does not suggest that their inclusion in a risk assessment instrument would constitute an Equal Protection violation. . . . [S]trict scrutiny is only triggered if the individuals challenging the law can

---

current equal protection law, however, unless the intent behind using these types of factors is race- or age-motivated, such a claim is likely to fail.").

84.   *Accord* Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 213 (1995); Washington v. Davis, 426 U.S. 229, 240 (1976) ("[T]he basic equal protection principle [is] that the invidious quality of a law claimed to be racially discriminatory must ultimately be traced to a racially discriminatory purpose."); *see also* Foster v. Chatman, 136 S. Ct. 1737, 1747–55 (2016) (reversing the Georgia Supreme Court's rejection of defendant's claim that the prosecution's use of peremptory strikes against black jurors was "motivated in substantial part by discriminatory intent" (quoting Snyder v. Louisiana, 552 U.S. 472, 485 (2008))). In *McCleskey v. Kemp*, the Supreme Court rejected a challenge to Georgia's capital punishment scheme despite statistical evidence showing large racial disparities in the receipt of death penalty because the evidence was "clearly insufficient to support an inference that any of the decisionmakers in [the defendant's] case acted with discriminatory purpose." 481 U.S. 279, 297–99 (1987).

85.   Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 264–65 (1977).

86.   *See supra* notes 73–75 and accompanying text.

87.   *See* Huq, *supra* note 21, at 1090 ("Without knowing the full spectrum of features that could, conceivably, have been included in the training data . . . it will be difficult or impossible to diagnose this kind of conduct absent direct evidence of discriminatory intent. It will, moreover, be especially difficult to show that, but for race, a specific feature would or would not have been included, as the doctrine requires." (citing Pers. Adm'r v. Feeney, 442 U.S. 256, 279 (1979))); *see also* Sidhu, *supra* note 22, at 699 ("To find that a facially neutral statute violates the Equal Protection Clause, the statute must be motivated by an impermissible purpose. Here, there is no indication that risk-assessment tools are driven by animus or any other illegitimate reason. Rather, these instruments are clearly used to control crime. As a result, facially neutral risk-assessments would likely survive a constitutional attack." (citation omitted)). Similar arguments have been made in the context of predictive algorithms and Title VII. *See* Barocas & Selbst, *supra* note 8, at 697–98 ("Except for masking, discriminatory data mining is by stipulation unintentional.").

show that it was also adopted with a racially discriminatory intent. If not, rational basis review applies, a highly deferential standard.[88]

In fact, some argue that proxy effects themselves *should* constitute disparate treatment under the law. Harned and Wallach, for example, argue that "the uniform application of a machine learning system—even one that is blinded to race—does not necessarily insulate against disparate treatment claims, because the system might inappropriately use proxies for race."[89]

Nevertheless, given the lack of current constitutional constraints on the use of proxies, many insead resort to normative judgments to determine whether certain racial proxies should be permitted. But the dividing line among legal scholars and policymakers between which proxies are problematic (and thus should be excluded) and which are not problematic (and thus can be included) is hard to define in theory. For example, Cathy O'Neil, author of *Weapons of Math Destruction*, has argued that figuring out which proxies are unacceptable and which are acceptable (if any) is no easy task. As she notes,

> [W]e shouldn't use race because essentially it creates this negative feedback loop, then you say, OK, well, OK, let's not use race, but should we use zip code, which of course is a proxy for race in our segregated society?
>
> And so once they acknowledge that zip code is just as good as race, then you're like, OK, so how do we choose our attributes? Because there are so many proxies to race. And it's really actually very tricky. It's tricky. And I'm not trying to claim that it's easy.[90]

Similarly, Ignacio Cofone writes:

> Blocking proxies for protected categories may be key for avoiding discriminatory outcomes. However, two central problems have been identified for doing that. The first problem is that we may not know which those proxies are and, if we did, it may be impossible to block all proxies. The second problem is that, even if it is possible to block proxies, it may be undesirable as those proxies could also contain valuable information.[91]

One possible dividing line is that correlated inputs should be excluded if the reason for the correlation is because of past discrimination or racial

---

88.   KEHL ET AL., *supra* note 67, at 24 (citing Pers. Adm'r v. Feeney, 442 U.S. 256 (1979) (holding that a statute is only invalid when the state has acted with the purpose of discriminating against a minority group, not when the statute merely has negative effects on such a group)).

89.   Harned & Wallach, *supra* note 72 (manuscript at 25).

90.   *When Not to Trust the Algorithm*, HARV. BUS. REV. (Oct. 6, 2016), https://hbr.org/ideacast/2016/10/when-not-to-trust-the-algorithm.html.

91.   Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389, 1413 (2019).

animus.[92] Otherwise, including these variables can result in discrimination being "baked in" to the algorithm, generating unjust or unwarranted disparities. In contrast, if the reason for the correlation between a variable and race is not due to discrimination, it should be included because any disparities that result may be "warranted." For example, Cass Sunstein has noted that:

> Especially difficult problems are presented if an algorithm uses a factor that is in some sense an outgrowth of discrimination. For example, a poor credit rating or a troubling arrest record might be an artifact of discrimination by human beings that occurred before the algorithm was asked to do its predictive work. There is a risk here that algorithms could perpetuate discrimination and extend its reach, by using factors that are genuinely predictive but products of unequal treatment. This might turn discrimination into a kind of self-fulfilling prophecy.[93]

Excluding proxies that are likely an "outgrowth of discrimination" is of course a challenging task, as it relies on normative judgments about the nature of discrimination. But even supposing that this principle could be implemented in theory, the current practice seems to deviate substantially from this idea. Commonly used inputs in many predictive algorithms often include racial proxies that are highly likely to be "an artifact of discrimination." For example, there is a plethora of empirical evidence suggesting that lengthier prior criminal histories among Black individuals could be due to discriminatory policing.[94] As a result, criminal history is consistently highly correlated with race.[95] Yet it is nearly universally embraced by legal scholars and policymakers and is almost always used in risk-assessment instruments.[96] In fact, criminal history is often portrayed as

---

    92.   *See, e.g.*, Prince & Schwarcz, *supra* note 18, at 1296–97 ("By allowing discriminators to indirectly but reliably take into account the ways in which historical discrimination impacts marginalized groups, proxy discrimination by AIs can cloak the reproduction of these historical hierarchies in seemingly neutral and objective structures.").

    93.   Cass R. Sunstein, *Algorithms, Correcting Biases*, 86 SOC. RSCH. 499, 509 (2019) (citation omitted).

    94.   *See, e.g.*, Decio Coviello & Nicola Persico, *An Economic Analysis of Black-White Disparities in NYPD's Stop and Frisk Program* (Nat'l Bureau of Econ. Rsch., Working Paper No. 18803, 2013); Roland G. Fryer Jr., *An Empirical Analysis of Racial Differences in Police Use of Force*, 127 J. POL. ECON. 1210 (2019); Felipe Goncalves & Steven Mello, *A Few Bad Apples? Racial Bias in Policing* (Princeton Univ. Indus. Rels. Section, Working Paper No. 608, 2017); Jeremy West, Racial Bias in Police Investigations (Oct. 2018) (unpublished manuscript) (on file with the *Michigan Law Review*).

    95.   Harcourt, *supra* note 76, at 238 ("Risk, today, is predominantly tied to prior criminal history, and prior criminality has become a proxy for race. The result is that decarcerating by means of risk instruments is likely to aggravate the racial disparities in our already overly racialized prisons.").

    96.   *See, e.g.*, Starr, *supra* note 38, at 231 ("In contrast to gender and socioeconomic variables, some other risk factors in the instruments are constitutionally permissible considera-

the counterpoint to protected characteristics such as race in terms of both legal and ethical permissibility. For example, Richard Berk and Jordan Hyatt claim that "[t]he explicit use of race, national origin, and other suspect classes for forecasting, regardless of the method, would likely fail to meet the necessary, strict scrutiny threshold. On the other hand, criminal history is relatively uncontroversial."[97] And as summarized by Mark Moore, the consensus view appears to be that

> [s]ome characteristics [used as risk factors for violence in sentencing], such as prior criminal conduct and current illegal drug use, are themselves crimes and therefore of direct interest to the criminal justice system. Others, such as race, religion, and political beliefs, are the opposite: they are specially protected against being used by criminal justice officials in making decisions.[98]

But we note that the view that criminal history is "uncontroversial" is increasingly under attack, with some commentators arguing that "[r]acism may well be a significant factor in the higher arrest and conviction rates among black people to begin with" such that "including racial proxies amounts—in effect, if not necessarily intent—to judging people by the color of their skin."[99]

Summary: Based on our review, we see the mainstream position as discouraging the use of proxies in predictive algorithms, primarily on normative grounds that using racial proxies is unfair and equivalent to using race directly. As noted above, there are likely weaker constitutional contraints on the use of proxies than the use of protected characteristics because the current equal protection doctrine has far less of a bite when dealing with facially neutral laws. As a result, deciding which proxies are permissible and which are not is often an ad hoc process, with substantial disagreement among legal scholars and policymakers.[100] Specifically, the

---

tions. These include criminal history as well as some demographic classifications, such as age, that do not trigger special constitutional scrutiny.").

97.   Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222, 226 (2015) (citing Carissa Byrne Hessick & F. Andrew Hessick, *Recognizing Constitutional Rights at Sentencing*, 99 CALIF. L. REV. 47 (2011)); *see also* Almendarez-Torres v. United States, 523 U.S. 224, 243–44 (1998).

98.   Mark H. Moore, *Purblind Justice: Normative Issues in the Use of Prediction in the Criminal Justice System*, *in* 2 CRIMINAL CAREERS AND "CAREER CRIMINALS" 314, 317 (Alfred Blumstein, Jacqueline Cohen, Jeffrey A. Roth & Christy A. Visher eds., 1986).

99.   Tafari Mbadiwe, *Algorithmic Injustice*, NEW ATLANTIS, Winter 2018, at 3, 19.

100.   A related debate is what nonrace controls should be included when testing for disparate impact in discrimination litigation. As Ian Ayres has noted, "in disparate impact testing, the primary statistical concern is most often 'included variable bias'—the worry that the statistical estimates of disparate impact are biased because the regression inappropriately *includes* non-race variables." Ian Ayres, *Testing for Discrimination and the Problem of "Included Variable Bias*," IAN AYRES 3 (2010), https://ianayres.yale.edu/sites/default/files/files/Testing%20for %20Discrimination.pdf [https://perma.cc/8GN2-FMSG]. Similarly, Jung et al. note that as

arguments in favor of or against certain inputs often rely on normative judgments of what is morally troubling and what is not.[101] These types of normative judgments include a wide range of perspectives, such as a determination of how predictive the proxy is of risk, whether the risk factor is appropriate in light of retributive goals, whether the risk factor is a product of discrimination, and whether the risk factor cannot be changed and is thus "static."[102] Because these normative judgments often conflict, scholars have summarized the legal position regarding the use of racial proxies in risk assessment as disjointed and inconsistent.[103] For example, Skeem and Lowenkamp write with respect to the legal field, "As is clear from this brief review, critics disagree in calling potentially race-related risk factors like criminal history 'in' or 'out' for the purposes of sentencing."[104]

   At the extreme, if one believes that all racial proxies should be excluded from predictive algorithms,[105] there remains no feasible way of designing an

---

   an extreme example, it is problematic to include control variables in a regression that are obvious proxies for protected attributes—such as vocal register as a proxy for gender . . . . Including such proxies will typically lead one to underestimate the true magnitude of discrimination in decisions. But what counts as a 'proxy' is not always clear. For example, given existing patterns of residential segregation, one might argue that zip codes are a proxy for race, and thus should be excluded when testing for racial bias. But one could also argue that zip code provides legitimate information relevant to a decision, and so excluding it would lead to omitted-variable bias.

Jongbin Jung, Sam Corbett-Davies, Ravi Shroff & Sharad Goel, *Omitted and Included Variable Bias in Tests for Disparate Impact*, SHARAD GOEL 2 (Aug. 29, 2019), https://5harad.com/papers/included-variable-bias.pdf [https://perma.cc/QBN5-43AE].

   101.   For example, Slobogin claims that nonrace factors should be included depending on "a normative judgment . . . about when a level of correlation is so low it requires a factor's exclusion." Slobogin, *supra* note 42, at 592–93 (arguing that age and gender are permissible because they improve accuracy, but that marital and employment status may not be). But how does one determine the "level of correlation" that determines whether a factor should be included or not? If the correlation is high, but the factor is a strong proxy for race, does that mean the input should nevertheless be included?

   102.   See Skeem & Lowenkamp, *supra* note 13, at 680–85, for a discussion of these different principles.

   103.   In the lending context, Talia Gillis similarly notes that excluding inputs that are proxies to protected characteristics "is not feasible when there is no agreed-upon definition of a proxy, and when complex interactions between variables are unidentifiable to the human eye. Even inputs that have traditionally been thought of as proxies for race, such as zip codes, may be less concerning than other ways in which a borrower's race can be recovered." Talia Gillis, *False Dreams of Algorithmic Fairness: The Case of Credit Pricing*, SCHOLARS HARV. 10–11 (Nov. 1, 2019), https://scholar.harvard.edu/files/gillis/files/gillis_jmp_191101.pdf [https://perma.cc/7RLD-SBSK].

   104.   Skeem & Lowenkamp, *supra* note 13, at 684.

   105.   *See, e.g.*, Prince & Schwarcz, *supra* note 18, at 1314; *see also* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. INSTITUTIONAL & THEORETICAL ECON. 98, 117–18 (2018) (noting that even seemingly "socially acceptable" inputs may themselves

algorithm because every possible input is likely correlated with race. As some have noted, "[i]f you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[106] We return to this question in our empirical results below.

C. *Trade-Off Between Fairness and Accuracy*

We conclude this Part by discussing an alternative view of protected characteristics that prioritizes algorithmic accuracy. The consensus view discussed above defines a predictive algorithm as "fair" if it is does not use any information stemming from membership in a protected class, either directly through the use of the protected characteristic or indirectly through the use of proxies. For example, some scholars have suggested that "antidiscrimination regimes could develop specific criteria for requiring firms that are at substantial risk of engaging in proxy discrimination to deploy 'ethical algorithms' that explicitly seek to eliminate the capacity of any facially-neutral considerations to proxy for prohibited characteristics."[107]

This definition of fairness comes with an important trade-off in terms of accuracy. Given a large literature that shows that traits like race and gender are often statistically correlated with risk,[108] choosing to exclude protected characteristics comes at the cost of predictive accuracy.[109] Removing correlated inputs that serve as proxies for protected characteristics also comes with a loss in accuracy.[110] Berk and Hyatt, for example, note the

---

proxy for race such that "because race and gender may affect everything, settling on pretreatment covariates (or socially acceptable predictors) is challenging to say the least").

106.   Nadya Labi, *Misfortune Teller*, ATLANTIC (Jan./Feb. 2012), https://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846/   [https://perma.cc/F35P-9QLP] (quoting Ellen Kurtz, director of research for the Philadelphia Adult Probation and Parole Department in 2012).

107.   Prince & Schwarcz, *supra* note 18, at 1266–67.

108.   *See, e.g.*, Paul Gendreau, Tracy Little & Claire Goggin, *A Meta-analysis of the Predictors of Adult Offender Recidivism: What Works!*, 34 CRIMINOLOGY 575, 576 (1996).

109*.   See, e.g.*, Pari McGarraugh, Note, *Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment Is Appropriate at Sentencing and Inappropriate at Parole*, 97 MINN. L. REV. 1079, 1102 (2013) ("In order to create a risk assessment instrument that does not offend the Constitution, race and ethnicity, factors closely overlapping with race and ethnicity, and gender must be purged from the list of inputs. But because race and gender are fairly reliable predictors of criminal behavior, removing them will reduce the predictive capability of risk assessments."); *see also* Kristy Holtfreter & Rhonda Cupp, *Gender and Risk Assessment: The Empirical Status of the LSI-R for Women*, 23 J. CONTEMP. CRIM. JUST. 363 (2007) (arguing for separate risk-assessment instruments for men and women given different pathways to crime for men and women).

110*.   See, e.g.*, Calders & Žliobaitė, *supra* note 23, at 54 ("The first possible solution is to remove the sensitive attribute from the training data. For example, if gender is the sensitive attribute in university admission decisions, one would first think of excluding the gender information from the training data. Unfortunately, . . . this solution does not help if some other attributes are correlated with the sensitive attribute. . . . The next step would be to remove the

concern that some algorithmic inputs may be proxies for race, but conclude that if "one could purge actuarial methods of all racial factors captured indirectly through proxy predictors[, i]t is almost certain that forecasting accuracy would decline."[111]

These two competing goals can lead to divergent views on the permissibility of including protected characteristics. As Oleson notes, there appear to be "two cultures," one which takes the stance that all predictive variables should be used, and another which takes the stance that traits like race and gender are "off-limits."[112]

In fact, the degree to which an input enhances an algorithm's accuracy may be a factor that is considered by courts.[113] For example, the degree to which a protected characteristic improves predictive accuracy may determine whether an algorithm survives strict or intermediate scrutiny because promoting accuracy can be a way of achieving a government's compelling interest.[114] In a string of recent state supreme court cases dealing with the constitutionality of algorithms in the criminal justice system, courts have generally emphasized the importance of accuracy in constructing risk-assessment instruments. Although none of these cases have dealt with equal protection challenges, courts have noted that personal characteristics, including protected characteristics like gender, may need to be taken into account in forming risk predictions because promoting accuracy is an important goal that serves both the state and criminal defendants.[115] In *State*

---

correlated attributes as well. This seems straightforward in our example dataset; however, it is problematic if the attribute to be removed also carries some objective information about the label.").

111.   Berk & Hyatt, *supra* note 97, at 227.

112.   Oleson, *supra* note 59, at 1352.

113.   *See, e.g.*, Malenchik v. State, 928 N.E.2d 564, 572–73 (Ind. 2010); State v. Loomis, 881 N.W.2d 749, 763–64 (Wis. 2016).

114.   Melissa Hamilton argues that if race and ethnicity significantly improve predictive accuracy,

> then including them would appear to be narrowly tailored to the government's compel-ling interests. . . . *If*, instead, . . . race or ethnicity was not a significant correlate . . . then developers should, practically and constitutionally, exclude it because there would be no fit with the policy's compelling need, and certainly the use of the classification would not be narrowly tailored.

Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 259 (2015). Even Starr claims that if there is a "marginal gain in predictive accura-cy" from adding characteristics like race and gender, her "constitutional objections . . . would be alleviated." Starr, *supra* note 38, at 232 (citing a few studies that purport to show that in-cluding demographic and socioeconomic factors does not significantly increase predictive ac-curacy).

115.   In *Malenchik v. State*, a 2010 case decided by the Supreme Court of Indiana, the de-fendant was sentenced to six years in prison (two years suspended) after pleading guilty to re-ceiving stolen property and admitting to being a habitual offender. 928 N.E.2d at 566. Prior to sentencing, the county probation department prepared a presentence investigation report. As

v. Loomis, for instance, a defendant was sentenced due in part to a risk-assessment tool known as COMPAS and argued, among other claims, that the algorithm's use of gender violated his due process rights.[116] The Supreme

---

part of this report, the probation department completed a Level of Service Inventory-Revised (LSI-R) risk assessment. *Id.* at 567. The probation department also conducted a Substance Abuse Subtle Screening Inventory (SASSI). On the basis of these risk assessments, the defendant was classified as high-risk/needs and as having a "high probability of having a Substance Dependence Disorder." *Id.* The scores from both LSI-R and SASSI were referenced two times by the judge at sentencing, who noted, among other things, "[Y]our LSIR score is high. Your SASSI score is high with a high probability of substance dependence disorder." *Id.* (alteration in original). After sentencing, the defendant appealed and argued that the trial court's consideration of the LSI-R score was erroneous for a variety of reasons, citing to the court of appeals's prior precedent in *Rhodes v. State*, where it had disapproved generally of the use of the LSI-R. 896 N.E.2d 1193, 1195 (Ind. Ct. App. 2008) (holding that "it is an abuse of discretion to rely on scoring models to determine a sentence").

As part of his claim that the trial court's consideration of the LSI-R was improper, the defendant argued that factors such as economic status and personal preferences, inputs into the LSI-R, are discriminatory. *Malenchik*, 928 N.E.2d at 574. However, the court rejected this argument, noting that Indiana's law required such factors to be included in the presentence investigation report and that "supporting research convincingly shows that offender risk assessment instruments, which are substantially based on such personal and sociological data, are effective in predicting the risk of recidivism and the amenability to rehabilitative treatment." *Id.* The Supreme Court of Indiana went on to laud the use of such risk assessments, stating that these "evidence-based sentencing practices [hold] considerable promise" and that they are "well supported by empirical data and provide target areas to change an individual's criminal behavior, thereby enhancing public safety." *Id.* at 569–70 (citing Christopher T. Lowenkamp & Kristin Bechtel, *The Predictive Validity of the LSI-R on a Sample of Offenders Drawn from the Records of the Iowa Department of Corrections Data Management System*, FED. PROB., Dec. 2007, at 25, 27–29).

116.   In another recent state court decision dealing with risk assessments, *State v. Loomis*, a 2016 decision by the Wisconsin Supreme Court, the defendant Eric Loomis was charged with five criminal counts related to a drive-by shooting. While he denied participating in the shooting, he pled guilty to "attempting to flee a traffic officer and operating a motor vehicle without the owner's consent." 881 N.W.2d 749, 754 (Wis. 2016). Prior to sentencing, a probation officer prepared a presentence investigation report, which included a COMPAS risk assessment. *Id.* at 755. At Loomis' sentencing, the trial judge referred to this COMPAS assessment, stating to the defendant:

> You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.

Id. at 755. Loomis was subsequently sentenced to six years in prison and five years of extended supervision. *Id.* at 756. The defendant filed a motion for postconviction relief requesting a new sentencing hearing. *Id.* Specifically, he challenged the court's consideration of the COMPAS algorithm, arguing that it violated his due process rights for several reasons, one of which was that the risk assessment improperly considered gender. *Id.* at 757. Notably, Loomis did not bring an equal protection claim regarding the use of gender. Ultimately, the court concluded that because the sentencing court essentially gave minimal weight to the COMPAS assessment and would have imposed the same sentence regardless of the risk score, the trial court's use of

Court of Wisconsin, however, determined that there was a "factual basis underlying COMPAS's use of gender . . . . [because] it appears that any risk assessment tool which fails to differentiate between men and woman will misclassify both genders."[117] As a result, the court concluded that "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose," but also found that the defendant had failed to show that the sentencing judge actually relied on gender as a factor in determining his sentence.[118]

As a result, some legal scholars have argued that exclusion of race and racial proxies would "compromis[e] the ability of the government to achieve its compelling interest (preventing crime)."[119] Thus, for an individual who seeks to maximize the accuracy of an algorithm, no input characteristics should be off-limits, including protected characteristics and their proxies.

II.   PREDICTIVE ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM

In this Part, we review the most commonly used predictive algorithms in the criminal justice system to determine how these algorithms deal with the direct and proxy effects of race.[120] We first describe the most commonly used predictive algorithms at each stage of the criminal justice system, from policing to pretrial decisions to sentencing to probation. While not meant to be an exhaustive survey of all the predictive algorithms available, we believe this review captures the most widely used and representative algorithms in the criminal justice system. We then describe how each of these predictive algorithms deals with direct and proxy effects of race.

A.   *Survey of Predictive Algorithms in the Criminal Justice System*

Policing: Predictive algorithms are increasingly used to predict crime in the United States, a phenomenon broadly known as predictive policing. The most commonly used predictive-policing algorithm is PredPol, which was created by the Los Angeles Police Department and UCLA in 2012 to predict when and where specific crimes are most likely to occur in Los Angeles.[121] The algorithm has subsequently been adopted by over sixty police

---

the algorithmic risk assessment did not violate the defendant's due process rights. *Id.* at 770–71.

117.   *Id.* at 766.

118.   *Id.* at 766–67.

