**Thursday, February 27, 2020**
**4:15 p.m.- 5:45 p.m.**
**Stanford Law School**
**Room 301**

# "Equal Protection Under Algorithms: A New Statistical and Legal Framework"

**by**

**Crystal S. Yang**

**(Harvard Law School)**

**Note: It is expected that you will have reviewed the speaker's paper before the seminar. Because it is longer than average, the speaker was asked to prepare a guide for readers who can't look at the whole paper. This is her response: "Thanks for taking the time to engage with my paper. Given its length, please feel free to focus on the abstract, Introduction and Section II (pp. 1-19), and Sections IV-V (pp. 26-38). I look forward to your feedback."**

# Equal Protection Under Algorithms: A New Statistical and Legal Framework[*]

Crystal S. Yang[†] Will Dobbie[‡]

February 2020

**Abstract**

*In this paper, 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 non-race 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 non-race 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 non-race correlates such as employment and education.*

*In the second part of the paper, we challenge the mainstream legal position that the use of a protected characteristic always violates the Equal Protection Clause. We first 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 non-minorities. While an overly formalistic legal solution requires exclusion of race and all potential non-race 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 blacks relative to whites. Both of our proposed algorithms substantially reduce the number of black defendants detained compared to these 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.*

[*] We thank Alma Cohen, Howell Jackson, Louis Kaplow, Steven Shavell, and numerous seminar participants for helpful comments and suggestions. Victoria Angelova, Claire Lazar, and Ashley Litwin provided excellent research assistance. The main data we analyze are provided by the New York State Division of Criminal Justice Services (DCJS), and the Office of Court Administration (OCA). The opinions, findings, and conclusions expressed in this publication are those of the authors and not those of DCJS. Neither New York State nor DCJS assumes liability for its contents or use thereof.

[†] Harvard Law School and NBER. Email: cyang@law.harvard.edu

[‡] Harvard Kennedy School and NBER. Email: will_dobbie@hks.harvard.edu

## Table of Contents

I Introduction 1

II Predictive Algorithms and the Equal Protection Clause 6
A Direct Effects of Protected Characteristics 6
B Proxy Effects of Protected Characteristics 13
C Trade-Off Between Fairness and Accuracy 17

III Predictive Algorithms in the Criminal Justice System 19
A Survey of Predictive Algorithms in the Criminal Justice System 19
B Algorithmic Inputs and the Direct and Proxy Effects of Race 24

IV A Statistical Framework for Predictive Algorithms 26
A Categorizing Algorithmic Inputs 26
B Benchmark Statistical Model 27
C The Direct and Proxy Effects of Algorithmic Inputs 27

V Legal and Statistical Solutions to Ensuring Race Neutrality 32
A Current Legal Solution: The Excluding-Inputs Algorithm 32
B Our First Solution: The Colorblinding-Inputs Algorithm 33
C Our Second Statistical Solution: The Minorities-as-Whites Algorithm 35
D Legality of our Two Statistical Solutions 36
E Racial Gaps Under our Two Statistical Solutions 38

VI Empirical Tests of Our Proposed Statistical Solutions 39
A The New York City Pretrial System 39
B Data Description 43
C Proxy Effects in Commonly-Used Algorithms 44
D Comparison of Different Predictive Algorithms 50

VII Extensions 58
A Additional Protected Characteristics 59
B More Complicated Algorithms 59
C Other Contexts 61

VIII Conclusion 64

A Appendix 66

# I. Introduction

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, predictive 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." 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.[1] Credit scores are also widely used in non-lending decisions, such as rental decisions for apartments.[2] 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.[3] In the criminal justice system, the focus of our paper, predictive algorithms use characteristics such as an individual's criminal history and age to predict the risk of future criminal behavior, with these "risk assessments" used to inform pretrial release conditions, sentencing decisions, and the dispatch of police patrols.[4]

The increasing use of predictive algorithms has contributed to an active debate on whether commonly-used predictive algorithms intentionally or unintentionally discriminate against certain groups, in particular 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. Yet, critics argue that the algorithmic inputs are themselves biased, resulting in violations of the Equal Protection doctrine and anti-discrimination law.[5]

[1] *See* Rob Berger, A Rare Glimpse Inside the FICO Credit Score Formula, DOUGHROLLER (Apr. 30, 2012), http://www.doughroller.net/credit/a-rare-glimpse-inside-the-fico-credit-score-formula. More recent examples include ZestFinance, which uses public credit report data, but also proprietary and social network data to predict the likelihood that a borrower will repay their debts. *See* http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html.