119.   Oleson, *supra* note 59, at 1386.

120.   For a general overview of risk assessments in the criminal justice system, see Brandon L. Garrett & John Monahan, *Judging Risk*, 108 CALIF. L. REV. 439 (2020).

121.   *See Overview*, PREDPOL, https://www.predpol.com/about/ [https://perma.cc/HX5S-NLFJ]; Ali Winston & Ingrid Burrington, *A Pioneer in Predictive Policing is Starting a Troubling New Project*, VERGE (Apr. 26, 2018, 1:36 PM), https://www.theverge.com/2018/4/26/17285058/predictive-policing-predpol-pentagon-ai-racial-bias [https://perma.cc/JRQ4-GG2L].

departments across the country, including departments in Kansas, Washington, and South Carolina.[122] PredPol currently uses only three input variables to predict the incidence and location of future crimes: crime types, crime locations, and crime dates and times from historical data.[123] The PredPol documentation explicitly states that "[n]o demographic, ethnic or socio-economic information is ever used. This eliminates the possibility for privacy or civil rights violations seen with other intelligence-led policing models."[124]

There are also a number of predictive-policing algorithms that are used in just one city. One of the most prominent city-specific algorithms is the Strategic Subject List (SSL), or "heat list," which was created in 2013 in Chicago to predict an individual's probability of involvement in gun violence, either as a perpetrator or victim.[125] Using data on arrestees from Chicago, the algorithm predicts the probability that individuals will be involved in a shooting and ranks individuals on a risk scale of zero to 500.[126] SSL currently uses eight input variables to predict the risk of gun violence: the number of times the individual has been the victim of a shooting incident; the number of times the individual has been the victim of an aggravated battery or assault; the number of prior arrests for violent offenses; the number of prior arrests for narcotics offenses; the number of prior arrests for unlawful use of a weapon; age as of the most recent arrest; gang affiliation; and trends in recent criminal activity.[127] SSL explicitly excludes race and gender as algorithmic inputs.[128]

Pretrial Decisions: In the context of the pretrial system, the most commonly used predictive algorithm is the Public Safety Assessment (PSA) tool created by Arnold Ventures, formerly the Laura and John Arnold

---

122.   Emily Thomas, *Why Oakland Police Turned Down Predictive Policing*, VICE (Dec. 28, 2016, 9:00 AM), https://www.vice.com/en_us/article/ezp8zp/minority-retort-why-oakland-police-turned-down-predictive-policing [https://perma.cc/26S8-4VHP]; Caroline Haskins, *Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed*, VICE (Feb. 14, 2019, 12:57 PM), https://www.vice.com/en_us/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed [https://perma.cc/AC82-B8HP].

123.   *Predictive Policing: Guidance on Where and When to Patrol*, PREDPOL, https://www.predpol.com/how-predictive-policing-works [https://perma.cc/NDN8-ETZ2].

124.   *Id.*

125.   *See* Asher & Arthur, *supra* note 5; *see also* Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put into Practice: A Quasi-experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347 (2016).

126.   *See* Asher & Arthur, *supra* note 5.

127.   *Strategic Subject List*, CHI. DATA PORTAL, https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np [https://perma.cc/EQG7-A8MV].

128.   *Id.*

Foundation, to predict the risk of pretrial misconduct.[129] The PSA has been rapidly adopted by at least forty jurisdictions to date, including Charlotte, Chicago, and Phoenix,[130] and promises to be one of the most influential criminal justice developments of the recent era. The PSA predicts the likelihood that an individual will be rearrested for a new crime if released before trial, as well as the likelihood that he or she will not return for a future court hearing. The PSA also identifies defendants with a high risk of being rearrested for a violent crime.[131]

The PSA currently uses nine inputs to predict each outcome of interest: age at current arrest; the pending charge at the time of the offense; whether the current charge is for a violent offense; whether the individual has a prior misdemeanor conviction; whether the individual has a prior felony conviction; whether the individual has a prior violent conviction; whether the individual has a prior failure to appear in the past two years; whether the individual has a prior failure to appear older than two years; and whether the individual has a prior incarceration spell.[132] The PSA explicitly excludes inputs such as race, gender, education, socioeconomic status, and neighborhood of residence.[133]

In creating the PSA, Arnold Ventures wanted to create an objective and fair pretrial decision tool, which it interpreted as "meaning that [the tool] should not contain factors that would lead defendants to be treated differently because of their race, gender, or socioeconomic status."[134] In addition, Arnold Ventures has stated that "[t]o design a risk assessment that violated any of these principles would not only conflict with our shared values of fairness and justice, in addition to the law, but would also do nothing to enhance the predictive accuracy of risk assessments."[135] Citing

---

129.   Emily Hamer, *Controversial Algorithms Help Decide Who Stays in Jail*, ASSOCIATED PRESS (Feb. 17, 2019), https://apnews.com/794c27c772ae4450acb978f4b84f4619 [https://perma .cc/792L-GMZS].

130.   *Pretrial Risk Assessment Now Available to All Interested Jurisdictions; Research Advisory Board Announced*, ARNOLD VENTURES (July 11, 2018), https://www.arnoldventures.org /newsroom/laura-and-john-arnold-foundation-makes-pretrial-risk-assessment-available-to- all-jurisdictions-announces-expert-panel-to-serve-as-pretrial-research-advisory-board [https://perma.cc/64C4-TETW]; *21 Cities, States Adopt Risk Assessment Tool to Help Judges Decide Which Defendants to Detain Prior to Trial*, ARNOLD VENTURES (June 26, 2015), https://www.arnoldventures.org/newsroom/more-than-20-cities-and-states-adopt-risk- assessment-tool-to-help-judges-decide-which-defendants-to-detain-prior-to-trial [https://perma.cc/J874-Q74U].

131.   *About the Public Safety Assessment*, ADVANCING PRETRIAL POL'Y & RSCH., https://advancingpretrial.org/psa/factors/ [https://perma.cc/VWW3-NJRX].

132.   *Id.*

133.   *See id.*

134.   Anne Milgram, Alexander M. Holsinger, Marie Vannostrand & Matthew W. Alsdorf, *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making*, 27 FED. SENT'G REP. 216, 220 (2015).

135.   *Id.*

research that shows that race and gender are not the best predictors of pretrial risk,[136] Arnold Ventures concludes that "there is simply no need to choose between the predictive accuracy of a risk assessment and the fair treatment of all individuals, regardless of race, gender, or socioeconomic status."[137]

There are also versions of pretrial risk-assessment tools that are used by just one city or state. One of the earliest is the Virginia Pretrial Risk Assessment Instrument (VPRAI), developed by the Virginia Department of Criminal Justice Services in 2003.[138] The VPRAI calculates the risk of pretrial misconduct using eight factors: whether the current charge is a felony; whether the defendant has another pending charge; the defendant's criminal history; whether the defendant has two or more failures to appear; whether the defendant has two or more violent convictions; whether the defendant lived at the current residence for less than one year; whether the defendant was employed at the time of arrest; and whether the defendant has a history of drug abuse.[139] These factors are then converted into a risk level, which is used as an input into the Praxis decisionmaking tool that provides recommendations for release and detention, as well as the appropriate terms of pretrial supervision.[140] Factors like race and gender are not included.

Sentencing: Risk-assessment tools are also commonly used at sentencing. One of the first risk-assessment tools used at sentencing was developed by the Virginia Sentencing Commission in 1995, known as the Nonviolent Risk Assessment (NVRA).[141] The risk-assessment tool was mandated by the Virginia General Assembly, with the goal of diverting 25% of nonviolent offenders to alternative sanctions in lieu of incarceration by identifying low-risk individuals.[142] Prior to 2012, the Commission included eleven factors to predict recidivism, including gender, age, marital status, employment status,

---

136. *Id.* (citing K Bechtel, Christopher T. Lowenkamp & Alex Holsinger, *Identifying the Predictors of Pretrial Failure: A Meta-analysis*, FED. PROB., Sept. 2011, at 129, 132–33, and Marie VanNostrand & Gena Keebler, *Pretrial Risk Assessment in the Federal Court*, FED. PROB., Sept. 2009, at 5, 18).

137. *Id.*

138. *Common Pretrial Risk Assessments*, MAPPING PRETRIAL INJUSTICE, https://pretrialrisk.com/the-basics/common-prai/ [https://perma.cc/4ZYN-E8YQ].

139. MONA J.E. DANNER, MARIE VANNOSTRAND & LISA M. SPRUANCE, RACE AND GENDER NEUTRAL PRETRIAL RISK ASSESSMENT, RELEASE RECOMMENDATIONS, AND SUPERVISION: VPRAI AND PRAXIS REVISED 4 (2016), https://university.pretrial.org /HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=7ebee4a7-4bde-62f5-c031-6a3df7a4bc13 [https://perma.cc/BT98-XBYF].

140. *Id.* at 1.

141. Brandon Garrett & John Monahan, *Assessing Risk: The Use of Risk Assessment in Sentencing*, JUDICATURE, Summer 2019, at 42, 45.

142. *See* BRIAN J. OSTROM, MATTHEW KLEIMAN, FRED CHEESMAN, II, RANDALL M. HANSEN & NEAL B. KAUDER, NAT'L CTR. FOR STATE CTS., OFFENDER RISK ASSESSMENT IN VIRGINIA 9, 17 (2002).

current   offense   information,   prior   record,   and   prior   juvenile
incarceration.[143]

In 2012, the Virginia Sentencing Commission revised the risk-
assessment instrument using data on eligible drug and property offenders,
and the instrument is currently administered only to offenders who would
otherwise be recommended for incarceration under the state's sentencing
guidelines.[144] As part of this retesting, the Commission further restricted the
factors used to predict risk, eliminating factors such as employment status
and marital status.[145]

The Commission also originally found that race was highly predictive of
recidivism.[146] However, it chose to exclude race from the risk assessment; it
viewed including race as "inappropriate" because "race was 'standing in' for
other factors that are difficult, and often impossible, to measure. . . . [such as]
economic deprivation, inadequate educational facilities, family instability,
and limited employment opportunities, many of which disproportionately
apply   to   the   African-American   population."[147]   Interestingly,   the
Commission noted that by excluding race, the "procedure inevitably led to
the loss of some predictive efficiency."[148]

In the past several years, other state legislatures and sentencing
commissions have expressed growing interest in the use of algorithms at
sentencing and have begun developing their own risk-assessment tools. For
example, the Pennsylvania legislature mandated the development of a risk-
assessment sentencing tool in a 2010 senate bill in an effort to reduce the
increasing prison populations by diverting low-risk offenders out of prison.
In developing its proposal—which has not yet been enacted—the
Pennsylvania Sentencing Commission considered including factors such as
age, gender, and the number and types of prior convictions.[149] Importantly,

---

143.   *Id.* at 27.

144.   John Monahan, Anne L. Metz & Brandon L. Garrett, *Judicial Appraisals of Risk As-
sessment in Sentencing*, 36 BEHAV. SCI. & L. 565, 567 (2018).

145.   *See* Garrett & Monahan, *supra* note 141 ("In 2012, the commission revised and re-
validated the NVRA on large samples of eligible drug and property offenders. For example, the
NVRA for the crime of larceny now consists of five risk factors: offender age at the time of the
offense; gender; prior adult felony convictions; prior adult incarcerations; and whether the of-
fender was legally restrained (e.g., on probation) at the time of the offense.").

146.   *See* OSTROM ET AL., *supra* note 142, at 27–28.

147.   *Id.*

148.   *Id.* at 28 n.10.

149.   *See* PA. COMM'N ON SENT'G, VALIDATION OF A RISK ASSESSMENT INSTRUMENT BY
OFFENSE GRAVITY SCORE FOR ALL OFFENDERS 1–2 (2016), http://pcs.la.psu.edu/publications-
and-research/research-and-evaluation-reports/risk-assessment/phase-ii-reports/interim-
report-2-validation-of-risk-assessment-instrument-by-ogs-for-all-offenses-february-2016/view
[https://perma.cc/BK6U-F8EK].

the Pennsylvania Sentencing Commission purposely excluded the use of race in its risk-assessment tool.[150]

   Parole: There are several generic risk-assessment tools designed for parole decisions, with many of these tools subsequently adapted for sentencing decisions as well. The most commonly used risk-assessment instrument in this context is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is used in many states across the country to assist in the placement and management of offenders.[151] Developed by a company called Northpointe (recently renamed Equivant), the COMPAS system uses answers from a 137-item questionnaire to predict the risk of committing a new crime within two years and then classifies offenders on a scale of one through ten.[152] Broadly speaking, these factors include questions regarding current charges, criminal history, history of noncompliance on probation or parole, family and peers, residential stability, education, employment, and traits such as anger and criminal

---

   150.   *See* Mitch Smith, *In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures*, N.Y. TIMES (June 22, 2016), https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html [https://perma.cc/5HCJ-B7YD]. Interestingly, the Pennsylvania Commission's interim report suggests that race may not be fully excluded in a statistical sentence, noting that "[w]hile race and county were found to be significant predictors of recidivism, they are not included in the risk scale. They are, however, statistically controlled for in the analyses, which means that the effects of the other factors are included only after eliminating the effects of race and county." PA. COMM'N ON SENT'G, *supra* note 149, at 12 n.8.

   151.   Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/NE4G-R4KF]. In recent years, the COMPAS algorithm has faced intense public scrutiny. In 2016, a ProPublica report analyzed the risk predictions on arrestees from Broward County, Florida and alleged that the COMPAS algorithm was biased against Black defendants, and specifically that Black individuals are almost twice as likely as white individuals to be labeled a higher risk but not actually reoffend. *See id.* Although this allegation was challenged by both the algorithm's creator Northpointe and various academics, who noted that the algorithm exhibited similar rates of recidivism among white and Black offenders who received the same score (i.e., equal predictive accuracy), the ProPublica story generated a large debate about the appropriate use of such algorithms in the criminal justice system and a discussion on competing notions of algorithmic fairness. Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016, 5:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ (on file with the *Michigan Law Review*); *see also* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807v2.pdf [https://perma.cc/3NVL-QVHW].

   152.   Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, SCI. ADVANCES, Jan. 17, 2018, at 1, 1; Corbett-Davies et al., *supra* note 151.

attitudes.[153] While the algorithm used by COMPAS is proprietary, it is known that COMPAS does not use an offender's race in generating predictions, although other demographic characteristics such as age and gender are used.[154]

A second commonly used risk-assessment instrument is the Level of Service Inventory Revised (LSI-R). Developed in the mid-1980s, the LSI-R is frequently used at both sentencing and probation stages of the criminal justice system to "guide sentencing decisions, placement in correctional programs, institutional assignments, and release from institutional custody."[155] The LSI-R uses fifty-four factors in the "areas of Criminal History, Education and Employment, Financial, Family, Accommodations, Leisure and Recreation, Companions, Alcohol and Drugs, Emotional and Personal Issues, and Attitudes and Orientation."[156] These factors then generate a risk prediction of each offender's likelihood of recidivism. Gender and race/ethnicity are not included among the various risk factors.[157]

A third commonly used parole risk-assessment tool is the Salient Factor Score (SFS), originally created by the U.S. Parole Commission for use in federal parole guidelines.[158] Designed to predict the risk of future offending, the most current iteration (issued in 1991) of the SFS includes factors such as prior convictions, incarcerations, age at commencement of current commitment, recent commitment-free period, parole revocation, and custody status.[159] Importantly, however, the creators of the SFS were concerned about fairness and deliberately chose to exclude characteristics that were deemed unfair. For example, gender and race were excluded from

---

153.   *See Sample COMPAS Risk Assessment: COMPAS "CORE,"* PROPUBLICA (May 23, 2016),   https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html [https://perma.cc/TF5W-492Y].

154.   Corbett-Davies et al., *supra* note 151; *see, e.g.*, Dressel & Farid, *supra* note 152, at 1 ("Although the data used by COMPAS do not include an individual's race, other aspects of the data may be correlated to race that can lead to racial disparities in the predictions.").

155.   CHRISTOPHER T. LOWENKAMP & EDWARD J. LATESSA, VALIDATING THE LEVEL OF SERVICE INVENTORY REVISED IN OHIO'S COMMUNITY BASED CORRECTIONAL FACILITIES 5–6, https://www.uc.edu/content/dam/uc/ccjr/docs/reports/project_reports/OHIOCBCFLSI-R.pdf [https://perma.cc/X9DL-S335]. Importantly, however, the creators of the LSI-R have noted that their risk-assessment tool "is not a comprehensive survey of mitigating and aggravating factors relevant to criminal sanctioning and was never designed to assist in establishing the just penalty." Malenchik v. State, 928 N.E.2d 564, 572 (Ind. 2010) (quoting D.A. ANDREWS & JAMES L. BONTA, THE LEVEL OF SERVICE INVENTORY–REVISED USER'S MANUAL 3 (2001)).

156.   *Malenchik*, 928 N.E.2d at 567; Alexander M. Holsinger, Christopher T. Lowenkamp & Edward J. Latessa, *Ethnicity, Gender, and the Level of Service Inventory-Revised*, 31 J. CRIM. JUST. 309, 310 (2003) (describing the LSI-R).

157.   Holsinger et al., *supra* note 156, at 312–13.

158.   *See* Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 FED. SENT'G REP. 167, 168 (2014).

159.   *Id.* at 168 tbl.1.

the SFS even though doing so weakened predictive accuracy.[160] Characteristics such as current age, employment, education, residential status, and family characteristics were initially included in earlier versions of the SFS but eventually discarded because they were deemed "heavily correlated with race," with the U.S. Parole Commission deciding that their use would be "unjust."[161]

    Risk Assessments in Other Contexts: Risk-assessment instruments are also increasingly common in a number of related contexts. For example, many jurisdictions are now using predictive algorithms to predict the risk of future violence in both criminal and civil settings. One prominent example is the Classification of Violence Risk ("COVR"), which was constructed using data from the MacArthur Violence Risk Assessment Study to predict the risk of future violence for individuals with mental disorders.[162] The study collected information on 134 risk factors on over 1,000 patients in acute civil psychiatric institutions, who were then followed after their discharge from the hospital.[163] These risk factors included characteristics such as the seriousness and frequency of past requests, age, gender, unemployment, and diagnosis of illnesses like antisocial personality disorder and schizophrenia.[164] Using these inputs, MacArthur researchers placed patients into one of five risk categories using a "classification tree" methodology.[165] The MacArthur researchers explicitly excluded race from the algorithm "[t]o avoid any possible misinterpretation of our risk assessment procedures as a form of 'racial profiling.'" The researchers also note that "[t]he revised models without race differed only trivially in accuracy from the original ones that included race."[166]

    A number of jurisdictions are also beginning to use predictive algorithms to identify children who are at risk of abuse and neglect. For example, in August 2016, the Allegheny County Department of Human Services implemented the Allegheny Family Screening Tool (AFST), a predictive algorithm to improve call-screening decisionmaking in the

---

160.   *Id.* at 172.

161.   *Id.* at 168; *see also* Peter B. Hoffman, *Screening for Risk: A Revised Salient Factor Score (SFS 81)*, 11 J. CRIM. JUST. 539 (1983).

162.   *See, e.g.*, Paul S. Appelbaum, Pamela Clark Robbins & John Monahan, *Violence and Delusions: Data from the MacArthur Violence Risk Assessment Study*, 157 AM. J. PSYCHIATRY 566, 566 (2000).

163.   *The MacArthur Violence Risk Assessment Study*, MACARTHUR RSCH. NETWORK ON MENTAL HEALTH & L. (Apr. 2001), https://macarthur.virginia.edu/risk.html [https://perma.cc/G272-3C4T].

164.   See JOHN MONAHAN, HENRY J. STEADMAN, ERIC SILVER, PAUL S. APPELBAUM, PAMELA CLARK ROBBINS, EDWARD P. MULVEY, LOREN H. ROTH, THOMAS GRISSO & STEVEN BANKS, RETHINKING RISK ASSESSMENT: THE MACARTHUR STUDY OF MENTAL DISORDER AND VIOLENCE 134 (2001) for a detailed description of method.

165.   *Id.* at 115, 124.

166.   *Id.* at 119 n.1.

county's child welfare system.[167] The AFST includes factors such as criminal history in the predictive algorithm,[168] but race is explicitly excluded as an input. Government reports justified the exclusion of race by explaining that "in conjunction with the researchers' finding that including race in the model did not significantly improve its accuracy, administrators, in conjunction with ethics and legal staff, determined that race would be omitted as a factor for determining the risk score."[169]

B. *Summary of Predictive Algorithms in the Criminal Justice System*

In this Section, we summarize the findings from our survey of commonly used algorithms in the criminal justice system. Specifically, we summarize how each of the algorithms appears to deal with race and racial proxies. Table 1 lists each of these commonly used predictive algorithms. For each algorithm, we list the setting in which the algorithm is generally employed, whether race is excluded as an input, and whether some notable nonrace correlates are excluded as inputs.

TABLE 1: PREDICTIVE ALGORITHMS IN THE CRIMINAL JUSTICE SYSTEM

Algorithm	Setting	Excludes Race	Excludes Education	Excludes Employment	Excludes Past Criminal History
1. PredPol	Policing	Yes	Yes	Yes	No
2. SSL	Policing	Yes	Yes	Yes	No
3. PSA	Pretrial	Yes	Yes	Yes	No
4. VPRAI	Pretrial	Yes	Yes	No	No
5. VA NVRA (pre-2012)	Sentencing	Yes	Yes	No	No
6. COMPAS	Sentencing & Parole	Yes	No	No	No
7. LSI-R	Sentencing & Parole	Yes	No	No	No
8. SFS	Parole	Yes	Yes	Yes	No

Note: This table summarizes the most commonly used predictive algorithms in the criminal justice system and how they deal with both race and nonrace correlates. See the text for additional details.

---

167. HORNBY ZELLER ASSOCS., INC., ALLEGHENY COUNTY PREDICTIVE RISK MODELING TOOL IMPLEMENTATION: PROCESS EVALUATION 3, 7 (2018).

168. TIM DARE & EILEEN GAMBRILL, ETHICAL ANALYSIS: PREDICTIVE RISK MODELS AT CALL SCREENING FOR ALLEGHENY COUNTY (2017).

169. HORNBY ZELLER ASSOCS., INC., *supra* note 167, at 7; DARE & GAMBRILL, *supra* note 168.

Based on our review, none of the most commonly used predictive algorithms in the criminal justice system directly use race as an input.[170] The universal approach is to explicitly exclude race as an algorithmic input, with some recognition that accuracy is reduced as a result. We view the exclusion of a race in all of these commonly used predictive algorithms as a consistent extension of the mainstream legal position that including race would likely be unconstitutional.[171] The decision to exclude race as an algorithmic input, despite the lack of settled legal precedent on the issue, is likely because the "[e]xplicit use of race, ethnicity, or religion . . . is widely regarded as unseemly."[172] As some have argued, the exclusion of race from predictive algorithms in the criminal justice system "appears to reflect corporate risk aversion, not an effort at legal compliance."[173] We believe that explicit exclusion of race is also likely due to a perception that inclusion would violate antidiscrimination law, as reviewed in Section I.A. As Deborah Hellman has argued, "algorithms are designed to be 'race blind' because their designers, as well as many legal scholars, assume that use of racial classifications within algorithms is legally prohibited."[174]

Predictive algorithms in the criminal justice system are much more varied in how they deal with nonrace correlates and the proxy effects of race. Six of the predictive algorithms we reviewed exclude at least employment or education, two nonrace correlates that commentators have expressed growing concerns with due to racial proxy effects, while the remaining algorithms do not explicitly exclude these nonrace correlates. PredPol, for example, uses "[n]o demographic, ethnic or socio-economic infor-

---

170.   The stance towards other protected characteristics, such as gender, is more varied, with some risk-assessment instruments explicitly including gender and others explicitly excluding gender. *Compare, e.g.*, Dressel & Farid, *supra* note 152 (COMPAS), *with* CHI. DATA PORTAL, *supra* note 127 (SSL).

171.   *See* Tonry, *supra* note 158, at 169. Despite what he perceives as "toothless" legal constraints, Tonry notes that "[r]ace, ethnicity, and religion are not to my knowledge anywhere used as an explicit factor in prediction instruments or in sentencing or parole policies" because "[e]xplicit use of race, ethnicity, or religion . . . is widely regarded as unseemly, and so the issue is unlikely to arise." *Id.* at 169, 170; *see also* Luis Daniel, *The Dangers of Evidence-Based Sentencing*, GOVLAB (Oct. 31, 2014), http://thegovlab.org/the-dangers-of-evidence-based-sentencing/ [https://perma.cc/AN8P-MBCW] ("Overwhelmingly, states do not include race in the risk assessments since there seems to be a general consensus that doing so would be unconstitutional."); Berk & Hyatt, *supra* note 97, at 227 ("[A]ctuarial methods need not include race as a predictor, and to the best of our knowledge, most do not."); Nicholas Scurich & John Monahan, *Evidence-Based Sentencing: Public Openness and Opposition to Using Gender, Age, and Race as Risk Factors for Recidivism*, 40 LAW & HUM. BEHAV. 36, 37 (2016) ("No risk assessment instrument explicitly includes race as a risk factor in sentencing . . . .").

172.   Tonry, *supra* note 158, at 170.

173.   Huq, *supra* note 21, at 1079. But as he notes, "[c]urrent law does not address whether the availability of race as an input into the deliberative process that results in state action violates the Equal Protection Clause on anticlassification grounds." *Id.* at 1097–98.

174.   Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 848 (2020).

mation . . . . This eliminates the possibility for privacy or civil rights violations."[175] The Arnold Ventures PSA also takes a principled stance against using any "factors that would lead defendants to be treated differently because of their race, gender, or socioeconomic status."[176] Based on this stance, the PSA excludes both race and nonrace correlates such as education, socioeconomic status, and neighborhood of residence.[177] The SFS also explicitly excludes characteristics such as age, employment, education, residential status, and family characteristics precisely because they were deemed "heavily correlated with race" such that their inclusion would be "unjust."[178]

We note that, interestingly, some of the reviewed algorithms have over time excluded more nonrace correlates that have the potential to generate racial proxy effects, potentially reflecting the mainstream position we discussed in Section I.B. For example, the current iteration of the SFS (1991) does not include employment or education, but earlier versions did include these inputs.[179] Similarly, the pre-2012 Virginia NVRA included factors such as employment and marital status, but these factors were removed after the post-2012 revision of the NVRA.[180]

In contrast, some of the predictive algorithms use many input factors that are likely to generate racial proxy effects, including employment, education, and other measures of socioeconomic status. As one example, COMPAS uses information regarding family and peers, residential stability, education, employment, and traits such as anger and criminal attitudes, all of which are likely to be correlated with race.[181]

There is also considerable variation in which nonrace correlates are considered problematic, with no clear principle guiding the choice of these nonrace correlates. Across algorithms, what governs why some algorithms use factors like education and employment while others reject them? Within the same algorithm, what governs the choice to use certain racial proxies while excluding other racial proxies? Specifically, while some of the above algorithms exclude factors like education or employment out of a view that these proxy effects are unfair, they also universally include characteristics related to the current offense or the defendant's criminal history, or measures of past crime in an area. It is almost certainly the case that past criminal history is a highly predictive measure of future recidivism, suggesting that predictive accuracy is a key concern to algorithmic designers.

---

175.   *Predictive Policing: Guidance on Where and When to Patrol*, *supra* note 123.

176.   Milgram et al., *supra* note 134, at 220.

177.   *Id.*

178.   Tonry, *supra* note 158.

179.   *Id.*

180.   *See* VA. CRIM. SENT'G COMM'N, 2012 ANNUAL REPORT 37–38, 46–48 (2012); Garrett & Monahan, *supra* note 141.

181.   *See* Angwin et al., *supra* note 151.

But as many have pointed out, race and measures of socioeconomic status are also highly predictive factors, and yet are often excluded.[182]

We view the universal inclusion of measures of past criminal history as consistent with the mainstream position that these inputs are legally permissible and valid. But, as we noted previously, it is almost certainly the case that current offense and prior criminal history are highly correlated with race. If an individual's current offense or prior criminal history are driven, for example, by racial biases in policing, then including these inputs in the algorithm may lead to predictions that are also racially biased and can result in what many perceive to be an unfair algorithm.

In summary, the most commonly used predictive algorithms in the criminal justice system exhibit two features relevant to our analysis. First, these algorithms follow an exclusionary approach when dealing directly with race, omitting race as an input regardless of whether race improves the accuracy of the underlying predictions. Second, these algorithms take a very haphazard approach to dealing with nonrace correlates and proxy effects, sometimes excluding inputs deemed to be correlated with race out of fairness concerns (even if a loss to accuracy) yet also retaining others that are also likely correlated with race, including in particular current offense and criminal history.

III.   A STATISTICAL FRAMEWORK FOR PREDICTIVE ALGORITHMS

In this Part, we provide a simple statistical framework that formalizes the mainstream legal consensus outlined in Part I. We then use this framework to illustrate how the direct and proxy effects of race can lead to algorithmic predictions that disadvantage one group relative to another. We illustrate these direct and proxy effects through the use of simple examples, showing exactly how both direct use of race and indirect use of nonrace correlates can generate unwarranted disparities.

A.   *Categorizing Algorithmic Inputs*

We begin by categorizing the potential algorithmic inputs into three mutually exclusive categories: (1) protected characteristics, (2) correlates of protected characteristics, and (3) noncorrelates of protected characteristics. This simple categorization will allow us to both formalize the mainstream legal consensus described in Part I and illustrate how the direct and proxy effects of race impact predictive algorithms. The definition of each category of algorithmic input is as follows.

Protected Characteristics: The first set of potential algorithmic inputs consists of protected characteristics, denoted by $X^{Protected}$. By definition, protected characteristics are algorithmic inputs that trigger heightened

---

scrutiny under the Equal Protection Clause, including both suspect and quasi-suspect classes. Examples include race, national origin, religion, and gender. We will focus on race as our canonical example of a protected characteristic in all our theoretical and empirical exercises moving forward, but all of our results are easily extended to consider other protected characteristics.[183]

Correlates of Protected Characteristics: The second set of potential algorithmic inputs consists of correlates of protected characteristics, denoted by $X^{Correlated}$. Correlated characteristics include all algorithmic inputs that are correlated with protected characteristics such as race. In the context of race and the criminal justice system, these nonrace correlates may include zip code of residence, education level, and employment status.[184] However, whether an algorithmic input is actually correlated with race is an empirical question that may vary across contexts.

Noncorrelates of Protected Characteristics: The third and final set of algorithmic inputs we consider consists of inputs that are not correlated with protected characteristics, denoted by $X^{Uncorrelated}$. For simplicity, we assume that $X^{Uncorrelated}$ are also uncorrelated with $X^{Correlated}$, but all of our results are easily extended to allow for some correlation between $X^{Uncorrelated}$ and $X^{Correlated}$. In the context of race and the criminal justice system, these uncorrelated characteristics may include criminal history, which some have argued is not a proxy for race.[185] However, as above, whether an algorithmic input is actually uncorrelated with race is an empirical question that may vary across contexts.

B. *Benchmark Statistical Model*

Let the statistical relationship between the outcome of interest and the full set of observable potential algorithmic inputs be equal to:

$$Y_i = \beta_0 + \beta_1 \cdot X_i^{Uncorrelated} + \beta_2 \cdot X_i^{Correlated} + \beta_3 \cdot X_i^{Protected} + \epsilon_i$$

(Equation 1)

where, for simplicity, we assume that each set of input characteristics enters linearly and additively. We begin with the most simple statistical framework for two main reasons. First, this linear framework helps to clearly illustrate the key concepts of this Article. Second, in the context of the criminal justice system, the focus of our paper, commonly used algorithms are in practice created using linear and additive statistical models, where each risk factor is

---

183.   *See infra* Section VI.A.

184.   *See, e.g.*, Starr, *supra* note 12, at 838 ("[S]ocioeconomic and family variables that [the instruments] include are highly correlated with race, as is criminal history, so they are likely to have a racially disparate impact.").

185.   Skeem & Lowenkamp, *supra* note 13.

associated with a numeric value and numeric values are summed to create a final "risk score." Thus, we view this simple linear model as reflective of current practice in this setting. We consider extensions to this framework in Part VI.

Under the statistical relationship in Equation 1, Y_i is the observed outcome for individual i. In the criminal justice context, this could be, for example, the likelihood that an individual would fail to appear at a future court appearance or that she would commit a crime in the future. $X_i^{Protected}$ includes all protected characteristics, $X_i^{Correlated}$ includes all correlated input characteristics, and $X_i^{Uncorrelated}$ includes all uncorrelated input characteristics. ϵ_i is an error term that includes both idiosyncratic noise and systematic unobserved characteristics of individual i. This error term represents the part of the outcome that is left unexplained by the full set of observable potential inputs.

In our statistical framework, β_0 represents the constant term, and β_1, β_2, and β_3 represent the predictive relationship between each set of potential algorithmic inputs and the outcome of interest. We assume that $\beta_1 \neq 0$, $\beta_2 \neq 0$, and $\beta_3 \neq 0$, such that each set of potential inputs has predictive power for the outcome of interest. In other words, we take as given that each category of potential algorithmic inputs helps predict the outcome of interest, holding aside the question of legal permissibility.

Following the definition of the potential algorithmic inputs outlined above, $X_i^{Correlated}$ is the set of potential inputs that is correlated with the set of protected characteristics $X_i^{Protected}$. To allow for this correlation, we assume that the relationship between $X_i^{Correlated}$ and $X_i^{Protected}$ is equal to:

$$X_i^{Protected} = \alpha_0 + \alpha_{Corr} \cdot X_i^{Correlated} + v_i$$

(Equation 2)

where α_{Corr} represents the relationship between $X_i^{Correlated}$ and $X_i^{Protected}$ and v_i is an error term. If $\alpha_{Corr} > 0$, this would indicate that $X_i^{Correlated}$ is positively correlated with $X_i^{Protected}$.

C.   *The Direct and Proxy Effects of Algorithmic Inputs*

We can now formalize how the direct and proxy effects of race can lead to algorithmic predictions that disadvantage one group relative to another. We will establish two important facts in this Section: (1) including a protected characteristic such as race will lead to predictions that allow for the direct effects of race, generating unwarranted disparities under the mainstream legal position; (2) including correlated characteristics will lead to predictions that allow for the indirect effects of race through proxy effects, even when race itself is excluded, again generating unwarranted disparities under the mainstream legal position. By design, we allow all correlated characteristics to have the potential to generate racial proxy effects that many would argue are unwarranted. This position is broad in defining all

proxy effects as unwarranted, but we view this choice as most consistent with the mainstream legal consensus[186] and principled because it does not rely on an ad hoc classification of which types of racial proxies are socially acceptable and which are socially unacceptable.[187] We will also illustrate these ideas by means of simple examples, showing exactly how both direct use of race and indirect use of nonrace correlates can generate unwarranted disparities.

Direct Effects and Unwarranted Disparities: The first important fact illustrated by our statistical framework is that including a protected characteristic such as race will lead to predictions that allow for the direct effects of race, generating unwarranted racial disparities.

To form predictions that incorporate the direct effects of protected characteristics such as race, we estimate the following statistical relationship using a standard ordinary least squares (OLS) regression:[188]

$$Y_i = \beta_0 + \beta_1 \cdot X_i^{Uncorrelated} + \beta_2 \cdot X_i^{Correlated} + \beta_3 \cdot X_i^{Protected} + \epsilon_i$$

(Equation 3)

With all inputs included in the regression, the estimated coefficients yield uncontaminated (in the statistical sense) estimates of $\beta_0, \beta_1, \beta_2,$ and β_3. We can then form the following prediction:

$$\hat{Y}_i^{Direct} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i^{Uncorrelated} + \hat{\beta}_2 \cdot X_i^{Correlated} + \hat{\beta}_3 \cdot X_i^{Protected}$$

(Equation 4)

where $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are the estimated relationship between each set of algorithmic inputs and the outcome of interest. Equation 3 and Equation 4 together illustrate that algorithmic predictions rely on a two-step procedure. First, we estimate the underlying statistical model in order to obtain our coefficient estimates $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$—the "estimation step." Second, we use those estimated coefficients to predict the outcome for each individual,

---

186.   *See, e.g.*, Mayson, *supra* note 17, at 2224 ("Among racial-justice advocates engaged in the debate, a few common themes have emerged. The first is a demand that race, and factors that correlate heavily with race, be excluded as input variables for prediction." (footnote omitted)).

187.   Specifically, we deviate from a classification scheme used by Pope and Sydnor, which groups inputs into those that are "socially acceptable," "socially unacceptable," and "contentious." *See* Pope & Sydnor, *supra* note 23. As noted by Altenburger and Ho, "such classification can be highly contested." Altenburger & Ho, *supra* note 105, at 118. We share the view of Altenburger and Ho that a "commonsense classification of 'socially acceptable' does not necessarily imply statistical independence. Many predictors that may superficially seem 'socially acceptable' are in fact highly correlated with race." *Id.* at 111.

188.   OLS is one of the most common methods of estimating a linear regression model.

relying on their own values of $X_i^{Uncorrelated}$, $X_i^{Correlated}$, and $X_i^{Protected}$—the "prediction step."

By design, predictions formed using Equation 4 lead to different predictions for otherwise similar individuals who differ only in terms of a protected characteristic. In the context of race and the criminal justice system, suppose that $X_i^{Protected}$ is an indicator variable that is equal to 1 if an individual is Black and equal to 0 if an individual is not Black. If $\hat{\beta}_3 > 0$, then Black individuals will receive higher risk scores than white individuals who are otherwise identical in terms of the other algorithmic inputs.

To provide a concrete illustrative example of these direct effects, suppose that there are one hundred total individuals (fifty Black and fifty white), with the distribution of characteristics as follows in Table 2.

TABLE 2: HYPOTHETICAL EXAMPLE TO ILLUSTRATE THE DIRECT AND PROXY EFFECTS OF RACE

Note: This table presents hypothetical relationships between failure to appear, prior criminal history, and race. Individuals who fail to appear if released are denoted in dark gray, while individuals who will appear at all court appearances are denoted in light gray. Each figure represents five individuals, for a total population of 100 individuals. See the text for additional details.

We are interested in predicting the probability that an individual fails to appear at a required future court appearance. In Table 2, individuals who will fail to appear (FTA) if released are denoted in dark gray, while individuals who will appear at all court appearances are denoted in light gray. In our hypothetical example, we have assumed a positive correlation between an individual being Black and the probability of FTA, as well as a

positive correlation between having a prior criminal record and FTA. These assumptions largely mirror the patterns observed in real-world data, but are not critical to the point we are making here. In the hypothetical example illustrated in Table 2, eight out of every ten Black individuals have a prior criminal history and three out of every ten white individuals have a prior criminal history.

Mapping the example in Table 2 to our statistical framework, Y_i is an indicator variable for FTA, $X_i^{Protected}$ is an indicator equal to 1 if an individual is Black and equal to 0 if an individual is white, and $X_i^{Correlated}$ is an indicator equal to 1 if an individual has a prior criminal history and equal to 0 if an individual does not have a prior criminal history. We first estimate Equation 3, where we control for race through $X_i^{Protected}$ and prior criminal history through $X_i^{Correlated}$. These estimates are reported in Table 3.

TABLE 3: HYPOTHETICAL EXAMPLE OF DIRECT EFFECTS

	Prob of FTA
	(1)
Prior Criminal History	0.541***
	(0.077)
Black	0.330***
	(0.076)
Constant	0.038
	(0.052)
Observations	100
R²	0.576

Note: This Table presents a hypothetical example of the direct effects of race. We report OLS estimates of the relationship between FTA, prior criminal history, and race using the hypothetical data from Table 2. *See supra* Table 2 and accompanying text.

The results from Table 3 reveal that, in our hypothetical example, having a prior criminal history increases the probability of FTA by 54.1 percentage points. Table 3 also shows that there is a direct effect of race in our hypothetical example, with Black individuals having a 33.0 percentage point higher probability of FTA than white individuals. In other words, the predicted relationship between the likelihood of FTA and both race and prior criminal history is:

$$Y_i^{FTA} = 0.038 + 0.541 \cdot X_i^{Correlated\ (Prior\ Criminal\ History)}$$
$$+ 0.330 \cdot X_i^{Protected\ (Black)} + \epsilon_i$$

Given this predicted relationship, allowing for a direct effect of race means that Black individuals will receive a predicted risk score that is 33.0 percentage points higher than white individuals with exactly the same prior

criminal history, at least in our hypothetical example. The possibility that Black individuals will be treated differently than otherwise identical white individuals is at the heart of the mainstream argument that including race in predictive algorithms would constitute a violation of the Equal Protection Clause.[189] To address this legal concern, most if not all predictive algorithms therefore exclude race as an input.[190]

The use of direct effects can result in large racial gaps in predicted risk. If predictions were race neutral, the average predicted risk for white individuals is 0.37 and the average predicted risk for Black individuals is 0.64.[191] When direct effects are used to predict risk, the average predicted risk for white individuals is 0.20 and the average predicted risk for Black individuals is 0.80. Thus, when direct effects are incorporated into algorithmic predictions, Black individuals are disadvantaged relative to white individuals. If release decisions are made on the basis of predictions that incorporate direct effects, fewer Black individuals will be released relative to release decisions made on the basis of predictions that are race neutral.

Proxy Effects and Unwarranted Disparities: The second important fact illustrated by our statistical framework is that including correlated characteristics will lead to predictions that allow for the indirect effects of race through proxy effects, even when race itself is excluded, again generating unwarranted racial disparities.

To form predictions that incorporate the proxy effects of protected characteristics such as race, we estimate the following statistical relationship, omitting the protected characteristics as inputs:

$$Y_i = \gamma_0 + \gamma_1 \cdot X_i^{Uncorrelated} + \gamma_2 \cdot X_i^{Correlated} + \epsilon_i$$

(Equation 5)

We can then form the following prediction:

$$\hat{Y}_i^{Proxy} = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot X_i^{Uncorrelated} + \hat{\gamma}_2 \cdot X_i^{Correlated}$$

(Equation 6)

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the estimated relationship between each set of inputs and the outcome of interest. Again, note that these predictions are formed in

---

189.   See Oleson, *supra* note 59, at 1386, and Kopf, *supra* note 66, at 213, for scholars who argue that direct effects of race should be included because they increase predictive accuracy, a compelling government interest.

190.   *See supra* Part II.

191.   Here, race neutrality is achieved using the proposed "colorblinding-inputs" algorithm described in Section IV.B. Race neutrality using the proposed "minorities-as-whites" algorithm described in Section IV.C would result in averaged predicted risk for white individuals of 0.20 and average predicted risk of Black individuals of 0.53. Both proposed race-neutral algorithms would reduce the disadvantage of Black individuals relative to white individuals, as compared to an algorithm that incorporates direct effects of race.

a two-step procedure that includes a first estimation step and second prediction step.

The coefficient $\hat{\gamma}_2$ estimated in Equation 5, is, in general, *not* identical to the estimated coefficient $\hat{\beta}_2$ estimated in Equation 3. Recall that we have assumed that $X_i^{Correlated}$ is correlated with $X_i^{Protected}$, and that $X_i^{Protected}$ is predictive of the outcome such that $\beta_3 \neq 0$. From these two assumptions, it is straightforward to show that $\hat{\gamma}_2$ will not equal to $\hat{\beta}_2$ due to proxy effects. In other words, because of proxy effects, the predictive relationship between the outcome of interest and the correlates of protected characteristics is not the same depending on whether one includes or excludes the protected characteristics in the estimation process.

The importance of these proxy effects can be expressed in terms of the standard omitted-variable-bias (OVB) formula from the economics literature, which describes the relationship between regression estimates in models with different sets of controls.[192] We can illustrate these proxy effects by substituting the expression for $X_i^{Protected}$ from Equation 2, which showed the statistical relationship between $X_i^{Protected}$ and $X_i^{Correlated}$, into Equation 1, which showed the statistical relationship between Y_i and $X_i^{Uncorrelated}$, $X_i^{Correlated}$, and $X_i^{Protected}$. Doing so yields the following expression:

$$Y_i = (\beta_0 + \beta_3\alpha_0) + \beta_1 \cdot X_i^{Uncorrelated} + (\beta_2 + \beta_3\alpha_{Corr}) \cdot X_i^{Correlated} + (\epsilon_i + \beta_3 v_i)$$

(Equation 7)

The standard OVB formula shows us that $\hat{\gamma}_2$, found in Equation 6, is not a consistent estimate of β_2, found in Equation 1, but rather the expression $(\beta_2 + \beta_3\alpha_{Corr})$. Intuitively, β_2 includes the portion of the correlated characteristics that is orthogonal to (or uncorrelated with) protected characteristics, and $\beta_3\alpha_{Corr}$ includes the portion of the correlated characteristics that is purely a proxy for protected characteristics. One can think of β_2 as capturing predictive variation in the correlated characteristics *within* a protected class, and $\beta_3\alpha_{Corr}$ as the predictive variation in the correlated characteristic *across* protected classes.

The estimated coefficient $\hat{\gamma}_2$ is therefore "contaminated" (again in the statistical sense) by the proxy effect of race, $\beta_3\alpha_{Corr}$. These kinds of proxy effects emerge precisely *because* protected characteristics are excluded from the estimating equation. As a result, the remaining correlated characteristics act as partial proxies for the protected characteristics. Indeed, the OVB formula shows us that proxy effects will emerge anytime $\beta_3 \neq 0$ and $\alpha_{Corr} \neq 0$, or when excluded protected characteristics and included inputs

192. *See, e.g.*, JOSHUA D. ANGRIST & JÖRN-STEFFEN PISCHKE, MOSTLY HARMLESS ECONOMETRICS: AN EMPIRICIST'S COMPANION 59 (2009).

are correlated and the protected characteristics are predictive of the outcome. In fact, one can think of including the protected characteristics in the estimation process as a way to remove the proxy effects from the correlated inputs.

These proxy effects emerge regardless of whether we are identifying the "true" underlying causal relationship between inputs and the outcome of interest. If an algorithm correctly identifies the true underlying causal relationship, then the coefficients obtained from the full specification in Equation 1 can be interpreted as identifying the *causal* effects of each set of inputs. In this scenario, excluding a protected characteristic leads to proxy effects from the correlated inputs. But even if the algorithm does not estimate the true causal relationship, such that Equation 1 does not represent the *causal* effects of each set of inputs, proxy effects still emerge when protected characteristics are omitted. This is because the OVB formula that characterizes the "bias" is a mechanical characterization of the relationship between the coefficents when protected traits are excluded (e.g., Equation 6) versus when protected traits are included (e.g., Equation 2).[193] Thus, the statistical fact of proxy effects is not dependent on whether an algorithm has identified causal estimates, and in fact, predictive algorithms do not seek or claim to be estimating causal relationships. The exercise of prediction is *not* generally about establishing causation.

To provide a concrete illustrative example of these proxy effects, we return to the hypothetical distribution of characteristics described in Table 2. Recall that our hypothetical example assumes a positive correlation between an individual being Black and the probability of FTA, as well as a positive correlation between having a prior criminal record and FTA. We also assume a positive correlation between having a prior criminal record and being Black, which is what leads to the emergence of proxy effects in our hypothetical example. A visual illustration of proxy effects in this hypothetical can be seen here:

---
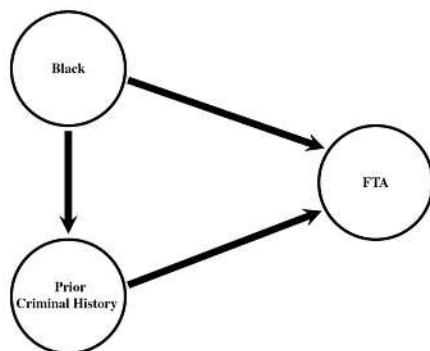
193.   *See, e.g.*, *id.* at 59 ("In fact, the OVB formula is a mechanical link between coefficient vectors that applies to short and long regressions whether or not the longer regression is causal.").

The above diagram presents a direct pathway between race and FTA and a direct pathway between prior criminal history and FTA. But the diagram also illustrates that there is a pathway between race and prior criminal history, which could reflect, for example, discriminatory policing. In theory, we want to capture only the direct pathway between prior criminal history and FTA, which we can obtain by including race in the estimation equation. But when we exclude race, prior criminal history enters into our estimation through a direct pathway and a proxy pathway.

To see how these proxy effects affect algorithmic predictions, Table 4 presents estimates from a series of OLS regressions of FTA on possible inputs using the hypothetical relationships described in Table 2. Column 1 of Table 4 controls only for prior criminal history, excluding race following mainstream practice. In this specification, the estimated coefficient on prior criminal history is equal to $\beta_2 + \beta_3 \alpha_{Corr}$, where $\beta_3 \alpha_{Corr}$ is the proxy effect of race. Column 2 adds an indicator for an individual being Black versus white, resulting in an estimated coefficient on prior criminal history that only reflects the race-orthogonal (or race-independent) predictive relationship between that input and the probability of FTA, β_2.

The results from Table 4 show that the proxy effects of race inflate the coefficient on prior criminal history, such that individuals with a prior criminal history will receive a predicted risk that is 70.7 percentage points higher than individuals with no prior criminal history. Recall that the race-independent predictive relationship is only 54.1 percentage points, meaning that the proxy effects of race add 16.6 percentage points to this estimated coefficient. As a result, the inflated coefficient on the prior criminal history variable will result in Black individuals receiving, on average, higher risk predictions due to the positive correlation between race and prior criminal history. Intuitively, this occurs because the predictive weight on criminal history will be overweighted relative to the race-independent predictive relationship when there are proxy effects. This inflation leads individuals with a criminal history to be penalized relative to those without a criminal history, and Black individuals are more likely to have criminal histories. Thus, because of proxy effects, membership in a racial group can still indirectly affect algorithmic predictions even when race itself is excluded as an input.

These proxy effects can also lead to racial gaps in predicted risk. Recall that if predictions were race neutral, the average predicted risk for white individuals is 0.37 and the average predicted risk for Black individuals is 0.64. When proxy effects are used to predict risk (even when direct effects are excluded), the average predicted risk for white individuals is 0.32 and the average predicted risk for Black individuals is 0.67. Thus, proxy effects in algorithmic predictions also disadvantage Black individuals relative to white individuals.

TABLE 4: HYPOTHETICAL EXAMPLE OF PROXY EFFECTS

	Prob of FTA		
	Proxy Effects	No Proxy Effects	Difference (1) - (2)
	(1)	(2)	(3)
Prior Criminal History	0.707*** (0.072)	0.541*** (0.077)	0.166*** (0.059)
Black		0.330*** (0.076)	
Constant	0.111** (0.054)	0.038 (0.052)	
Observations	100	100	---
R^2	0.494	0.576	---

Note: This Table presents a hypothetical example of the proxy effects of race. We report OLS estimates of the relationship between FTA, prior criminal history, and race using the hypothetical data from Table 2. *See supra* Table 2 and accompanying text.

Summary: We have demonstrated that the use of individual race can lead to direct effects that result in unwarranted disparities. We have also shown that excluding race but including any race correlate can lead to substantial proxy effects that also lead to racial disparities. Thus, simply excluding race is insufficient at guaranteeing that risk predictions are truly race neutral.

IV.   FORMALISTIC AND STATISTICAL SOLUTIONS TO ENSURING RACE NEUTRALITY

In this Part, we discuss three potential solutions that can eliminate the direct and proxy effects of race and nonrace correlates in predictive algorithms. The first formalistic solution follows the mainstream legal consensus by excluding both race and all nonrace correlates from the predictive algorithm, an approach that we argue is unlikely to work in practice because nearly all algorithmic inputs are correlated with race. Even if there remain some inputs that are uncorrelated with race, the set of permissible inputs under this formalistic solution is likely so small that the accuracy of the algorithm will be substantially degraded. We then propose two statistical solutions that build on the fact that algorithmic predictions are formed through an estimation step and prediction step. Our first recommended solution purges all algorithmic inputs of the proxy effects of race in the estimation step of the predictive algorithm, and then uses these "colorblind" inputs to predict outcomes in the prediction step. Our second recommended solution instead uses only white individuals in the estimation step of the predictive algorithm, and then uses these "colorblind" estimates to predict outcomes for both white and Black individuals in the prediction

step. Our two recommended solutions allow us to address direct and proxy effects of race without jettisoning all race-correlated inputs.

A. *Formalistic Solution: The Excluding-Inputs Algorithm*

We have shown that because the mainstream practice advocates for an outright exclusion of race, algorithms will automatically generate proxy effects if any correlated input is used. Taking these positions as given, we now identify the type of algorithm supported by legal scholars who seek to eliminate both direct and proxy effects of race from predictive algorithms.

We call this solution the "excluding-inputs" algorithm. This algorithm explicitly excludes using race directly, $X_i^{Protected}$, *and* excludes using any correlated inputs, $X_i^{Correlated}$. By excluding race and all race correlates, the excluding-inputs model is mechanically fair in that it does not use race in forming predictions, either directly or through proxy effects. The only remaining inputs that are permissible under the excluding-inputs model are uncorrelated inputs, or $X_i^{Uncorrelated}$. We believe that this solution most intuitively follows from legal definitions of fairness given our survey of risk-assessment tools as reviewed in Part II. These tools generally exclude race explicitly and often exclude factors that are correlated with race out of a view that their inclusion would be illegal, unethical, and/or unjust. As a result, some hold the view that the fewer the inputs, the better, as the legal position is based on excluding as many problematic inputs as possible. For example, practitioners have claimed that "[a]n effective risk assessment must be gender and race neutral . . . . The more risk factors you have, the less likely you'll be able to eliminate gender and racial bias."[194] Thus, the excluding-inputs algorithm is likely to be very parsimonious.

Estimation of the Algorithm: To illustrate how this algorithm would form predictions, we return to the two-step process described above in Section III.C. We first estimate the following statistical relationship, using only uncorrelated inputs:

$$Y_i = \delta_0 + \delta_1 \cdot X_i^{Uncorrelated} + \epsilon_i$$

(Equation 8)

We can then form the following predictions:

$$\hat{Y}_i^{ExcludingInputs} = \hat{\delta}_0 + \hat{\delta}_1 \cdot X_i^{Uncorrelated}$$

(Equation 9)

---

194.   *See* Issie Lapowsky, *One State's Bail Reform Exposes the Promise and Pitfalls of Tech-Driven Justice*, WIRED (Sept. 5, 2017, 7:00 AM), https://www.wired.com/story/bail-reform-tech-justice/ [https://perma.cc/SZ3B-E6ZB].

where $\hat{\delta}_1$ is the estimated relationship between the uncorrelated characteristics and the outcome of interest. The estimated coefficient $\hat{\delta}_1$ from this model is not affected by any direct or proxy effects of race, as we have assumed that $X_i^{Uncorrelated}$ are uncorrelated with the other input factors. As a result, the predictions from the excluding-inputs algorithm will not generate unwarranted racial disparities in predicted outcomes.

However, an important concern with this algorithm is that it comes with a substantial cost in terms of predictive accuracy. This model will generally be much less accurate than models that use $X_i^{Protected}$ and/or $X_i^{Correlated}$ because it purposely excludes the largest set of factors that are predictive of the outcome of interest. The loss in predictive accuracy can be large, with the exact loss depending on the statistical usefulness of the inputs that are excluded.

In the most extreme case, the excluding-inputs algorithm is infeasible if *all* characteristics are either protected or correlated, as is likely to be the case in settings such as the criminal justice system.[195] This is because avoiding proxy effects through the excluding-inputs algorithm requires that predictive algorithms only use inputs that are completely uncorrelated with race, a nearly impossible task given the influence of race in nearly every aspect of American life today. In that scenario, there would be no way of using an algorithm to form predictions.

How do commonly used algorithms fare compared to this traditional solution? Perhaps because of the likely impossibility of finding uncorrelated inputs, most if not all predictive algorithms today likely fail to meet the standard of race neutrality under the "excluding-inputs" solution. Recall from our survey of commonly used risk-assessment instruments in Part II that some algorithms include socioeconomic factors such as education or employment, and all reviewed algorithms included measures of past criminal history. The inclusion of these factors, which are likely to be highly correlated with race, will result in algorithms that some may argue are unfair. As a result, there is no guarantee that the estimates from these commonly used algorithms rely only on $X_i^{Uncorrelated}$ and are truly race neutral.

Indeed, the approach taken by commonly used algorithms is simultaneously overinclusive and underinclusive, in the sense that it is not satisfying to someone who prioritizes fairness and not satisfying to someone who priorities accuracy. Excluding inputs correlated with race because of equity concerns throws out all the predictive power from race-correlated inputs, even the predictive power that is *independent* of race. Including inputs correlated with race because of accuracy concerns results in unwarranted racial disparities because of proxy effects.

The challenge of finding uncorrelated inputs puts algorithmic creators in an understandably difficult situation of trying to minimize unwarranted

---

195. *See infra* Section IV.C.

racial disparities without jettisoning all possible inputs. With this practical concern in mind, we now turn to our two recommended solutions, which allow algorithmic creators to retain all predictive inputs while simultaneously eliminating the direct and proxy effects of race.

B. *Our First Solution: The Colorblinding-Inputs Algorithm*

We call our first statistical solution the "colorblinding-inputs" algorithm. Like the excluding-inputs algorithm, this solution also eliminates both direct and proxy effects of race when forming predictions, thereby eliminating unwarranted racial disparities. Unlike the excluding-inputs algorithm, however, the colorblinding-inputs algorithm does not exclude race and race correlates in the estimation step. In fact, it uses *all* inputs to estimate predictive relationships, in contrast to the current approach of using ad hoc human judgment to decide which race-correlated inputs are permissible, which we believe leaves much to be desired from either a fairness or accuracy perspective. Because the colorblinding-inputs algorithm allows us to use all possible correlated characteristics purged of their proxy effects, this statistical solution can achieve fairness without as large a sacrifice on predictive accuracy compared to the formalistic solution of wholly excluding correlated inputs. Thus, the colorblinding-inputs algorithm can preserve a large amount of predictive accuracy because race-correlated inputs often contain information orthogonal of race that is predictive of the outcome of interest. At the extreme, our solution allows one to use an algorithm even if every possible input is correlated with race, a scenario in which the formalistic solution would be impossible to implement.

As we will demonstrate below, the key feature of the colorblinding-inputs algorithm is that it explicitly uses race in the estimation step in order to colorblind all nonrace inputs, and then ignores individual race information in the prediction step. In other words, all possible inputs are used to estimate the algorithm, but only nonrace information from each given individual is used when the algorithm is applied to their specific case.

Estimation of the Algorithm: To construct our colorblinding-inputs model, we follow the approach developed by Devin Pope and Justin Sydnor, which utilizes only the predictive power from input variables that is *orthogonal* to (or uncorrelated with) protected characteristics.[196] For example, we want to utilize only the variation from each input that is independent of its association with race, allowing us to purge predictions of all proxy effects.

Formally, this model is estimated again using a two-step procedure. In the first estimation step, we estimate the benchmark statistical case from Equation 1 that includes the full set of observable input characteristics:

---

196.   Pope & Sydnor, *supra* note 23.

$$Y_i = \beta_0 + \beta_1 \cdot X_i^{Uncorrelated} + \beta_2 \cdot X_i^{Correlated} + \beta_3 \cdot X_i^{Protected} + \epsilon_i$$

<div align="center">(Equation 10)</div>

where, as discussed previously, the estimates from this model yield the coefficients $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. Estimating this benchmark model allows us to obtain predictive weights on correlated characteristics ($\hat{\beta}_2$) that are not contaminated by proxy effects, exactly because we explicitly include $X_i^{Protected}$. Thus, this first estimation step ensures that we eliminate all proxy effects from including $X_i^{Correlated}$. Intuitively, we ensure that the estimated relationship between our outcome of interest and $X_i^{Correlated}$ is uncontaminated by only keeping the predictive power from $X_i^{Correlated}$ that is orthogonal to (or uncorrelated with) $X_i^{Protected}$. As a result, we are able to "colorblind" the correlated inputs, $X_i^{Correlated}$.

In the second prediction step, we use these "colorblind" inputs to form predictions. We also ensure that no direct effects of race are used to make predictions. To do so, we use the predictive power contained in $\hat{\beta}_1$ and $\hat{\beta}_2$ (purged of proxy effects), but *not* $\hat{\beta}_3$ (the direct effect of protected characteristics), to form risk predictions. To do this, we form predictions that use the average value of $X_i^{Protected}$ across all individuals (rather than individual-level information), $\overline{X}^{Protected}$, but the actual input values of $X_i^{Uncorrelated}$ and $X_i^{Correlated}$ for each specific individual. We therefore form the following prediction:

$$\hat{Y}_i^{ColorblindingInputs} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i^{Uncorrelated}$$
$$+ \hat{\beta}_2 \cdot X_i^{Correlated} + \hat{\beta}_3 \cdot \overline{X}^{Protected}$$

<div align="center">(Equation 11)</div>

By using $\overline{X}^{Protected}$ instead of $X_i^{Protected}$, we ensure that two individuals who differ only in terms of a protected characteristic will not receive different predictions under the model. And we ensure that protected characteristics do not contaminate the predictions through $X_i^{Correlated}$. Our colorblinding-inputs model therefore eliminates racial disparities driven by both direct or proxy effects, achieving race neutrality.

To provide a concrete example, we return to our hypothetical example from Table 2. In that hypothetical, Y_i is an indicator variable for FTA, $X_i^{Protected}$ is an indicator equal to 1 if an individual is Black, and $X_i^{Correlated}$ is an indicator equal to 1 if an individual has a prior criminal history. In the first estimation step, we estimate a model of FTA controlling for both race and prior criminal history, yielding the coefficients reported in Table 3. Specifically, having a prior criminal history increases the predicted risk of FTA by 54.1 percentage points, and being Black increases the predicted risk

of FTA by 33.0 percentage points. By including race, we ensure that the weight on prior criminal history is not contaminated by proxy effects. In the second prediction step, rather than use the real individual-level values for race, which would lead to higher predicted risk for Black individuals compared to otherwise similar white individuals, we input the same race value, $\overline{R}$, for all individuals. Here, as race is an indicator variable, $\overline{R}$ is simply the average rate of Black individuals in the hypothetical population, which is 50% by construction.[197] Thus, both white and Black individuals with no priors receive the same risk prediction, and both white and Black individuals with priors receive a predicted risk that is 54.1 percentage points higher than individuals with no prior criminal history. These risk predictions statistically ensure that Black and white individuals who are otherwise identical will receive the same predicted risk, eliminating both direct and proxy effects of race.

C.   *Our Second Solution: The Minorities-as-Whites Algorithm*

We call our second solution the "minorities-as-whites" algorithm. This solution also eliminates both direct and proxy effects of race when forming predictions, thereby eliminating unwarranted racial disparities. Unlike the excluding-inputs algorithm, this approach does not exclude any race-correlated inputs in the estimation step, allowing us to achieve fairness without as large a loss in predictive accuracy.

In much the same way as the colorblinding-inputs algorithm, the minorities-as-whites algorithm uses only the predictive power from each input within race. In a scenario in which the relationship between each input and the outcome of interest is identical for white and nonwhite individuals, the minorities-as-whites algorithm and colorblinding-inputs algorithm will yield *identical* predictive weights on $X_i^{Uncorrelated}$, $X_i^{Correlated}$, and $X_i^{Protected}$. But the two algorithms will differ when the relationship between inputs and outcome of interest are different for white and nonwhite individuals (e.g., the effect of age on risk is different for white versus nonwhite individuals). One can think of the predictive weights that emerge from the colorblinding-inputs algorithm as a weighted average of the "minority" weights and "white" weights.

In this situation where minority weights and white weights differ, the difference between the two algorithms is that the minorities-as-whites algorithm uses only the predictive power from inputs among *white* individuals, not both white and minority individuals, thereby ensuring that the algorithm treats minority individuals exactly the same way it treats white individuals. That is, we use only white individuals in the first estimation step of the predictive algorithm, then rely on the resulting relationships in the second prediction step to predict outcomes for both white *and* nonwhite

---

197.   *See* Table 2.

individuals. In spirit, our minorities-as-whites algorithm is based on what economists call a "Blinder-Oaxaca decomposition,"[198] a method typically used to explore the role of group differences versus discrimination in driving racial differences in outcomes. In other settings, researchers have used similar methods to generate "unbiased" or "proxy-free" AFQT scores by predicting AFQT scores for white and nonwhite individuals through an estimation equation using coefficients/weights for white individuals.[199]

Which baseline population to use in the first estimation step is a choice. Just as one can design a minorities-as-whites algorithm, one can also easily design a whites-as-minorities algorithm. Or one can choose a particular weighted average, as in the colorblinding-inputs algorithm. But why might we want to limit the estimation step to white individuals? By focusing only on white individuals as the baseline population in the first estimation step, there may be less concern that inputs like criminal history are an outgrowth of discrimination. For example, one might believe that measured criminal history is not a true reflection of past criminality among nonwhite individuals because of certain policing practices. But if one believes that bias in policing is not an issue among white defendants and that criminal history is an accurate reflection of past criminality for these individuals, estimating the relationship between criminal history and future risk using white individuals alone can eliminate any proxy effects. In addition, focusing only on white individuals in the first estimation step can also address concerns about measurement error in the outcome of interest—say rearrest for a new crime. If risk is measured as rearrest, an outcome that can also be plagued by bias, focusing on a group where measured rearrest is more objective may eliminate another form of discrimination being baked into the algorithm.

Estimation of the Algorithm: To construct our minorities-as-whites model, we estimate the predictive relationship between each input and outcome of interest for the population of white individuals, and then apply these predictions equally to both white and nonwhite individuals.

Formally, our model is estimated in two steps. In the first estimation step, we estimate the benchmark statistical model from Equation 1 that includes the full set of observable input characteristics, but for *white individuals only*:

$$Y_i^W = \beta_0^W + \beta_1^W \cdot X_i^{Uncorrelated} + \beta_2^W \cdot X_i^{Correlated}$$
$$+ \beta_3^W \cdot X_i^{Protected} + \epsilon_i^W$$

(Equation 12)

198.   Alan S. Blinder, *Wage Discrimination: Reduced Form and Stuctural Estimates*, 8 J. HUM. RES. 436 (1973); Ronald Oaxaca, *Male-Female Wage Differentials in Urban Labor Markets*, 14 INT'L ECON. REV. 693 (1973).

199.   William M. Rodgers III & William E. Spriggs, *What Does the AFQT Really Measure: Race, Wages, Schooling and the AFQT Score*, REV. BLACK POL. ECON., Spring 1996, at 13.

where the estimates from this model yield the coefficients for white individuals $\hat{\beta}^W_1$, $\hat{\beta}^W_2$, and $\hat{\beta}^W_3$.

In the second prediction step, we ensure that no direct effects of race are used to make predictions, that is, that a white and nonwhite individual who are otherwise identical receive the same risk predictions. To do so, we form the following predictions for white and nonwhite defendants:

$$\hat{Y}_i^{Minorities as Whites} = \hat{\beta}_0^W + \hat{\beta}_1^W \cdot X_i^{Uncorrelated} + \hat{\beta}_2^W \cdot X_i^{Correlated} + \hat{\beta}_3^W \cdot X_i^{Protected}$$

(Equation 13)

by applying the *same* coefficients $\hat{\beta}^W_1$, $\hat{\beta}^W_2$, and $\hat{\beta}^W_3$ for all races.

To provide a concrete example, return again to our hypothetical example from Table 2. Under this hypothetical, recall that Y_i is an indicator variable for FTA, $X_i^{Protected}$ is an indicator equal to 1 if an individual is Black, and $X_i^{Correlated}$ is an indicator equal to 1 if an individual has a prior criminal history. In the first estimation step, we estimate predictions of FTA controlling for prior criminal history among only the population of white individuals. This first step yields the statistical relationship that having a prior criminal history increases the risk of FTA by 66.6 percentage points. In the second prediction step, we apply this relationship equally for both white and Black individuals, such that white and Black individuals with a prior criminal history receive risk predictions that are 66.6 percentage points higher than individuals with no prior criminal history. As a result, we ensure that Black and white individuals who are otherwise identical will receive the same predicted risk.

D. *Legality of Our Two Statistical Solutions*

Before we move on to an empirical assessment of how much our two proposed statistical solutions improve upon commonly used algorithms, we discuss the legality of our proposed solutions, the colorblinding-inputs and minorities-as-whites algorithms. The most salient distinction (from a legal perspective) of our two statistical solutions relative to the formalistic excluding-inputs solution (which prohibits the use of race and all race correlates) is that both our statistical proposals explicitly require the consideration of race in the estimation step. Specifically, the colorblinding-inputs algorithm uses race in the first estimation step in order to eliminate proxy effects but does not use individual race information in the second prediction step. Similarly, the minorities-as-whites algorithm considers race in the first estimation step in order to remove nonwhite individuals from the sample used to estimate the model, before proceeding to the second prediction step. Critically, the use or consideration of race in the first estimation step serves the distinct purpose of achieving a race-neutral prediction. To the best of our knowledge, neither of our two proposed

algorithms are used in practice today. This is likely because of the formalistic prohibition on the use of protected characteristics under the Equal Protection Clause, a position that fails to take into account the statistical reality of how algorithms work.

Nevertheless, a growing number of scholars, like us, have emphasized the need for the law to permit the use of protected characteristics in the first estimation step, as specified in our two proposals. For example, Indrė Žliobaitė and Bart Custers have argued that "in order to make sure that decision models are non-discriminatory, for instance, with respect to race, the sensitive racial information needs to be used in the model building process. Of course, after the model is ready, race should not be required as an input variable for decision making"[200]—just as we propose in our statistical solutions. Thus, the authors conclude that "collecting sensitive personal data is needed in order to guarantee fairness of algorithms, and law making needs to find sensible ways to allow using such data in the modeling process."[201]

Similarly, in the Title VII context, Harned and Wallach advocate for the use of a "middle ground . . . in which the system is blinded to sensitive attributes only during deployment and not during training,"[202] precisely the procedure underlying our two statistical solutions. In contrast to a formalistic approach that forbids the sensitive attribute (which they claim is legally compliant but does not mitigate discrimination given proxy effects) and an approach that fully uses the attribute (which they argue would likely be perceived as direct evidence of disparate treatment and thus illegal), Harned and Wallach argue that the "middle ground" of using the attribute in the first estimation step but *not* the second prediction step should be legally permissible.[203] Indeed, the authors analogize these algorithms to legally accepted auditing procedures, found in contexts like employment and university admissions.[204] As a result, they claim that this type of algorithm "avoids disparate treatment claims because it does not use racial classifications for decision making."[205] Ignacio Cofone also analogously argues that an algorithm that mitigates bias in the first estimation step and not the second prediction step "addresses discrimination by dealing with the input data not the decision process . . . . Therefore, it would not fall under the constitutional challenges based on disparate treatment."[206]

---

200.   Indrė Žliobaitė & Bart Custers, *Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models*, 24 ARTIFICIAL INTEL. & L. 183, 183 (2016).

201.   *Id.* at 199.

202.   Harned & Wallach, *supra* note 72 (manuscript at 18).

203.   *Id.* (manuscript at 20–22).

204.   *Id.* (manuscript at 22).

205.   *Id.*

206.   Cofone, *supra* note 91, at 1421–24, 1429–31.

For the same reasons, we also firmly believe that our two proposed solutions should be permissible under the equal protection doctrine and antidiscrimination law. One might at first object to our proposals because they appear "counterintuitive" in considering a protected characteristic in the design of the algorithm. Thus, a naive observer may argue that our proposals are illegal because they run up against the anticlassification principle, which many argue drives our understanding of the equal protection doctrine.[207] But this formalistic objection fails to take into account the previously demonstrated statistical reality of direct and proxy effects and unnecessarily distorts the anticlassification principle. At its heart, the anticlassification principle rests on a view that "the Constitution protects individuals, not groups, and so bars all racial classifications, except as a remedy for specific wrongdoing."[208] But what the principle prohibits is differential treatment on the basis of a protected characteristic.[209] And our two proposed algorithms are built to prevent precisely this type of differential treatment: to ensure that decisions are not made with respect to a protected trait, either directly or indirectly, which statistically requires the use of race in the estimation step but *not* the prediction step.[210] Thus, we argue that if a formalistic

---

207.   For example, some commentators have observed of the colorblinding-inputs algorithm that "[c]ounterintuitively, the first step in this process is for the statistical model under consideration to be re-estimated in a way that explicitly includes data on legally prohibited characteristics." Prince & Schwarcz, *supra* note 18, at 1314; *see also* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV (Aug. 14, 2018), https://arxiv.org/pdf/1808.00023.pdf [https://perma.cc/9UJ3-CTQS] ("It can feel natural to exclude protected characteristics in a drive for equity . . . . In contrast to the principle of anti-classification, it is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics.").

208.   Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278, 1281 (2011); *see also* Missouri v. Jenkins, 515 U.S. 70, 120–21 (1995) (Thomas, J., concurring) ("At the heart of this interpretation of the Equal Protection Clause lies the principle that the government must treat citizens as individuals, and not as members of racial, ethnic, or religious groups. It is for this reason that we must subject all racial classifications to the strictest of scrutiny . . . ."); Charles R. Lawrence III, Essay, *Two Views of the River: A Critique of the Liberal Defense of Affirmative Action*, 101 COLUM. L. REV. 928, 950 (2001) (associating the anticlassification principle with "[l]iberal legality[,] [which] sees the equality principle as primarily concerned with protecting individuality, and views racial discrimination as unjust because when we judge a person based on her race we disregard her unique human individuality").

209.   *See, e.g.*, Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 205, 227 (1995) (applying strict scrutiny to a federal program designed to provide highway contracts to disadvantaged business enterprises, where it was presumed that socially and economically disadvantaged individuals include "Black Americans, Hispanic Americans, Native Americans, Asian Pacific Americans, and other minorities" (quoting 15 U.S.C. § 637(d)(2)–(3) (1988))); *see also* Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (applying strict scrutiny when "the government distributes burdens or benefits on the basis of individual racial classifications").

210.   Anya Prince and Daniel Schwarcz note, with respect to these algorithms, that "[i]n a very real sense, the process explicitly discriminates with respect to membership in a legally pro-

interpretation of the Equal Protection Clause prohibits our proposals, it does more harm than good by undermining the very objective of the anticlassification principle. As Harned and Wallach put it, in the context of Title VII,

> This tension arises from our attempt to stretch human laws to apply to machines even though human decision-making processes are quite different from automated decision-making processes. In other words, when stretched to apply to machines, laws designed to regulate human behavior may even be detrimental to the very people that they were designed to protect.[211]

In recent work, Benjamin Eidelson also argues that treating people as individuals, a core commitment underlying the anticlassification approach to race and equal protection, may require shedding the requirement of colorblindness. In laying out a new account of what it means to treat people as individuals, Eidelson notes:

> Indeed, in a society characterized by racial bias, attending to race will often be *necessary* to treating a person respectfully as an individual—because race will mediate evidential connections between her record of choices or achievements and what the Court calls "her own essential qualities." Colorblindness, which is so often justified as a way of respecting people as individuals, thus stands in the way of doing exactly that.[212]

Further, we believe that the use of race in the first estimation step does not even constitute the type of "express racial classification" that triggers strict scrutiny. Legal scholars have noted courts have not always applied strict scrutiny to governmental actors' use of race, such as the use of racial descriptions of criminal suspects by law enforcement,[213] the Census' collection of race and ethnicity information,[214] and race-conscious adoption placements by social service agencies.[215] As Jack Balkin and Reva Siegel state, for example, "[w]hile state action doctrine may limit the reach of the anticlassification principle, it is commonly assumed that all use of race by state actors is subject to strict scrutiny. This is not in fact the case."[216]

---

tected group in order to prevent the effects of such discrimination from being felt by these individuals." Prince & Schwarcz, *supra* note 18, at 1315. We, however, contest the view that using a protected trait in the first estimation step is a form of explicit discrimination.

211.   Harned & Wallach, *supra* note 72 (manuscript at 21).

212.   Eidelson, *supra* note 57, at 1607.

213.   Brown v. City of Oneonta, 221 F.3d 329 (2d Cir. 2000) (upholding use of race in suspect descriptions).

214.   *See* Morales v. Daley, 116 F. Supp. 2d 801, 814–15 (S.D. Tex. 2000) (holding that census questions concerning race and ethnicity do not violate the Fifth Amendment).

215.   *See* R. Richard Banks, *The Color of Desire: Fulfilling Adoptive Parents' Racial Preferences Through Discriminatory State Action*, 107 YALE L.J. 875, 904–05 & n.135 (1998).

216.   Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 19 (2003).

Similarly, Richard Primus has stated that "many practices that do involve government actors' identifying people by race are not always subject to strict scrutiny. . . . [N]ot all instances in which the government explicitly considers the race of individuals are 'express racial classifications' for purposes of equal protection doctrine. Some are, and some are not."[217] Because our two proposed algorithms do not use racial classifications for the purposes of *decisionmaking*—rather, our algorithms use race as a factor precisely to ensure that race is never used as a criterion based on which individuals are treated differently—our approaches should not be considered "express racial classifications." As with the use of collection of race information for the Census,

> "Statistical information as such is a rather neutral entity which only becomes meaningful when it is interpreted." . . . [There is a] distinction between collecting demographic data so that the government may have the information it believes at a given time it needs in order to govern, and governmental use of suspect classifications without a compelling interest.[218]

In addition, our two proposed algorithms should not trigger strict scrutiny because race is never the sole criterion used in any step of the design of the algorithm, let alone never a categorical determinant of government decisiomaking. As stated by R. Richard Banks, "[a]lthough the members of the Court disagree on the threshold use of race that triggers strict scrutiny, a majority of the Court seems to embrace the view that a minimal reliance on race does not trigger the racial classification rule."[219] To support this proposition, Banks cites to litigation under the Voting Rights Act, where the Supreme Court has stated that "[s]trict scrutiny does not apply merely because redistricting is performed with consciousness of race. . . . For strict scrutiny to apply, the plaintiffs must prove that other, legitimate districting principles were 'subordinated' to race."[220] Thus, while our two proposals are "race conscious" in some sense, they do not allow any decisionmaking to be based fundamentally on race.

Furthermore, we view our proposed solutions as consistent with an alternative conception of equal protection, the antisubordination principle. As summarized by David Strauss,

> This principle holds that the evil of discrimination does not lie in the use of a racial (or other similar) criterion for distinguishing among people. Rather the evil of discrimination is the particular kind of harm that it inflicts on the disadvantaged group—in varying formulations, it subordinates

---

217.   Primus, *supra* note 37, at 505–06.

218.   *Morales*, 116 F. Supp. 2d at 814 (quoting United States v. New Hampshire, 539 F.2d 277, 280 (1st Cir. 1976)).

219.   Banks, *supra* note 215, at 904–05.

220.   *See* Bush v. Vera, 517 U.S. 952, 958–59 (1996) (plurality opinion) (quoting Miller v. Johnson, 515 U.S. 900, 916 (1995)).

them . . . . According to the anti-subordination principle, where that particular kind of harm is absent, there is no unlawful discrimination, even if a racial classification is used.[221]

Many argue that these antisubordination principles have been reflected in case law, such as in *Strauder v. West Virginia*, where the Supreme Court condemned "discriminations which are steps towards reducing [Black individuals] to the condition of a subject race."[222] Similarly, in *Loving v. Virginia*, the Court struck down antimiscegenation laws, in part because they are "justifi[ed only]as measures designed to maintain White Supremacy."[223] Siegel has argued, with respect to *Brown v. Board of Education*, that "many justifications offered for *Brown* sounded like an antisubordination defense of the opinion might today."[224] We view both of our statistical proposals, which consider race in the estimation step race precisely to avoid inflicting harm on disadvantaged groups through proxy effects that can reflect past discrimination (such as discriminatory policing), as fully consistent with the antisubordination principle.

Ultimately, we believe that our approaches do not violate the core tenets that underlie the equal protection doctrine and should be legally permissible. Because our proposed solutions consider race only in the first estimation step and do so precisely to purge algorithmic predictions of any proxy effects—thus not permitting individual predictions to be based on racial-group membership—our approaches are very much in line with the goal of treating citizens as individuals and the goal of not inflicting harm on disadvantaged groups.[225]

---

221.   Strauss, *supra* note 20, at 1; *see also* Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N.Y.U. L. REV. 1003, 1007–08 (1986) ("Th[e] [anti-subordination] approach seeks to eliminate the power disparities between men and women, and between whites and non-whites . . . . From an anti-subordination perspective, both facially differentiating and facially neutral policies are invidious only if they perpetuate racial or sexual hierarchy."); Siegel, *supra* note 208, at 1288–89 ("[T]he antisubordination principle is concerned with protecting members of historically disadvantaged groups from the harms of unjust social stratification. . . . Because the antisubordination principle focuses on practices that disproportionally harm members of marginalized groups, it can tell the difference between benign and invidious discrimination.").

222.   100 U.S. 303, 308 (1880).

223.   388 U.S. 1, 11 (1967).

224.   Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over* Brown, 117 HARV. L. REV. 1470, 1474 (2004).

225.   An algorithm that uses race in the second prediction step is a potentially different story, an issue we address in Section VI.B. But even in that context, where the algorithm may constitute an "express racial classification," some scholars have argued that the algorithm could withstand strict scrutiny under a form of algorithmic affirmative action. *See, e.g.*, Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 849–51 (2020).

E.   *Racial Disparities Under Our Two Statistical Solutions*

Above, we have presented two statistical solutions, the colorblinding-inputs and minorities-as-whites algorithms. We believe that these two statistical solutions, in contrast to the formalistic excluding-inputs algorithm, not only are implementable in practice but also better advance the widespread goal of policymakers and legal advocates who seek to eliminate both direct and proxy effects of race in predictive algorithms.

It is important to note, however, that neither of our two statistical solutions would result in complete *racial balance* in terms of resulting algorithmic predictions or outcomes. Specifically, predictions for the minority population and predictions for the white population are not guaranteed to be identical. Why is this? Recall that we have defined predictions as race neutral if algorithmic predictions have been purged of both direct and proxy effects of race, a view that we believe best captures the leading legal consensus. Any remaining racial disparities after elimination of direct and proxy effects are thus, by definition, not unwarranted. Under both our statistical solutions, algorithmic predictions would still result in some racial gaps so long as characteristics of individuals vary by protected class. For example, even if we eliminate the direct and proxy effects of race, it may still be the case that having a prior criminal history leads to higher predicted risk,[226] as is almost always the case in commonly used algorithms. If Black individuals, on average, are more likely to have a prior criminal history compared to white individuals, risk predictions may still be higher on average for Black individuals relative to white individuals. In our view, these remaining racial gaps are not unwarranted, and we are not aware of any definition of algorithmic fairness that requires predictions by race to be equal. In fact, as legal scholars have pointed out, "racial balance . . . is not legally mandated, and efforts to pursue that goal might themselves be struck down on constitutional grounds."[227]

Some may argue that the overrepresentation of prior criminal records for Black individuals relative to white individuals is not due to valid differences in criminal behavior, but rather discrimination—a critique that is sometimes referred to as "measurement error" in predictive inputs that is correlated with race. We are sympathetic to this critique, but unfortunately, we are not aware of any systematic approach to dealing with these measurement issues, when dealing with either algorithms or human decisionmakers. For example, we are not aware of any government that attempts to correct for mismeasurement of, say, prior conviction records by adjusting what it means to have a prior conviction for Black offenders versus white offenders. The only real solution to this specific issue is to understand the possible sources of measurement error and find inputs that do not suffer

---

226.   *See supra* Table 4.

227.   Sunstein, *supra* note 93, at 509.

from measurement error, a worthwhile goal when dealing with both algorithms and human decisionmaking and an area that we believe is ripe for future research.

Our two statistical solutions also do not directly address measurement error in the *outcome of interest* or "label," focusing instead on direct and proxy effects from *inputs*. But a challenge for almost all algorithms is measurement error in the outcome itself, which may be systematically correlated with race, and thus exhibit another source of potential unfairness. For example, in the criminal justice context, we rarely observe actual criminal activity, relying instead on arrests or convictions. These observable outcomes may be imperfect measures of underlying activity that are differentially mismeasured by race due to structural inequalities, as would be the case if criminal justice actors were racially biased. The solution to this particular form of bias is challenging but requires carefully choosing observable outcomes that are good stand-ins for the underlying outcome of interest and that are less likely to suffer from mismeasurement.

V.   EMPIRICAL TESTS OF OUR PROPOSED STATISTICAL SOLUTIONS

In this Part, we present our main empirical results using data from the pretrial system in New York City. We begin with a brief overview of the New York City pretrial system and our data. We then demonstrate that commonly used algorithms are likely to generate unwarranted racial gaps by including variables that, in practice, are all highly correlated with race. We then show that, as a result, these algorithms generate economically meaningful proxy effects and unwarranted racial disparities. We conclude by showing that our two proposed statistical solutions substantially reduce the number of Black defendants detained compared to more commonly used algorithms.

A.   *The New York City Pretrial System*

Background on Arraignment and Bail: In the United States, the bail system is meant to allow all but the most dangerous criminal suspects to be released from custody while ensuring their appearance at required court proceedings and, in some jurisdictions, also ensuring the public's safety. The federal right to nonexcessive bail is guaranteed by the Eighth Amendment to the U.S. Constitution,[228] with almost all state constitutions granting similar rights to defendants.[229] In New York, the state constitution states that "[e]xcessive bail shall not be required nor excessive fines imposed."[230] New

---

228.   U.S. CONST. amend. VIII.

229.   Ariana Lindermayer, Note, *What the Right Hand Gives: Prohibitive Interpretations of the State Constitutional Right to Bail*, 78 FORDHAM L. REV. 267, 283–84 (2009).

230.   N.Y. CONST. art. I, § 5.

York's bail statute also grants a right to some form of bail for most defendants. According to § 510.10 of New York Criminal Procedure Law (CPL),

> When a principal, whose future court attendance at a criminal action or proceeding is or may be required, initially comes under the control of a court, such court must, by a securing order, either release him on his own recognizance, fix bail or commit him to the custody of the sheriff.[231]

Excepting cases wherein the defendant is charged with a Class A felony or has two previous felony convictions, the court may order recognizance or bail for a defendant.[232] If the defendant only has charges that are less than felony grade, the court must order recognizance or bail.[233] New York law also states that the sole purpose of bail is to ensure that the defendant returns to court such that the *only* consideration at arraignment is the defendant's risk of failure to appear, and not dangerousness to the community.[234]

In New York City, the pretrial process generally starts when a police officer brings the arrestee to the precinct for processing, where the defendant is photographed and fingerprinted. While the defendant's criminal history is being processed, the arresting officer meets with an assistant district attorney to draft a complaint to begin the prosecution process. Meanwhile, the defendant is interviewed for a bail recommendation by the Criminal Justice Agency (CJA), which has created a pretrial risk-assessment instrument that predicts the risk of failing to appear for future court dates, known as the "CJA score."[235] The DCJS and CJA reports, along with the complaint, are then delivered to court arraignment clerks to file the defendant's information, a docket number is assigned, and the case is initialized in the court's computerized records. The arraignment process cannot proceed until all of these documents are submitted into the system. The defendant's counsel is finally given an opportunity to interview the defendant prior to arraignment.[236]

---

231.   N.Y. CRIM. PROC. LAW § 510.10 (McKinney 2020).

232.   *Id.* § 530.20(2)(a).

233.   *Id.* § 530.20.

234.   *See id.* § 510.30 ("With respect to any principal, the court in all cases, unless otherwise provided by law, must impose the least restrictive kind and degree of control or restriction that is necessary to secure the principal's return to court when required."); *see also* Sardino v. State Comm'n on Jud. Conduct, 448 N.E.2d 83, 85 (N.Y. 1983) (holding that in New York, the "only matter of legitimate concern" when setting bail is "whether any bail or the amount fixed was necessary to insure the defendant's future appearances in court").

235.   Sainath R. Iyer, *Will NYC's New Pre-Trial Risk Assessment Be Race Neutral?*, N.Y.U. J. LEGIS. & PUB. POL'Y QUORUM (Nov. 5, 2017), https://nyujlpp.org/quorum/iyer-nyc-pretrial-risk-assessment-race/ [https://perma.cc/BNB3-4GV5].

236.   OFF. OF THE CHIEF CLERK OF THE N.Y.C. CRIM. CT., CRIMINAL COURT OF THE CITY OF NEW YORK: 2013 ANNUAL REPORT 18–20 (2014) [hereinafter NYC 2013 ANNUAL REPORT].

During this period between arrest and arraignment, most arrestees in New York City are transferred to holding cells in each borough's criminal court, with arraignments usually taking place within twenty-four hours of arrest.[237] However, not all arrested individuals will be held in holding cells prior to arraignment. For certain individuals with no outstanding warrants at the time of arrest, the arresting police officer may use his or her discretion to issue a Desk Appearance Ticket (DAT).[238] This DAT allows the arrested individual to be released but requires them to return to court for a later prescheduled arraignment. Twenty-eight percent of all misdemeanor arrests were issued a DAT in 2016.[239] Between DATs and non-DATs, in 2016, 249,776 criminal cases were arraigned in New York City, with these cases largely comprised of misdemeanor charges (82%).[240]

At arraignment, the first court appearance in the criminal justice process in New York City, an arraignment judge notifies the defendant of the charges he faces and the rights he has.[241] In contrast to some other jurisdictions, almost half of all case filings are disposed of at arraignment in New York City.[242] For many misdemeanor defendants, the case is dismissed at arraignment or adjourned in contemplation of dismissal (ACD).[243] In 2013, for example, about 80% of first-time nonviolent-misdemeanor youth had their cases resolved with an outright dismissal or ACD.[244]

For the cases that are not disposed of at arraignment, the assigned arraignment judge has several options when setting the pretrial-release conditions. First, defendants who show a minimal risk of flight may be released on their promise to return for all court proceedings, known broadly as release on recognizance (ROR). In practice, about 70% of defendants in New York City are released via ROR at arraignment such that no bail is set and no other court conditions are mandated.[245] Second, defendants may be required to post some sort of bail payment to secure release if they pose an appreciable risk of flight, which makes up most of the remaining 30% of

---

237. INDEP. COMM'N ON N.Y.C. CRIM. JUST. & INCARCERATION REFORM, A MORE JUST NEW YORK CITY 37 (2017) [hereinafter THE LIPPMAN REPORT], https://static1.squarespace .com/static/5b6de4731aef1de914f43628/t/5b96c6f81ae6cf5e9c5f186d/1536607993842/Lippman %2BCommission%2BReport%2BFINAL%2BSingles.pdf [https://perma.cc/JKM4-KTVY].

238. *Id.* This is permitted if the arrest charge is a violation, misdemeanor, or Class E felony. DATs are not permitted for other types of felonies (e.g., Class A–D felonies).

239. *Id.*

240. *Id.*

241. NYC 2013 ANNUAL REPORT, *supra* note 236, at 18–19.

242. *Id.* at 19.

243. THE LIPPMAN REPORT, *supra* note 237, at 37.

244. *See id.*

245. *Id.* at 42.

cases.[246] If the defendant is remanded or is unable to make the set bail, he or she is detained until the adjudication of their case.[247] For more serious crimes, the arraignment judge may require that the defendant be detained pending trial by denying bail altogether.[248] Bail denial is often mandatory in first- or second-degree murder cases, but can be imposed for other crimes when the bail judge finds that no set of conditions for release will guarantee appearance.[249] For example, prior to recent bail reform in New York City, a Class A felony, which includes murder, kidnapping, arson, and high-level drug possession and sale, almost always resulted in a denial of bail. These cases make up about 0.8% of all cases in New York City.[250] Finally, about 1.5% of cases are sent to a supervised release program as an alternative to pretrial detention.[251]

The assigned arraignment judge is granted considerable discretion in evaluating each defendant's circumstances when making decisions about release. With the exception of circumstances as detailed in NY CPL § 530 that prohibit discretion altogether, the assigned judge is meant to base his or her decision on the following mandated factors:

> The principal's activities and history; . . . the charges facing the principal; . . . The principal's criminal conviction record if any; The principal's record of previous adjudication as a juvenile delinquent . . . or a youthful offender, if any; The principal's previous record with respect to flight to avoid criminal prosecution; If monetary bail is authorized, according to the restrictions set forth in this title, the principal's individual financial circumstances . . . . Where the principal is charged with a crime or crimes against a member or members of the same family or

---

246.   In New York City, arraignment judges are required by law to set at least two forms of bail in these cases. *Id.* at 44 (citing People *ex rel.* McManus v. Horn, 967 N.E.2d 671 (N.Y. 2012)). The two most common bail options used are cash bail and insurance-company bail bond, despite there being nine forms of bail authorized by law. N.Y. CRIM. PROC. LAW § 520.10 (McKinney 2020); THE LIPPMAN REPORT, *supra* note 237, at 44. Cash bail requires the individual to pay the full bail amount up front in order to secure release while insurance-company bail bond requires an individual to deposit 10% of the bond amount as collateral with a bail-bond company. Infrequently used alternatives include credit-card bail, which allows an individual to use a credit card to pay bail of $2,500 or less; partially secured bonds, which require the individual to pay only a percentage of the total bail amount up to 10%; and unsecured bonds that do not require up-front payment. For both secured and unsecured bonds, the defendant is only liable for the rest of the bond if he or she fails to return to court. THE LIPPMAN REPORT, *supra* note 237, at 44. In New York, there is a 3% surcharge on all cash bail if the defendant is convicted, which the government keeps. N.Y. State Unified Ct. Sys., *Bail*, NYCOURTS.GOV (Oct. 31, 2016), https://www.nycourts.gov/courthelp/Criminal /bail.shtml [https://perma.cc/N89J-PPLE].

247.   N.Y. State Unified Ct. Sys., *supra* note 246.

248.   *Id.*

249.   *Id.*

250.   *See* THE LIPPMAN REPORT, *supra* note 237, at 41 n.23.

251.   *See id.*

household . . . any violation by the principal of an order of protection issued by any court . . . and the principal's history of use or possession of a firearm.[252]

Much of this information will be available in the defendant's rap sheet, DCJS, and CJA reports. While New York's bail statute also requires that judges take into account a defendant's "financial circumstances" when setting bail,[253] many have noted that there is little evidence that judges consider individual ability to pay in practice.[254] In considering these factors and arguments made by both prosecutors and defense counsel, it is estimated that the average arraignment in New York City lasts only six minutes given the caseload and number of arraignment judges available.[255]

Changes to the NYC Pretrial System: There have been several important changes to the pretrial system in New York City in recent years. Several charitable bail funds have, for example, started operating in New York since the enactment of a 2012 law that allows for the operation of bail funds that post bail in misdemeanor cases where bail is set at $2,000 or less.[256] These bail funds include the Bronx Freedom Fund, the Brooklyn Community Bail Fund, and the Liberty Fund.[257] In 2016, the Mayor's Office of Criminal Justice also created a supervised-release program with the goal of diverting 3,000 defendants each year who would otherwise be detained due to inability to pay bail to community supervision.[258] Under this supervised-release program, individuals receive supervision and conditions that are based on a risk-assessment screening created by the NY Criminal Justice Agency. Individuals charged with most misdemeanor and nonviolent felony charges are eligible for the program.[259] In 2018, the Mayor's Office also announced the creation of an online bail-payment system out of recognition of the extensive and difficult process for paying bail in person during business

---

252.   N.Y. CRIM. PROC. LAW § 510.30 (McKinney 2020).

253.   *Id.* § 510.30(1)(f).

254.   THE LIPPMAN REPORT, *supra* note 237, at 44 ("[I]f a person is on public assistance and you know they are receiving $300 a month, and you give them a $5,000 bail . . . that's a ransom—not a bail.").

255.   *See* Emily Leslie & Nolan G. Pope, *The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments*, 60 J.L. & ECON. 529, 533 (2017). The Criminal Court and its judges have the responsibility of conducting arraignments in New York City. *Id.*

256.   THE LIPPMAN REPORT, *supra* note 237, at 44–46.

257.   *Id.*

258.   THE LIPPMAN REPORT, *supra* note 237, at 45; *Supervised Release Program*, CTR. FOR CT. INNOVATION, https://www.courtinnovation.org/node/20042/more-info [https://perma.cc/BZ5Z-MP82].

259.   THE LIPPMAN REPORT, *supra* note 237, at 45.

hours.[260] Under the new online system, sureties no longer need to pay bail in person, individuals living out of state can pay bail on behalf of a defendant, and payment can now be shared across multiple people and multiple credit cards.[261]

Most recently, in April 2019, New York passed legislation on bail reform, effective January 2020.[262] Among these reforms, the most notable include the elimination of money bail and pretrial detention for nearly all misdemeanors and nonviolent felonies.[263] Electronic monitoring is also prohibited in most misdemeanor cases.[264] Among violent felonies, judges may set bail if they find that less restrictive means of release, such as nonmonetary conditions or electronic monitoring, are not sufficient to assure a person's return to court.[265] The law also mandates consideration of ability to pay when imposing bail.[266] In addition, the law gives a grace period for the issuance of bench warrants, prohibiting courts from issuing a warrant for forty-eight hours whenever a defendant fails to appear, unless the defendant is charged with a new crime or there is evidence of a "willful" failure to appear.[267] With respect to risk-assessment instruments, the law requires that such instruments be publicly available, free of bias due to "race, national origin, sex, or any other protected class," and validated for predictive accuracy.[268]

B. *Data Description*

This Section summarizes the most relevant information regarding our administrative-court data from New York City and provides summary statistics. We have data on all arraignments in New York City between November 1, 2008 and November 1, 2013, totaling 1,460,462 cases in all.[269] These data contain information on a defendant's gender, race, date of birth,

---

260.   *See Mayor de Blasio Announces Launch of Online Bail in New York City*, NYC (Apr. 27, 2018), https://www1.nyc.gov/office-of-the-mayor/news/226-18/mayor-de-blasio-launch-online-bail-new-york-city [https://perma.cc/F2HC-DWJ4].

261.   *Id.*

262.   MICHAEL REMPEL & KRYSTAL RODRIGUEZ, BAIL REFORM IN NEW YORK: LEGISLATIVE PROVISIONS AND IMPLICATIONS FOR NEW YORK CITY 1 (2019), https://www.courtinnovation.org/sites/default/files/media/document/2019/Bail_Reform_NY_full_0.pdf [https://perma.cc/62QN-8JMY].

263.   *Id.*

264.   *Id.* at 5.

265.   *Id.* at 2, 5.

266.   *Id.* at 4.

267.   *Id.* at 7.

268.   *Id.* at 6.

269.   These data exclude undocketed arrests as well as the substantial number of arrests for nonfingerprintable charges such as violations, infractions, and many misdemeanors (i.e., VTL 511s).

and county of arrest. The data also include extensive information that is available to the arraignment judge at the time of bail, including detailed information on the charge in the current offense, history of prior criminal convictions obtained from the rap sheet, and a history of past failures to appear. We also observe whether the defendant was released on recognizance at the time of arraignment or was assigned some form of bail, as well as whether the defendant eventually secured release on bail prior to case disposition. Finally, we can measure whether a defendant subsequently failed to appear for a required court appearance or was arrested for a new crime before case disposition because the data contain defendant identifiers that allow us to match the same individual across different cases. Given that nonappearance at court is the only legitimate concern taken into account at the time of setting bail in New York, our primary measure for pretrial misconduct is an indicator for failing to appear.

We make three restrictions to our final estimation sample. First, we limit the sample to non-Hispanic Black and non-Hispanic white male defendants charged with either a felony or misdemeanor (N = 718,305 cases from 345,940 unique defendants). Thus, our empirical results will focus on Black-white male disparities to illustrate our key concepts but our proposed statistical solutions can be easily extended to allow for other racial/ethnic comparisons or gender disparities, as we will discuss in Section VI.A. Second, we further limit the sample to cases that were not adjudicated or disposed of at arraignment and where we are not missing any information on background characteristics (N = 379,048 cases from 212,000 unique defendants). This restriction allows us to observe the sample of defendants who had a bail hearing. Finally, we further limit the sample to the approximately 85% of defendants who are released before trial and, as a result, who are relevant for our analysis (N = 264,379 cases from 180,887 unique defendants).[270] The final sample thus contains 264,379 cases from 180,887 unique defendants.

---

270.   Focusing on the sample of defendants actually released before trial raises a common issue known as "selection." Selection occurs because outcome data can be missing in a nonrandom way, where here we only observe pretrial-misconduct outcomes for individuals that judges decided to release before trial. The predictive relationship between inputs and outcome may thus be biased by selection. This is a common issue with risk-assessment instruments, which are validated on a sample of released defendants. *E.g.*, Shawn Bushway & Jeffrey Smith, *Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us*, 23 J. QUANTITATIVE CRIMINOLOGY 377, 378 (2007) ("[T]he data used in the formal risk assessments in the criminal justice and criminology literature are generated by processes of informal risk assessment and treatment assignment. Current sentences embody efforts by judges, prosecutors, parole boards and other actors in the criminal justice system to use the information available to them (only some of which is observed by researchers) to predict risk and to assign punishments based on those predictions. As a result, it is impossible without additional strong assumptions to distinguish the 'true' behavior of individual offenders from the behavior that results from their non-random treatment within the existing system.").

Table 5 reports summary statistics for our estimation sample, both overall and separately by race. The typical released defendant in New York City is 31.8 years old and has 2.0 prior misdemeanor convictions, 0.5 prior felony convictions, and 1.6 prior failures to appear. Fifty percent of released defendants also have a prior violent felony conviction, with 12% having a violent felony charge on the current case. Nineteen percent are charged with at least one drug charge, 6.0% with at least one DUI charge, 9.0% with at least one property charge, and 43.0% with at least one violent charge. Twenty-three percent are charged with other types of offenses, including prostitution, gambling, and public order offenses.

In terms of outcomes, 82.0% of released defendants are released ROR at arraignment, with the remaining 18.0% released on money bail of some sort. Fifteen percent of released defendants do not appear at one or more court appearances on the current case, while 27.0% are rearrested prior to case disposition.

Compared to released white defendants, released Black defendants have 1.1 more prior misdemeanor convictions, 0.4 more prior felony convictions, and 1.0 more prior failures to appear. Released Black defendants are also 4.0 percentage points more likely to have a violent felony charge on the current case. Released Black defendants are also arrested in counties with $8,200 lower income than released white defendants, largely reflecting the difference in where these defendants reside. Finally, released Black defendants are 1.0 percentage point more likely to be released ROR compared to released white defendants, but are 6.0 percentage points more likely to not appear at court and 11.0 percentage points more likely to be rearrested prior to case disposition.

C. *Proxy Effects in Commonly Used Algorithms*

This Section argues that commonly used algorithms in the criminal justice system result in unwarranted racial gaps under the mainstream legal position. These commonly used algorithms do so because they include variables that, in practice, are highly correlated with race, such as criminal history and current charge. In doing so, these algorithms use inputs that are "almost tantamount to using race,"[271] which introduces proxy effects in forming predictions, generating arguably unwarranted racial disparities.

To demonstrate how proxy effects infiltrate commonly used algorithms, we focus on a statistical model that is inspired by one of the most prominent models in the pretrial context, the Arnold Ventures PSA. The PSA was designed with the goal of being both objective and fair, "not contain[ing] factors that would lead defendants to be treated differently because of their race, gender, or socioeconomic status."[272] For this reason, the PSA excludes

---

271.   O'Neil, *supra* note 71.
272.   Milgram et al., *supra* note 134, at 220.

factors that Arnold Ventures deem to be inconsistent with fairness under the law, including race, gender, socioeconomic status, and neighborhood, or what we might call $X_i^{Protected}$. However, the PSA does include inputs such as prior criminal history and detailed charge characteristics that may or may not be correlated with protected characteristics such as race, or what we call $X_i^{Correlated}$ and $X_i^{Uncorrelated}$. Under the legal consensus of fairness, we believe the PSA's mission statement of not treating individuals differently because of race depends on whether inputs like prior criminal history are correlated with race. If, in fact, inputs like prior criminal history are not correlated with race, proxy effects will not be present, allowing us to form risk predictions that are truly race neutral. If, however, these inputs are correlated with race, unwarranted disparities will emerge as a result of proxy effects.

TABLE 5: DESCRIPTIVE STATISTICS

	All Defendants	White Defendants	Black Defendants
Panel A: Defendant Characteristics	(1)	(2)	(3)
Defendant Age	31.8	34.0	31.1
Violent Felony Charge	0.12	0.09	0.13
Prior Misdemeanor Convictions	2.00	1.09	2.29
Prior Felony Convictions	0.50	0.22	0.59
Prior Violent Felony Convictions	0.15	0.06	0.17
Prior Failures to Appear	1.64	0.86	1.89
County Income	78,300	84,500	76,300
Drug Charge	0.19	0.18	0.20
DUI Charge	0.06	0.12	0.04
Property Charge	0.09	0.10	0.09
Violent Charge	0.43	0.39	0.44
Other Charge	0.23	0.21	0.23
Panel B: Arraignment Outcomes			
Released Before Trial	1.00	1.00	1.00
ROR at Arraignment	0.82	0.82	0.83
Money Bail at Arraignment	0.18	0.18	0.17
Panel C: Pretrial Outcomes			
Failure to Appear	0.15	0.10	0.16
Rearrest Prior to Disposition	0.27	0.19	0.30
Observations	264,379	63,880	200,499

Note: This table reports descriptive statistics for the sample of defendants from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. See the text for additional details on the specification and sample.

The key question we now consider here is whether the inputs similar to those used in the Arnold Ventures PSA are, in fact, $X_i^{Uncorrelated}$ or $X_i^{Correlated}$ in real-world data. We test whether the types of input variables used in the PSA are $X_i^{Uncorrelated}$ or $X_i^{Correlated}$ in two ways. First, we examine whether each potential input variable is correlated with race by regressing an indicator for a defendant being Black on each of these variables. These regressions allow us to assess whether being Black is significantly associated or correlated with other characteristics, such as having a prior conviction. Due to data constraints, we are not able to use identical inputs as the PSA. To be as consistent as possible with the PSA, we consider the following input variables available in our data: defendant age; an indicator for whether the current charge is for a violent felony; the number of past misdemeanor convictions; the number of past felony convictions; the number of past violent felony convictions; the number of prior failures to appear; average income in the county of arrest; and

indicators for whether the current charge includes a drug, DUI, property, or violent charge. Our statistical model does not include inputs like education or employment status, which some commonly used algorithms use.[273] Again, if a potential input variable is uncorrelated with race (but correlated with pretrial misconduct), then it is $X_i^{Uncorrelated}$ in our statistical framework. If, on the other hand, a variable is correlated with race, then it is $X_i^{Correlated}$ in our statistical framework.

Table 6 presents the results from this first empirical test using our dataset on released male Black and released male white defendants from New York City. Columns 1–8 present tests of the independent correlation between defendant race and the listed input variables using these data. Column 9 presents a test of the joint correlation between defendant race and all of the listed input variables. The results show that *all* of the listed input variables are significantly correlated with defendant race, both independently and jointly. We find, for example, that Black defendants are both younger and more likely to have a violent felony charge compared to white defendants, correlations that will lead to proxy effects were these input variables to be included in an algorithm. Black defendants also tend to have more prior convictions, come from counties with lower incomes, be less likely to have DUI and property charges, and be more likely to be charged with a violent offense—again, correlations—that will lead to proxy effects if these inputs are included.

---

273.   *See supra* Table 1.

TABLE 6: CORRELATION BETWEEN RACE AND ALGORITHMIC INPUTS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Dependent Variable: Indicator for being Black				
Defendant Age in 10s	-0.034***								-0.051***
	(0.001)								(0.001)
Violent Felony Charge		0.071***							0.039***
		(0.002)							(0.002)
Prior Misdemeanor Convictions			0.006***						-0.002***
			(0.000)						(0.000)
Prior Felony Convictions				0.052***					0.039***
				(0.001)					(0.001)
Prior Violent Felony Convictions					0.096***				0.036***
					(0.002)				(0.002)
Prior Failures to Appear						0.021***			0.019***
						(0.000)			(0.000)
County Income in 10,000s							-0.019***		-0.017***
							(0.000)		(0.000)
Drug Charge								0.001	-0.028***
								(0.003)	(0.002)
DUI Charge								-0.283***	-0.223***
								(0.004)	(0.004)
Property Charge								-0.041***	-0.052***
								(0.003)	(0.003)
Violent Charge								0.006**	0.002
								(0.002)	(0.002)
Constant	0.868***	0.750***	0.746***	0.732***	0.744***	0.725***	0.911***	0.776***	1.020***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.003)
Observations	264379	264379	264379	264379	264379	264379	264379	264379	264379
R²	0.010	0.003	0.008	0.019	0.011	0.021	0.016	0.025	0.085
Independent Variable Mean	3.18	0.12	2.00	0.50	0.15	1.64	7.83	-	-

Note: This table reports the correlation between race and algorithmic inputs using information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. The dependent variable is an indicator for the defendant being Black. Each column reports results from an OLS regression of an indicator for being Black on the listed inputs. See the text for additional details on the specification and sample.

Our second test examines how the weight on each input variable changes when we account for proxy effects by regressing an indicator for failing to appear at court on all input variables, both with and without an additional control for defendant race that removes any potential proxy effects. Recall from our hypothetical example in Table 4 that there are no proxy effects when we control for all input variables *and* defendant race. Thus, we can test whether an input variable is contaminated by race by comparing how the coefficient on an input variable changes once we control for defendant race compared to when we do not control for defendant race. The magnitude of the change in coefficients is captured by the standard omitted-variable-bias (OVB) formula described previously in Equation 7.

Table 7 presents the results from this second empirical test using the same dataset on released Black and released white defendants from New York City. Column 1 presents results that include the full set of input variables, including defendant race—the benchmark statistical model. Each input variable is significantly associated with the outcome variable: failure to appear. In particular, Column 1 of Table 7 shows that there is a statistically significant relationship between race and the probability of failure to appear, with our estimates suggesting that Black defendants are 3.5 percentage points more likely to not appear at court compared to otherwise similar white defendants, the direct effect of race.

Column 2 presents results from the commonly used algorithm in the spirit of the PSA that uses the same set of nonrace input variables, but excluding defendant race. We call this the "excluding-race" model. Column 3 reports the difference between the estimated coefficients for the two statistical models. Consistent with our results from Table 6, we see that models like the PSA include significant information about defendant race through the proxy effects of other input variables. For example, being ten years older is associated with a 2.6 percentage point lower probability of failure to appear in the benchmark model, but is associated with a 2.8 percentage point lower probability of failure to appear when race is excluded. Another coefficient that changes substantially is the weight given to a current DUI charge. Compared to other charges, a defendant charged with a DUI is 6.4 percentage points less likely to fail to appear under the benchmark model, but 7.2 percentage points less likely to fail to appear when race is excluded. These predictive weights change across the two models precisely because our input variables are contaminated by race. In other words, simply excluding race from a regression, as done under commonly used algorithms, does not eliminate the proxy effects of race when correlated inputs are included, and can generate unwarranted racial disparities.

TABLE 7: COMPARISON OF BENCHMARK AND RACE-BLIND STATISTICAL
MODELS

	Dependent Variable: Failure to Appear		
	Benchmark Model	Excluding Race	Difference (1) - (2)
	(1)	(2)	(3)
Defendant Age in 10s	-0.026*** (0.001)	-0.028*** (0.001)	0.002*** (0.000)
Violent Felony Charge	-0.056*** (0.002)	-0.054*** (0.002)	-0.001*** (0.000)
Prior Misdemeanor Convictions	-0.003*** (0.000)	-0.003*** (0.000)	0.000*** (0.000)
Prior Felony Convictions	-0.010*** (0.001)	-0.008*** (0.001)	-0.001* (0.000)
Prior Violent Felony Convictions	0.007*** (0.002)	0.008*** (0.002)	-0.001*** (0.000)
Prior Failures to Appear	0.024*** (0.000)	0.024*** (0.000)	-0.001*** (0.000)
County Income in 10,000s	0.005*** (0.000)	0.004*** (0.000)	0.001*** (0.000)
Drug Charge	-0.038*** (0.002)	-0.039*** (0.002)	0.001*** (0.000)
DUI Charge	-0.064*** (0.003)	-0.072*** (0.003)	0.008*** (0.000)
Property Charge	-0.013*** (0.003)	-0.015*** (0.003)	0.002*** (0.000)
Violent Charge	-0.056*** (0.002)	-0.056*** (0.002)	-0.000 (0.000)
Black	0.035*** (0.002)		
Constant	0.179*** (0.003)	0.215*** (0.003)	-0.036*** (0.002)
Observations	264379	264379	---
R^2	0.045	0.043	---

Note: This table uses information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. The dependent variable is an indicator for failing to appear. Columns 1 and 2 report results from an OLS regression of an indicator for pretrial failure to appear on the listed inputs. Column 3 reports the difference in the coefficients for each variable between Column 1 and Column 2. See the text for additional details on the specification and sample.

Overall, the results from this Section tell us that commonly used algorithms such as the PSA likely include information about defendant race

through the proxy effects of other input variables. Even input variables that are relatively noncontroversial in the law and policy sphere, such as current charge and prior criminal history, are contaminated by these proxy effects because of their strong correlation with race and can lead to unwarranted disparities when used in predictive algorithms. More concretely, if the goal is to have an algorithm that is free of all direct and proxy effects of race, commonly used algorithms fail to deliver. Thus, these results suggest that commonly used algorithms that purport to satisfy race neutrality through the formalistic solution of excluding problematic inputs do not in fact attain this goal.

These results also demonstrate that there are likely *no* truly uncorrelated input variables in real-world data, and, as a result, that likely *all* of the commonly used algorithms may violate core principles underlying antidiscrimination law by allowing race to contaminate predictions of risk. Thus, the results indicate that we must use alternative algorithms if we want to purge predictions of all direct and proxy effects of race.

D.   *Comparison of Different Predictive Algorithms*

We conclude this Part by showing how our two statistical solutions fare in terms of racial disparities and predictive accuracy compared to commonly used predictive algorithms using our data on released defendants from New York City.

Predictive Weights on Colorblinding-Inputs and Black-as-White Algorithms: We begin by identifying the predictive weights on each input under the colorblinding-inputs algorithm in comparison to other statistical models. Table 8 shows how the weight on each input factor used to predict pretrial risk changes depending on the type of predictive algorithm. Column 1 presents the benchmark statistical model, which includes the full set of input variables, including defendant race. Column 2 presents results from commonly used algorithms that use the same set of nonrace input variables, but exclude defendant race. Finally, Column 3 presents results under our proposed colorblinding-inputs algorithm, which also requires the inclusion of all defendant characteristics including race in the first estimation step, precisely to eliminate nonrace inputs of their proxy effects. Column 4 reports the difference in the predictive weight on each input between Columns 1 and 3.

The key takeaway from Table 8 is that the coefficients in the colorblinding-inputs model (Column 3) are, by design, identical to those under the benchmark statistical model (Column 1). This is because the coefficients from Table 8 reflect the predictive relationships that come out of the first estimation step. The colorblinding-inputs model requires that we include race, just as in the benchmark model, when estimating the coefficients on all other input variables in order to eliminate racial proxy effects, as described previously in Section IV.B. As we will show in simulations below, however, the predictions for individuals under the colorblinding-inputs model and the benchmark statistical model will not be

the same. That is because the colorblinding-inputs model does not use individual race information in the second prediction step, thereby ensuring that two individuals who differ only in terms of race will not receive different predictions under the model.

Note also that, as shown previously in Table 7, the predictive weights on each input are in general different between the benchmark statistical model (Column 1); the commonly used approach, which excludes race (Column 2); and the colorblinding-inputs model (Column 3). Again, this difference is attributable to the proxy effects that emerge when race is excluded as in Column 2, but correlated nonrace inputs are nevertheless included.

TABLE 8: COMPARISON OF BENCHMARK, COMMONLY USED, AND
COLORBLINDING-INPUTS STATISTICAL MODELS

	Dependent Variable: Failure to Appear			
	Benchmark Model	Excluding Race	Colorblinding Inputs	Difference (1) - (3)
	(1)	(2)	(3)	(4)
Defendant Age in 10s	-0.026***	-0.028***	-0.026***	0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Violent Felony Charge	-0.056***	-0.054***	-0.056***	0.000
	(0.002)	(0.002)	(0.002)	(0.000)
Prior Misdemeanor Convictions	-0.003***	-0.003***	-0.003***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Prior Felony Convictions	-0.010***	-0.008***	-0.010***	0.000
	(0.001)	(0.001)	(0.001)	(0.000)
Prior Violent Felony Convictions	0.007***	0.008***	0.007***	0.000
	(0.002)	(0.002)	(0.002)	(0.000)
Prior Failures to Appear	0.024***	0.024***	0.024***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
County Income in 10,000s	0.005***	0.004***	0.005***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Drug Charge	-0.038***	-0.039***	-0.038***	0.000
	(0.002)	(0.002)	(0.002)	(0.000)
DUI Charge	-0.064***	-0.072***	-0.064***	0.000
	(0.003)	(0.003)	(0.003)	(0.000)
Property Charge	-0.013***	-0.015***	-0.013***	0.000
	(0.003)	(0.003)	(0.003)	(0.000)
Violent Charge	-0.056***	-0.056***	-0.056***	0.000
	(0.002)	(0.002)	(0.002)	(0.000)
Black	0.035***		0.035***	0.000
	(0.002)		(0.002)	(0.000)
Constant	0.179***	0.215***	0.179***	0.000
	(0.003)	(0.003)	(0.003)	(0.000)
Observations	264379	264379	264379	---
R^2	0.045	0.043	0.045	---

Note: This table reports the correlation between failure to appear and algorithmic inputs using information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. The dependent variable is an indicator for failing to appear. Columns 1–3 report results from an OLS regression of an indicator for pretrial failure to appear on the listed inputs. Column 4 reports the difference in the coefficients for each variable between Column 1 and Column 3. See the text for additional details on the specification and sample.

We now show the predictive weights on each input under the minorities-as-whites algorithm (or Black-as-white algorithm in our setting), in comparison to other statistical models. Table 9 presents these results. Column 1 presents the benchmark statistical model, which includes the full set of input variables, including defendant race. Column 2 presents results from commonly used algorithms that exclude defendant race. And Column 3 presents results under our proposed Black-as-white algorithm, which applies the whites-only predictive relationship between each input and the outcome of interest for all defendants, both white and Black. Column 4 reports the difference in the predictive weight on each input between Columns 1 and 3.

Table 9 reveals that in general, a Black-as-white algorithm will yield substantially different predictive weights on each input relative to both the benchmark statistical model and the commonly used approach that excludes race. Intuitively, these weights will differ because the Black-as-white algorithm is only estimating the relationship between each input and the outcome of interest within one population of defendants. For example, under the benchmark statistical model, a defendant who is ten years older is associated with a 2.6 percentage-point reduction in the probability of failing to appear. And under the commonly used approach, being ten years older is associated with a 2.8 percentage-point decrease in the probability of failing to appear. But under the Black-as-white model, a defendant who is ten years older is associated with only a 1.2 percentage-point reduction in the probability of failing to appear.

TABLE 9: COMPARISON OF BENCHMARK, COMMONLY USED, AND BLACK-AS-WHITE STATISTICAL MODELS

	Dependent Variable: Failure to Appear			
	Benchmark Model	Excluding Race	Black as White	Difference (1) - (3)
	(1)	(2)	(3)	(4)
Defendant Age in 10s	-0.026***	-0.028***	-0.012***	-0.014***
	(0.001)	(0.001)	(0.001)	(0.001)
Violent Felony Charge	-0.056***	-0.054***	-0.031***	-0.025***
	(0.002)	(0.002)	(0.004)	(0.003)
Prior Misdemeanor Convictions	-0.003***	-0.003***	-0.004***	0.001
	(0.000)	(0.000)	(0.001)	(0.001)
Prior Felony Convictions	-0.010***	-0.008***	-0.014***	0.004
	(0.001)	(0.001)	(0.003)	(0.003)
Prior Violent Felony Convictions	0.007***	0.008***	0.002	0.005
	(0.002)	(0.002)	(0.005)	(0.005)
Prior Failures to Appear	0.024***	0.024***	0.028***	-0.004***
	(0.000)	(0.000)	(0.001)	(0.001)
County Income in 10,000s	0.005***	0.004***	0.002***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
Drug Charge	-0.038***	-0.039***	-0.002	-0.036***
	(0.002)	(0.002)	(0.004)	(0.004)
DUI Charge	-0.064***	-0.072***	-0.042***	-0.022***
	(0.003)	(0.003)	(0.004)	(0.003)
Property Charge	-0.013***	-0.015***	0.026***	-0.039***
	(0.003)	(0.003)	(0.005)	(0.005)
Violent Charge	-0.056***	-0.056***	-0.028***	-0.028***
	(0.002)	(0.002)	(0.003)	(0.003)
Black	0.035***			
	(0.002)			
Constant	0.179***	0.215***	0.122***	0.057***
	(0.003)	(0.003)	(0.006)	(0.005)
Observations	264379	264379	63880	---
R^2	0.045	0.043	0.033	---

Note: This table reports the correlation between failure to appear and algorithmic inputs using information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. The dependent variable is an indicator for failing to appear. Columns 1–3 report results from an OLS regression of an indicator for pretrial failure to appear on the listed inputs. Column 4 reports the difference in the coefficients for each variable between Column 1 and Column 3. See the text for additional details on the specification and sample.

Racial Disparities and Predictive Accuracy: We now evaluate the performance of our proposed statistical solutions relative to other algorithms by measuring racial disparities in pretrial release. To evaluate our statistical algorithms, we use the estimates from Table 8 and Table 9 (which reflect the first estimation step) to construct risk predictions for every defendant in our sample under each algorithm (the second prediction step). Recall from Part IV that both our proposed algorithms do not use individual-level race information to form risk predictions in the second step.

Having formed risk predictions under each algorithm, we then simulate different release policies, calculating the fraction of Black (versus white) defendants among those released and the FTA rate among the released under each hypothetical policy. The goal of each algorithm is to have the lowest possible FTA rate and *no* unwarranted disparities between Black and white defendants. Recall that we define unwarranted racial disparities as differences in the treatment of otherwise similar individuals due solely to membership in a particular racial group, either through direct or proxy effects of race. Again, we view this definition as most consistent with the mainstream legal view of fairness. The goal of each algorithm is not, however, to release an equal number of Black and white defendants. Under the law, racial disparities are not illegal per se,[274] but rather only those disparities driven by race or motivated by a discriminatory purpose.

To examine racial disparities in pretrial release, Figure 1 reports the share of released defendants who are Black if we were to make pretrial release decisions using each of the different predictive algorithms. The x-axis in Figure 1 varies the percentage of all defendants that are released, ranging from 0 to 100%—what we call the "release threshold." The y-axis reports the fraction of released defendants that are Black at each release threshold. As a reference, 76% of the estimation sample is comprised of Black defendants, which is what the y-axis would report when the x-axis is at 100% release. We consider four total algorithms: (1) the benchmark statistical model that uses all inputs, including race, (2) the commonly used model that uses all inputs but excluding race, (3) our proposed colorblinding-inputs model, and (4) our proposed Black-as-white model.

This figure reveals that the benchmark statistical model always results in the lowest share of Black defendants among the released at each release threshold. This occurs because the benchmark statistical model uses race as a predictive input, giving rise to direct race effects. Given that being Black is positively associated with the risk of failing to appear (see Table 7),[275] Black defendants will receive higher risk predictions than otherwise similar white defendants, resulting in lower rates of pretrial release for Black individuals.
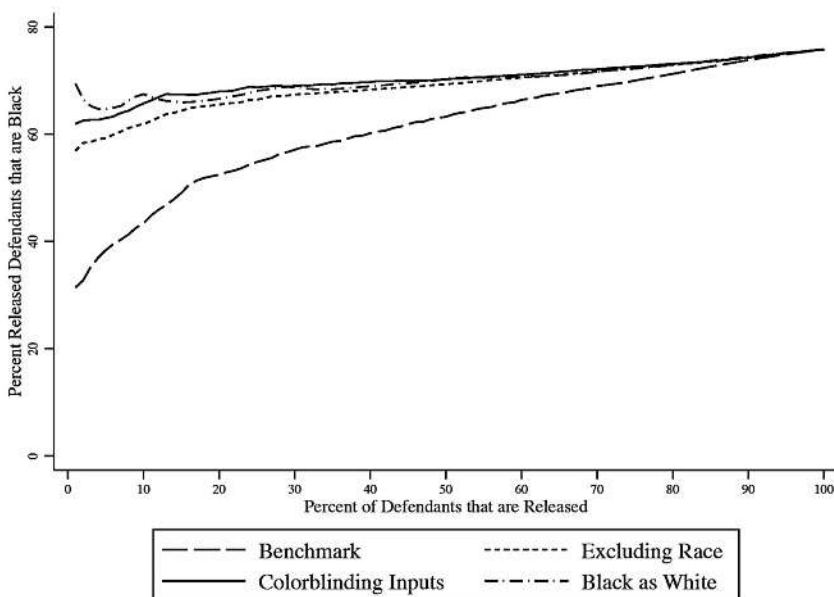
---

274. *See, e.g.*, Sunstein, *supra* note 93, at 509 ("In terms of existing law, racial balance, as such, is not legally mandated, and efforts to pursue that goal might themselves be struck down on constitutional grounds.").

275. *See supra* Table 7.

The commonly used model that excludes race improves upon this benchmark statistical model by increasing the share of released defendants that are Black. At all possible release thresholds, the commonly used model results in a higher share of released defendants that are Black relative to the benchmark statistical model. This occurs because the direct effects of race are eliminated when race is excluded as a predictive input.

However, our proposed colorblinding-inputs model results in an even higher share of Black defendants being released relative to both the benchmark model and the commonly used model. This pattern holds for all possible release rates, with the largest differences at particularly low overall release rates. The reason that our proposed colorblinding-inputs model increases the fraction of Black defendants released, regardless of the overall release rate, is that it purges all the input variables of racial proxy effects. These proxy effects are exactly what lead to the relative overdetention of Black defendants in the commonly used algorithm. Similarly, our proposed Black-as-white algorithm generally results in a higher share of Black defendants released relative to the commonly used model. These results indicate that racial disparities in pretrial detention can be further reduced under our proposed statistical solutions relative to the typical algorithm used in practice today.

FIGURE 1: RACIAL DISPARITIES UNDER DIFFERENT PREDICTIVE ALGORITHMS

Note: This figure plots the percent of released defendants who are black under different predictive algorithms and release rates using information from the New York City pretrial system. The sample consists of male black and white defendants who were arrested and charged between 11/2018 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. See the text for additional details on the specification and sample.

To provide some more concrete examples of the differences in the racial composition of released defendants across the various algorithms, Table 10 presents a selected subset of these simulations to precisely illustrate the differences in racial disparities among the types of algorithms. We consider hypothetical scenarios where we release 50, 70, or 90% of all individuals in our data and report the share of Black defendants among the released. Columns 1–4 report the fraction of individuals released that are Black under the benchmark, excluding race, colorblinding-inputs, and Black-as-white models, respectively. Column 5 reports the difference in share released that are Black between the commonly used excluding-race model and the colorblinding-inputs model. Column 6 reports the difference in share released that are Black between the commonly used excluding-race model and the Black-as-white model.

These results again show that our proposed colorblinding-inputs statistical model would significantly increase the fraction of Black individuals released compared to both the benchmark statistical model and the commonly used model that simply excludes race as a predictive input. The use of the benchmark model would, for example, lead to 63.6% of released defendants being Black if the overall release rate was set at a threshold of 50% (Column 1). The commonly used model would increase the fraction of released defendants who are Black to 69.5% (Column 2), consistent with the fact that the benchmark model penalizes Black defendants by allowing for direct race effects.

However, our proposed colorblinding-inputs model further increases the fraction of released defendants who are Black to 70.4% (Column 3), almost a full percentage-point increase compared to the commonly used algorithm (Column 5). If applied citywide, a back-of-the-envelope calculation implies this model would release an additional 3,500 Black defendants during our sample period compared to the typical algorithm used today if 50% of all defendants are released. In a similar nature, our proposed Black-as-white model increases the fraction of released defendants who are Black by 0.8 percentage points relative to the commonly used algorithm (Columns 4 and 6), which could lead to the release of an additional 3,100 Black defendants during our sample period. As a reference point, a typical jursidiction releases between 30 and 70% of its pretrial population,[276] so we view these changes as realistic.

---

276.   For example, "[b]etween 1990 and 2004, 62% of felony defendants in State courts in the 75 largest counties were released prior to the disposition of their case." THOMAS H. COHEN & BRYAN A. REAVES, U.S. DEP'T OF JUST., NCJ 214994, BUREAU OF JUSTICE STATISTICS SPECIAL REPORT: PRETRIAL RELEASE OF FELONY DEFENDANTS IN STATE COURTS (2007), https://www.bjs.gov/content/pub/pdf/prfdsc.pdf [http://perma.cc/2VQA-APWJ]. Average pretrial detention rates range from 30% to 70% for misdemeanor defendants in Harris County. Paul Heaton, Sandra Mayson & Megan Stevenson, *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 STAN. L. REV. 711, 736–37 tbl.1 (2017).

We find that these increases in the number of released Black defendants persist even at very high release rates. For example, if a city wanted to release 90% of all defendants, a release threshold that is substantially higher than currently used in most jurisdictions, both our proposed algorithms would continue to lead to nonnegligible increases in the number of released Black defendants relative to the commonly used algorithm.

Recall that our measure of pretrial misconduct is failing to appear in court given that nonappearance at court is by law the only legitimate concern taken into account at the time of setting bail in New York City.[277] However, our results are similar if we use our proposed algorithms to predict the risk of being arrested for a new crime prior to case disposition. The Appendix presents these results and simulations. For instance, if the overall release rate was set at a threshold of 50%, our proposed colorblinding-inputs model would lead to the release of an additional 3,300 Black defendants and our proposed Black-as-white algorithm would lead to the release of an additional 7,600 Black defendants compared to the commonly used algorithm.

TABLE 10: SIMULATIONS OF RACIAL DISPARITIES UNDER DIFFERENT PREDICTIVE ALGORITHMS

	Share of Black Defendants Among Released					
	Benchmark Model	Excluding Race	Colorblinding Inputs	Black as White	Difference (2) - (3)	Difference (2) - (4)
	(1)	(2)	(3)	(4)	(5)	(6)
50% Release Rate	63.61	69.49	70.40	70.29	-0.92	-0.81
70% Release Rate	69.01	71.83	72.20	71.75	-0.37	0.08
90% Release Rate	73.84	74.28	74.33	74.49	-0.05	-0.21

Note: This table reports the percent of released defendants who are Black versus white under different prediction models and release rates using information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. Column 1 reports the percent Black released among released defendants under the benchmark statistical model. Column 2 reports the percent Black released among released defendants under the commonly used model. Column 3 reports the percent Black released among released defendants under the colorblinding-inputs model. Column 4 reports the percent Black released among released defendants under the Black-as-white model. Column 5 reports the difference in the percent Black released defendants between the commonly used and the colorblinding-inputs model. Column

---

277.   *See* N.Y. CRIM. PROC. LAW § 510.30 (McKinney 2020) ("With respect to any principal, the court in all cases, unless otherwise provided by law, must impose the least restrictive kind and degree of control or restriction that is necessary to secure the principal's return to court when required."); *see also* Sardino v. State Comm'n on Jud. Conduct, 448 N.E.2d 83, 84 (N.Y. 1983) (in New York, the "only matter of legitimate concern" when setting bail is "whether any bail or the amount fixed was necessary to insure the defendant's future appearances in court").

6 reports the difference in the percent Black released defendants between the commonly used and Black-as-white model. See the text for additional details on the specification and sample.

FIGURE 2: ACCURACY UNDER DIFFERENT PREDICTIVE ALGORITHMS
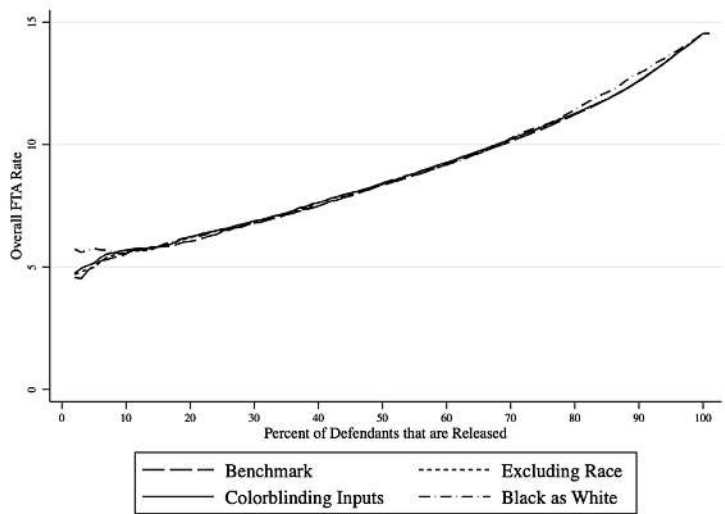
Note: This figure simulates the failure to appear rates for defendants who would be released under each predictive model using information from the New York City pretrial system. The sample consists of male black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. See the text for additional details on the specification and sample.

We would be remiss to not also illustrate that the choice of predictive algorithm comes with trade-offs in terms of accuracy, as mentioned in Section I.C. Recall that the benchmark statistical model, which uses all inputs, maximizes predictive accuracy. As we begin to eliminate direct effects of race (as under the commonly used algorithm), and then both direct and proxy effects of race (as under our proposed statistical solutions), accuracy decreases. Reducing unwarranted disparities requires the statistical model to "ignore" potentially relevant information, such as race or other inputs that are correlated with race. Under the particular definition of fairness outlined in this Article, an algorithm that eliminates both direct and proxy effects of race, thereby increasing the number of released Black defendants, is "fair" even if it comes at a cost to predictive accuracy.

To illustrate how accuracy changes across the different algorithms, Figure 2 reports the overall FTA rates if we were to make pretrial-release decisions using each of the four different predictive algorithms. Here, we measure FTA rates as our outcome, where one algorithm is more accurate than another if the FTA rate among released defendants is lower. For example, if 50% of defendants are released, and algorithm A results in a 20% FTA rate among the released and algorithm B results in a 30% FTA rate, we would say that algorithm A is superior in terms of predictive accuracy.

Consistent with our statistical framework, FTA rates are lowest for the benchmark statistical model that uses all available information, followed by the commonly used model and then the colorblinding-inputs and Black-as-white model. These patterns generally hold at all release thresholds. The reason that the benchmark statistical algorithm is most accurate is precisely because it explicitly uses race to generate predictions, and race is highly correlated with risk of FTA. For a similar reason, the commonly used model is generally more accurate than our proposed solutions because it retains some information on defendant race through proxy effects.

Table 11 presents a selected subset of our simulations to precisely illustrate the trade-off between our definition of fairness and predictive accuracy. Again, we consider hypothetical scenarios where we release 50, 70, or 90% of all individuals in our data. We present the simulated FTA rate among all released defendants under each hypothetical, where Columns 1–4 report the simulated FTA rates under the different predictive algorithms. Column 5 reports the difference in FTA rates between the commonly used excluding-race model and the colorblinding-inputs model, and Column 6 reports the difference in FTA rates between the commonly used model and the Black-as-white model. The results again show that predictive accuracy is maximized by the benchmark algorithm that explicitly includes race.

We note that the differences in accuracy among the models, in particular between the commonly used excluding-race algorithm and our proposed algorithms, are economically small. For example, if applied citywide in New York City, the colorblinding-inputs model would result in an additional seventy-six failures to appear during our sample period compared to the commonly used algorithm if the city decided to release 50% of all defendants.

TABLE 11: SIMULATIONS OF ACCURACY UNDER DIFFERENT PREDICTIVE
ALGORITHMS

	FTA Rate Among Released Defendants					
	Benchmark Model	Excluding Race	Colorblinding Inputs	Black as White	Difference (2 - (3)	Difference (2) - (4)
	(1)	(2)	(3)	(4)	(5)	(6)
50% Release Rate	8.35	8.38	8.40	8.42	-0.02	-0.04
70% Release Rate	10.13	10.20	10.23	10.29	-0.03	-0.09
90% Release Rate	12.58	12.61	12.60	12.90	0.01	-0.29

Note: This table simulates the failure to appear rates for defendants who would be released under each predictive model using information from the New York City pretrial system. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. Column 1 reports the FTA rate under the benchmark statistical model. Column 2 reports the FTA rate under the commonly used model. Column 3 reports the FTA rate under the colorblinding-inputs model. Column 4 reports the FTA rate under the Black-as-white model. Column 5 reports the difference in FTA rates between the commonly used and the colorblinding-inputs model. Column 6

Column 1 reports the FTA rate under the benchmark statistical model. Column 2 reports the FTA rate under the commonly used model. Column 3 reports the FTA rate under the colorblinding-inputs model. Column 4 reports the FTA rate under the Black-as-white model. Column 5 reports the difference in FTA rates between the commonly used and the colorblinding-inputs model. Column 6 reports the difference in FTA rates between the commonly used and the Black-as-white model. See the text for additional details on the specification and sample.

VI. EXTENSIONS

In this Part, we describe a number of potential extensions to our analysis and results. We begin by discussing how our proposed algorithms can incorporate additional protected characteristics such as gender. We then discuss nonlinear prediction models, prediction models with nonrace interactions, and prediction models that explicitly allow for race interactions. We conclude by discussing how our proposed statistical models are relevant to other contexts outside the criminal justice system.

A. *Additional Protected Characteristics*

Our proposed solutions can be easily adapted to deal with other protected characteristics. For example, many scholars have also decried the use of gender in forming predictions of risk, where generally men receive higher risk predictions than women.[278] Our proposed solutions can remove both the direct and proxy effects of both race and gender, or any other protected characteristic.

Specifically, a color- *and* genderblinding-inputs algorithm could be estimated in a similar two-step procedure as described previously in Section IV.B. In the first step, we would again estimate the benchmark statistical model that includes the full set of input characteristics (including both race and gender and all correlates). Including race and gender allows us to eliminate the proxy effects on all other inputs that are correlated with both race and gender. Then, in the second prediction step, to ensure that no direct effects of race and gender are used, we would simply use the average race or gender across all individuals to form predictions. This algorithm purges the predictions of both direct and proxy effects along both racial and gender dimensions, allowing for a race- and gender-neutral model.

Our minorities-as-whites solution can also be adapted to deal with other protected characteristics. If gender were a concern, we could construct an algorithm that treated all individuals the same as, say, white women, using the two-step procedure described in Section IV.C. Specifically, we would estimate the relationship between each input and the outcome of interest within a white female population, and then apply the same estimations to form predictions for all nonwhite and nonfemale individuals. Or, one could construct an algorithm that treated all individuals the same, as say, white males. As before, this algorithm does rely on a normative choice of which

---

    278.   *See, e.g.*, Starr, *supra* note 12, at 806; Sidhu, *supra* note 22, at 699–700.

group should be the baseline population for the first estimation step, a decision that we think should be informed by which population is least likely to have mismeasurement in inputs or outcomes (due to discrimination or otherwise).

B.   *More Complicated Algorithms*

Our proposed solutions can also be easily adapted to deal with more complicated predictive algorithms. Here, we consider four such extensions.

Nonlinear Prediction Models: Our main proposed algorithms assume that there is a linear relationship between each predictive input and the outcome of interest, that is, a linear probability model accurately captures the underlying statistical relationship. This modeling choice assumes, for example, that an additional year of age always has the same association with the outcome of interest (i.e., age has the same marginal effect). But one might imagine that there are nonlinearities in this relationship. A nice feature of both our proposed algorithms is that they can be easily adapted to allow for nonlinearities. In the context where the outcome of interest is a binary variable, as is almost always the case (e.g., whether a defendant fails to appear), one can estimate the underlying statistical model using a nonlinear model, such as a logit or probit model, and still be able to eliminate both direct and proxy effects of race.[279]

Prediction Models with Nonrace Interactions: For simplicity, our main proposed algorithms assumes a benchmark statistical model where each input affects the outcome of interest in a linear and additive way. This simplifying assumption assumes that there are no interaction effects between different nonrace predictive inputs. This modeling choice assumes, for example, that the relationship between age and the outcome of interest is linear and the same for all individuals. In other words, suppose that being ten years older was associated with a 5 percentage-point reduction in risk. Our proposed models assume that this relationship is true for all individuals. If, however, one believed that the relationship between age and the outcome of interest differed for groups of individuals (e.g., the relationship between age and risk is different for individuals with a prior criminal history and individuals with no priors), our approaches could easily be adapted to allow for these dynamics. Technically, one would allow for these relationships by adding interaction terms between age and prior criminal history. There is no constraint on the interactions that can be allowed between any input in $X_i^{Uncorrelated}$ or $X_i^{Correlated}$. For example, suppose that one wants to use this benchmark statistical model:

---

279.   For a discussion of how to purge both proxy and direct effects of protected charac- teristics from a logit or probit model, see Pope & Sydnor, *supra* note 23, at 215.

$$Y_i = \beta_0 + \beta_1 \cdot X_i^{Uncorrelated} + \beta_2 \cdot X_i^{Correlated}$$
$$+ \beta_3 \cdot X_i^{Uncorrelated} \cdot X_i^{Correlated} + \beta_4 \cdot X_i^{Protected} + \epsilon_i$$

(Equation 14)

where β_1 captures the uninteracted effect of $X_i^{Uncorrelated}$, β_2 captures the uninteracted effect of $X_i^{Correlated}$, and β_3 captures the interacted effect of $X_i^{Uncorrelated}$ and $X_i^{Correlated}$. Both our proposed algorithms can be readily adapted to allow for these interactions and still purge predictions of both direct and proxy effects of race. Under the colorblinding-inputs algorithm, for example, one could estimate this benchmark model controlling for race in the first estimation step and then use $\overline{X}^{Protected}$ in the second prediction step as described in Section IV.B.

Prediction Models Explicitly Allowing for Race Interactions: One might also want to explicitly allow interaction effects between a nonrace predictive input and race itself in the underlying statistical model. For example, one might want to generate predictions assuming that the relationship between age and risk differs by defendant race. At the extreme, if one were to fully interact each nonrace input with race, the predictive algorithm would be equivalent to estimating separate risk predictions for white and Black individuals. Unlike our proposed solutions, this approach would be akin to using race in both the first estimation step and second prediction step. Such an approach is similar to that proposed by Jon Kleinberg et al., which fully utilizes the predictive power of all input factors, including protected characteristics.[280] Under their approach, one would allow the coefficients, or "slopes," on the full set of input factors to differ by race.[281] This "race-interacted" algorithm will therefore have a higher level of predictive accuracy compared to our proposed models, as it allows for a more flexible relationship between input factors and the outcome of interest.

This race-interacted model, however, poses legal issues because it explicitly allows otherwise similar individuals to receive different predictions on the basis of race. It also highlights the important trade-off between our ability to eliminate unwarranted racial disparities ("fairness") and predictive accuracy, as we previously described in Section I.C. Most often, an unconstrained race-interacted model will yield larger racial disparities in predictions if Black defendants are statistically "riskier" than white defendants. But predictive accuracy is likely enhanced by allowing a more flexible underlying statistical model.

However, new approaches in computer science and economics indicate that there are ways of eliminating the theoretical trade-off between accuracy

---

280. Kleinberg et al., *supra* note 23, at 22–23.

281. *Id.* at 24.

and fairness, which sharply contrasts with the general view among legal scholars that the "only way to ensure that decisions do not systematically disadvantage members of protected classes is to reduce the overall accuracy of all determinations."[282] For example, the existing computer science and economics literature has regularly argued that "[a]bsent legal constraints, one should include variables such as gender and race for fairness reasons. . . . [T]he inclusion of such variables can increase both equity and efficiency."[283] How could an algorithm that uses race in forming predictions not increase racial disparities? One approach, as suggested by Kleinberg et al., is to achieve the desired racial composition by "setting a different threshold for different groups,"[284] an approach that explicitly requires disparate treatment of individuals and constitutes ex post racial balancing. Under this approach, one can use the more accurate risk predictions from the race-interacted algorithm but fix the racial composition ex post to the desired level, which can improve upon predictive accuracy because "society is still served best by ranking as well as possible using the best possible predictions."[285]

While we are in favor of this approach from a statistical perspective (in the sense that it can achieve the same desired racial composition at higher accuracy), we have concerns about its legality given that it would require explicit ex post racial balancing or fixing of a racial quota. As Deborah Hellman as similarly argued, "[t]he use of different cut scores would consti-tute disparate treatment on the basis of race."[286] Such an approach may run into a potential challenge given the Supreme Court's 2009 decision in *Ricci v. DeStefano*.[287] In *Ricci*, the city of New Haven, Connecticut administered exams to be used in promoting the city's firefighters. After exams were taken, the city noted that using the exams would result in a racially disparate impact because no Black candidates would have been immediately eligible for promotion on the basis of the exam results.[288] Thus, to avoid disparate impact liability under Title VII, the city decided to throw out the exams after some firefighters threatened to sue if promotions were based on the exam scores, alleging that the tests were discriminatory.[289] A group of white and Hispanic firefighters who would have been promoted based on their exam performance then sued the city, alleging that the city's refusal to use the exams subjected them to disparate treatment on the basis of race in violation

---

282.   Barocas & Selbst, *supra* note 8, at 721–22.

283.   *See, e.g.*, Kleinberg et al., *supra* note 23, at 23.

284.   *Id.* at 23–25.

285.   *Id.* at 23.

286.   Hellman, *supra* note 174, at 848.

287.   557 U.S. 557 (2009).

288.   *Ricci*, 557 U.S. at 562.

289.   *Id.*

of both Title VII and the Equal Protection Clause.[290] A five-person majority of the Court held that the city's race-based action violated Title VII, constituting disparate treatment, because there was no strong basis in evidence that the city would have been subject to disparate impact liability had it not thrown out the exams.[291] Thus, *Ricci* suggests that typically a decisionmaker cannot engage in disparate treatment on the basis of a protected characteristic in order to avoid a disparate impact.

Other scholars like Harned and Wallach share the view that directly using race to make decisions (such as in the second prediction step) would constitute disparate treatment and thus be illegal.[292] Moreover, these authors argue that even if the ultimate goal of directly using protected traits is to mitigate bias, "the stigma involved in racial classifications can constitute a cognizable harm" and "[o]nly in very limited circumstances can two different standards . . . be legally applied on the basis of a sensitive attribute."[293] In contrast, Jason Bent has argued the direct use of race to make decisions should be a form of legally permissible "[a]lgorithmic [a]ffirmative [a]ction," and that there is a "significant difference between *discarding* a biased algorithm and *fixing* a biased algorithm by introducing a race-aware fairness constraint."[294]

Machine-Learning Models: Recall that we have illustrated our two proposed solutions when the underlying statistical model is linear, which we believe captures the state of most algorithms used in the criminal justice system. Above, we described how our proposals can also work easily using nonlinear models such as logit or probit. But a natural question is how our approaches might work using a machine-learning algorithm. In recent work, Talia Gillis argues that one of our proposed solutions, the colorblinding-inputs algorithm (what she refers to as the "orthogonalization" method), "goes wrong in the machine learning context."[295] Using a lasso regression, a form of machine learning that selects inputs so as to avoid overfitting the model, Gillis shows that race is chosen as an input in some training datasets but not chosen in other training datasets.[296] And even when race was chosen by the algorithm, Gillis notes that the lasso regression can give race a differ-

---

290.   *Id.* at 562–63.

291.   *Id.* at 563, 584 ("If an employer cannot rescore a test based on the candidates' race, § 2000e-2(*l*), then it follows *a fortiori* that it may not take the greater step of discarding the test altogether to achieve a more desirable racial distribution of promotion-eligible candidates—absent a strong basis in evidence that the test was deficient and that discarding the results is necessary to avoid violating the disparate-impact provision."). The Court reserved the question of whether fear of disparate impact is ever sufficient to justify discriminatory treatment under the Equal Protection Clause of the Constitution. *Id.*

292.   Harned & Wallach, *supra* note 72 (manuscript at 21–22).

293.   *Id.* at 23.

294.   Bent, *supra* note 225, at 35.

295.   Gillis, *supra* note 103, at 72.

296.   *Id.* at 69–72.

ent weight depending on the training dataset.[297] Given this lack of "stability," Gillis argues that "[t]he orthogonalization method, which uses the coefficient or weight on 'race' for the screening stage, will therefore yield different results based on the random draw."[298] As a result, Gillis argues that lawmakers must give up on "input scrutiny."[299]

We view the approach of ceding full control to the machine as unnecessary and inconsistent with what antidiscrimination law demands. We believe input scrutiny is not only essential but also feasible with machine learning, such that Gillis is incorrect to state that "[i]n the case of machine learning, the algorithm itself determines which inputs to use and what weights to assign them in reaching an accurate prediction."[300]

If one acknowledges that the law requires human choice in the design of the algorithm, we believe that the principles of our proposed algorithms continue to work even in the context of machine learning, although more work is needed in this area. For example, as applied to our colorblinding-inputs approach, Žliobaitė and Custers, like us, believe that while

> [f]ormal conclusion for non-linear models remains a subject for future investigation. . . . [O]ur intuition from working in this field and observing the behaviour of various data mining and machine learning models is that similar principles apply, but to what extent, and what models are more or less sensitive, remains to be researched.[301]

Given the input-selection instability that might arise from certain forms of machine learning, the procedure for implementing the colorblinding-inputs algorithm needs to be slightly modified. But it is easy to imagine this variant of the core procedure: (1) first, we colorblind the inputs by residualizing/orthogonalizing the effect of race from each input characteristic;[302] (2) then we give the machine-learning algorithm the colorblinded inputs in a modified dataset to form estimations and predictions. This modified procedure, which at its core mirrors the two-step procedure we described in Section IV.B. when using a linear regression, is likely to work in the machine-learning context because it takes away the choice of the machine to select race as an input (as shown in Gillis's naive simulation); rather, the machine-learning algorithm is given all possible nonrace inputs that have already been purged of proxy effects. And if we turn to our second proposed "minorities-as-whites" algorithm, unaddressed by Gillis, we see no concerns with using

---

297.   *Id.*

298.   *Id.* at 72.

299.   *Id.* at 12.

300.   *Id.* at 36.

301.   Žliobaitė & Custers, *supra* note 200, at 193.

302.   Formally, one could convert all input variables into indicator variables, which addresses concerns about nonlinearities. Then one could individually "colorblind" each input by residualizing/orthogonalizing each input with respect to race. These modified inputs can then be used by the machine-learning algorithm.

machine-learning algorithms; the white baseline population data would be given directly to the algorithm, and thus the machine-learning algorithm does not have the choice to select race as an input. As a result, both our approaches can be adapted to the machine-learning context by altering the inputs or data that are given to the algorithm.

None of the ideas we propose are new to the machine-learning community. The core idea underlying our algorithms is well known to the computer science literature and often referred to as "preprocessing" the training data.[303] This concept and approach have been shown to work even with complex machine-learning models. For example, Faisal Kamiran and Toon Calders show that preprocessing techniques such as "[m]assaging the dataset" or "[r]eweighing" or "[s]ampling" the dataset can "lead to an effective decrease in discrimination with a minimal loss in accuracy," particularly as compared to simply excluding protected traits and/or their proxies.[304] Given such developed preprocessing methods, even legal scholars have advocated for these techniques as a way of regulating machine-learning algorithms to be compliant with antidiscrimination law. Cofone, for example, states that "while blocking data on protected categories is unhelpful, *shaping* the information that includes protected categories can be effective at eliminating bias from the data that decision-making models are trained with and, in turn, eliminating discrimination from such models."[305] Thus, we believe that arguments that our approaches cannot work in the context of machine learning are, at best, overstated and rely on a faulty and narrow conception of how antidiscrimination law can constrain machine learning.

C.   *Other Contexts*

Our proposed statistical solutions can be easily applied to other contexts that face similar concerns about direct and proxy effects of protected characteristics, including credit and lending decisions and employment and hiring decisions.

Credit and Lending: Take, for example, credit and lending, where federal laws prohibit discrimination on the basis of protected characteristics. For instance, the Equal Credit Opportunity Act (ECOA) of 1974 prohibits discrimination on the basis of protected characteristics such as race, gender,

---

303.   *See, e.g.*, Faisal Kamiran & Toon Calders, *Classifying Without Discriminating*, IEEE XPLORE 1–6 (2009) [hereinafter Kamiran & Calders, *Classifying Without Discriminating*], https://ieeexplore.ieee.org/document/4909197?arnumber=4909197 [https://perma.cc/LZ2W-FPM9] (describing a preprocessing approach that modifies the training data to yield an unbiased dataset); Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE INFO. SYS. 1 (2012) [hereinafter Kamiran & Calders, *Data Preprocessing Techniques*].

304.   *See* Kamiran & Calders, *Data Preprocessing Techniques*, *supra* note 303, at 2–3.

305.   Cofone, *supra* note 91, at 1432.

or national origin.[306] Regulation B of the ECOA lists many factors that cannot be used in empirically derived credit-scoring systems, including public-assistance status, marital status, race, color, religion, national origin, and sex.[307] In fact, Regulation B states that, generally, creditors may not even *request or collect* information about an applicant's race, color, religion, national origin, or sex.[308]

Scholars have summarized these laws as follows: "In essence, the law requires that lenders make decisions about mortgage loans as if they had no information about the applicant's race, regardless of whether race is or is not a good proxy for risk factors not easily observed by the lender."[309] These laws have also been interpreted to prohibit the use of "redlining," or geographic discrimination using zip codes as proxies for the racial composition of neighborhoods.[310]

As applied to predictive algorithms, legal scholars have generally interpreted these laws to preclude the direct consideration of protected characteristics such as race and gender in credit-scoring algorithms.[311] In addition, many are worried about proxy effects of these protected characteristics, noting that other traits used in credit scoring, such as social-media practices (used by newer companies to determine creditworthiness), may be proxies for protected characteristics.[312] Even arguably neutral factors commonly considered, such as amounts owed, new credit, length of credit history, credit mix, and payment history, may be highly correlated with race, generating racial proxy effects even when race itself is not directly used.[313]

---

306.   15 U.S.C. § 1691(a)(1).

307.   12 C.F.R. § 202.5 (2020).

308.   *Id.*

309.   Helen F. Ladd, *Evidence on Discrimination in Mortgage Lending*, J. ECON. PERSP., Spring 1998, at 41, 43.

310.   *Id.*

311.   *See, e.g.*, Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 182 (2016); Gillis & Spiess, *supra* note 25, at 467 ("One aspect of many antidiscrimination regimes is a restriction on inputs that can be used to price credit. Typically, this means that protected characteristics, such as race and gender, cannot be used in setting prices. Indeed, many antidiscrimination regimes include rules on the exclusion of data inputs as a form of discrimination prevention.").

312.   Hurley & Adebayo, *supra* note 311, at 163, 182–83.

313.   *See id.* at 182; Berger*, supra* note 1. In contrast, FICO does not consider characteristics such as race, color, religion, national origin, gender, and marital status, or neighborhood. *See What's Not in My FICO Scores*, MYFICO, https://www.myfico.com/credit-education /whats-not-in-your-credit-score [https://perma.cc/K2DT-SLB6]. In recent years, alternative credit-scoring companies like ZestFinance have emerged, relying on much more information than traditionally used under its motto "All data is credit data." *See* James Rufus Koren, *Some Lenders Are Judging You on Much More Than Finances*, L.A. TIMES (Dec. 19, 2015, 10:00 AM), http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html [https://perma .cc/UP4X-XQ9P]. ZestFinance uses thousands of pieces of consumer data to predict the likelihood that a borrower will repay their debts. *Id.*

Some scholars have cautioned that lenders may even deliberately target certain racial or ethnic groups by using "facially-neutral proxy variables in its scoring model as stand-ins for characteristics like race."[314] Thus, scholars like Mikella Hurley and Julius Adebayo propose that credit scores "not treat as significant any data points or combinations of data points that are highly correlated to immutable characteristics."[315]

A formalistic excluding-inputs algorithm may prohibit credit-scoring companies from using race and correlates of race from algorithms, or those deemed "highy correlated." But again, we note that this is likely to be impractical given that many, if not all, inputs are highly correlated with race. These nonrace inputs are also likely to have substantial predictive power, even independent of their correlation within race.[316] Indeed, as shown in a lending simulation by Gillis and Spiess, "if there are other variables that are correlated with race, then predictions may strongly vary by race even when race is excluded, and disparities may persist" such that "[t]o the extent that disparate impact plays a social role beyond acting as a proxy for disparate treatment, we may not find it sufficient to formally exclude race from the data considered."[317]

In contrast, our two proposed statistical solutions could reduce racial disparities in credit scoring relative to commonly used algorithms while preserving predictive power. But for our proposals to work, policymakers must shed their naive understanding of statistics, as some regulations (like Regulation B of the ECOA) prohibit creditors from even requesting or collecting information such as race. If this information cannot be collected and thus used in the first estimation step as required under our proposals, there is no way of truly eliminating racial proxy effects.

Employment and Hiring: In the employment context, Title VII of the Civil Rights Act of 1964 prohibits discrimination on the basis of race, color, religion, sex, and national origin.[318] Broadly speaking, there are two types of Title VII claims. Under the first theory, known as "disparate treatment" discrimination, an intentional policy of discriminating against a protected

---

314. Hurley & Adebayo, *supra* note 311, at 191.

315. *Id.* at 206.

316. For example, a 2007 study by the Federal Trade Commission found that credit information is highly predictive of risk even *within* racial groups, suggesting that credit information is not solely proxying for race. FED. TRADE COMM'N, CREDIT-BASED INSURANCE SCORES: IMPACTS ON CONSUMERS OF AUTOMOBILE INSURANCE 23 (2007), https://www.ftc.gov /sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-auto mobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance _scores.pdf [https://perma.cc/S355-GBRN] ("Credit-based insurance scores appear to have little effect as a 'proxy' for membership in racial and ethnic groups in decisions related to insurance.").

317. Gillis & Spiess, *supra* note 25, at 469, 471.

318. 42 U.S.C. § 2000e-2.

group is prohibited.[319] Under the second theory, known as the "disparate impact" doctrine, a facially neutral policy that nonetheless leads to an adverse impact on a protected class is prohibited unless the employer can offer a sufficient explanation that the practice in question is job related and consistent with business necessity.[320] Even if the employer meets that burden, plaintiffs can still win if they can demonstrate the existence of an available alternative employment practice that has less disparate impact and serves the employer's legitimate needs.[321]

Although many legal scholars have questioned whether Title VII is sufficient for dealing with the types of issues introduced by the use of algorithms,[322] a similar debate arises surrounding the direct and proxy effects of protected characteristics like race. Pauline Kim argues that a "formalist reading of Title VII might appear to prohibit any use of variables capturing sensitive characteristics in a data model. Certainly, a simple model that relied on race or other protected characteristics as the basis for adverse decisions would run afoul of Title VII's prohibitions."[323] Similarly, Solon Barocas and Andrew Selbst have stated with respect to algorithms in the employment context that "considering membership in a protected class as a potential proxy is a legal classificatory harm in itself" and that "[u]nder formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds."[324]

But as we noted above and as some of these scholars acknowledge, excluding these variables is problematic due to the existence of proxy effects that stem from other inputs.[325] As Kim has noted:

> [R]estricting access to sensitive information is not likely to be effective in preventing classification bias that results from data analytic models. If the data being mined is rich enough, other seemingly neutral factors may closely correlate with a protected characteristic, permitting a model to effectively sort along the lines of race or another protected characteristic. Factors such as where someone went to school or where they currently live may be highly correlated with race. Behavioral data, such as an individual's

---

319.   *See* Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324 (1977).

320.   *See* Griggs v. Duke Power Co., 401 U.S. 424, 430–31 (1971) (holding that the requirement that applicants have a high school diploma or a passing score on a written test is forbidden unless it has "a demonstrable relationship to successful performance"). Twenty years after *Griggs*, the Civil Rights Act of 1991 was enacted, which included a provision codifying the prohibition on disparate-impact discrimination. Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071 (1991) (codified as amended at 42 U.S.C. § 1981).

321.   42 U.S.C. § 2000e-2(k)(1)(A)(ii).

322.   *See, e.g.*, Barocas & Selbst, *supra* note 8, at 694 (concluding that Title VII is "not well equipped" to address data mining).

323.   Kim, *supra* note 21, at 917–18.

324.   Barocas & Selbst, *supra* note 8, at 695, 719.

325.   Kim, *supra* note 21, at 918.

Facebook 'likes,' can also predict sensitive characteristics like race and sex with a high degree of accuracy. Because other information contained in large datasets can serve as a proxy for race, disability, or other protected statuses, simply eliminating data on those characteristics cannot prevent models that are biased along these dimensions.[326]

As a result, she argues that "a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies."[327] Similarly, Barocas and Selbst note that when there is correlation between a protected characteristic and other traits, "data mining . . . can indirectly determine individuals' membership in protected classes and unduly discount, penalize, or exclude such people accordingly."[328]

Once again, however, our two statistical proposals could be used in employment algorithms. Rather than forbidding the use of protected characteristics and their correlates under a formalistic interpretation of antidiscrimination law, our solutions would allow employers to retain some predictive power while eliminating direct and proxy effects from predictions.[329] Our proposals therefore have the potential to reduce race or gender disparities in employment and hiring.

CONCLUSION

In this Article, we provide a new statistical and legal framework to understand the legality and fairness of using protected characteristics in predictive algorithms under the Equal Protection Clause. We challenge the mainstream legal position that the use of a protected characteristic always violates the Equal Protection Clause. We are also highly skeptical of the current legal push toward adopting a formalistic view of algorithms that requires the exclusion of race and all nonrace correlates in predictive algorithms, as nearly all potential algorithmic inputs are likely to be correlated with race. Our skepticism is supported by our empirical tests using information from the New York City pretrial system, where we find that all commonly used algorithmic inputs are correlated with race in our data. These results suggest that the formalistic legal position of excluding race and all race correlates from predictive algorithms is impractical, and may actually undermine the goals of equal protection if implemented incorrectly.

---

326.   *Id.* at 898.

327.   *Id.* at 918.

328.   Barocas & Selbst, *supra* note 8, at 692.

329.   For a strong legal defense of these ideas applied to Title VII, see Harned & Wallach, *supra* note 72.

Our Article offers two more practical solutions to eliminate unwarranted racial disparities in predictive algorithms, both grounded in the underlying statistical properties of algorithms and the practical reality that most, if not all, potential inputs are correlated with race. We argue that our proposed algorithms are fully consistent with the principles of the equal protection doctrine because they ensure that individuals are not treated differently on the basis of membership in a protected class, in stark contrast to commonly used algorithms that unfairly disadvantage Black individuals relative to white individuals despite the exclusion of race. We also demonstrate that our proposed algorithms could have large consequences for the racial composition of detained defendants. In empirical tests from the New York City pretrial system, our algorithms substantially reduce the number of Black defendants detained compared to commonly used algorithms.

Our findings require a fundamental rethinking of the equal protection doctrine as applied to predictive algorithms. To fully ensure that individuals are not treated differently solely on the basis of membership in a protected class, the equal protection doctrine must shed its overly formalistic interpretation of equal treatment that requires predictive algorithms to be blinded to race through exclusion. The equal protection doctrine must instead embrace the statistical reality that virtually all algorithmic inputs are correlated with race and allow for new statistical approaches that can truly ensure that all individuals are treated equally under the law.

APPENDIX

TABLE A1: SIMULATIONS OF RACIAL DISPARITIES – PREDICTING REARREST FOR
NEW OFFENSE

	Share of Black Defendants Among Released					
	Benchmark Model	Excluding Race	Colorblinding Inputs	Black as White	Difference (2) - (3)	Difference (2) - (4)
	(1)	(2)	(3)	(4)	(5)	(6)
50% Release Rate	58.55	65.95	66.83	68.00	-0.88	-2.05
70% Release Rate	67.26	70.16	70.82	71.51	-0.66	-1.35
90% Release Rate	73.71	74.23	74.34	74.49	-0.10	-0.26

Note: This table reports the percent of released defendants who are Black versus white under different prediction models and release rates using information from the New York City pretrial system. The outcome variable is whether a defendant is arrested for a new crime prior to case disposition. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. Column 1 reports the percent Black released among released defendants under the benchmark statistical model. Column 2 reports the percent Black released among released defendants under the commonly used model. Column 3 reports the percent Black released among released defendants under the colorblinding-inputs model. Column 4 reports the percent Black released among released defendants under the Black-as-white model. Column 5 reports the difference in the percent Black released defendants between the commonly used and the colorblinding-inputs model. Column 6 reports the difference in the percent Black released defendants between the commonly used and Black-as-white model. See the text for additional details on the specification and sample.

FIGURE A1: RACIAL DISPARITIES – PREDICTING REARREST FOR NEW OFFENSE
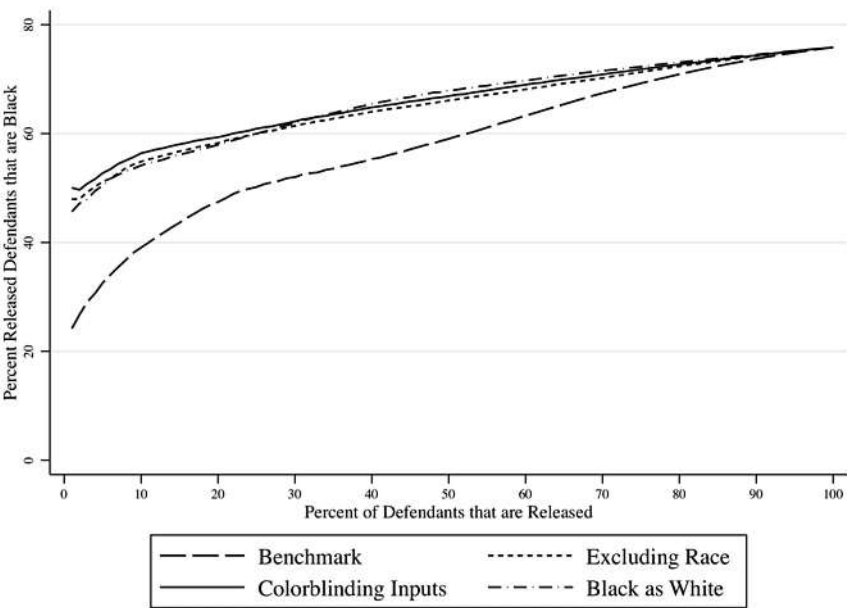
Note: This figure plots the percent of released defendants who are Black under different predictive algorithms and release rates using information from New York City pretrial system. The outcome variable is whether a defendant is arrested for a new crime prior to case disposition. The sample consists of male Black and white defendants who were arrested and charged between 11/2008 and 11/2013, whose cases were not adjudicated at arraignment, and who were released before trial. See the text for additional details on the specification and sample.