[2] *See* 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.

[3] *See, e.g.,* George Anders, *Who Should You Hire? LinkedIn Says: Try Our Algorithm*, FORBES (Apr. 10, 2013), http://www.forbes.com/sites/georgeanders/2013/04/10/whoshould-you-hire-linkedin-says-try-our-algorithm; Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, N.Y. TIMES: THE UPSHOT (June 25, 2015), http://www.nytimes.com/2015/06/26/upshot/can-an-algorithmhire-better-than-a-human.html; Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.

[4] *See, e.g.,* N.Y. TIMES (June 13, 2017), *Inside the Algorithm That Tries to Predict Gun Violence in Chicago*, https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html; Ellora Thadaney Israni, *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.

[5] 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. 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 "under 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). These authors have also noted that even seemingly neutral traits can, because of their correlation with protected characteristics, end up "indirectly determin[ing] individuals' membership in protected classes and unduly discount, penalize, or exclude such people accordingly." Barocas & Selbst at 692.

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 discrimi-

The debate on whether commonly-used predictive algorithms discriminate against minorities has been particularly heated in the criminal justice system, where risk assessment tools are increasingly utilized.[6] 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 demographic 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 who already disproportionately bear its brunt, including people of color."[7] 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.

These concerns are echoed in statements made by prominent public officials, including former Attorney General Eric Holder, who argue that "by 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."[8] Even commonly used algorithmic inputs such as current charge and prior criminal

nation using zip codes as proxies for the racial composition of neighborhoods. *See* 15 U.S.C. §1691(a)(1) (2012). 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 (*see* 12 C.F.R. §202.5 (2013), 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*, 12 Journal of Economic Perspectives 41, 43 (1998). Most recently, 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. *See* Andrew D. Selbst, *A New HUD Rule Would Effectively Encourage Discrimination by Algorithm*, Slate (Aug. 19, 2019), https://slate.com/technology/2019/08/hud-disparate-impact-discrimination-algorithm.html.

[6] 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* American Bar Association, Criminal Justice Section, State Policy Implementation Project at 18, available at https://www.americanbar.org/content/dam/aba/administrative/criminal_justice/spip_handouts.authcheckdam.pdf. 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 Courts, Conference of Chief Justices and Conference of State Court Adm'rs, Resolution 7: In Support of the Guiding Principles on Using Risk and Needs Assessment Information in the Sentencing Process (Aug. 3, 2011), http://www.ncsc.org/~/media/Microsites/FILES/CSI/Resolution-7.ashx. 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) (2016) (sentencing judges in Kentucky 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) (2015-16) (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. Cons. Stat. §2154.7(a) (2016) (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. tit. 22, §988.18(B) (2016) (an assessment and evaluation instrument designed to predict risk to recidivate is required to determine eligibility for any community punishment), with many other states permitting the use of such algorithmic tools, *See, e.g.,* Idaho Code §19-2517 (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).

[7] Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 Stan. L. Rev. 803, 806 (2014).

[8] *See* Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meet-

history, which many argue are both relevant and legally permissible,[9] 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.[10]

In this paper, 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 paper proceeds in three main steps. In the first part of the paper, 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 anti-classification 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 "objective" inputs can still result in unwarranted disparities if those inputs act as proxies for race. For example, zip code is highly correlated with race in many datasets, leading some to argue that using zip code as an algorithmic input is therefore "tantamount to using race." As a result, numerous legal scholars and policymakers have urged jurisdictions using predictive algorithms to exclude race and factors correlated with race as inputs.[11] As noted by some scholars,

ing and 13th State Criminal Justice Network Conference, available at https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th. 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 (Sep. 25, 2017), https://billypenn.com/2017/09/25/can-phillys-new-technology-predict-recidivism-without-being-racist/.

[9] *See, e.g.,* Attorney General Holder Remarks, *supra* note 8 ("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.").

[10] *See, e.g.,* Stephen Goldsmith and Chris Bousquet, *The Right Way to Regulate Algorithms*, City Lab (Mar. 20, 2018) ("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."), https://www.citylab.com/equity/2018/03/the-right-way-to-regulate-algorithms/555998/; *see also* Beth Schwartzapfel, *Can Racist Algorithms Be Fixed?*, The Marshall Project (July 1, 2019) ("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 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.), https://www.themarshallproject.org/2019/07/01/can-racist-algorithms-be-fixed.

[11] *See, e.g.* Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L.J. forthcoming at 3 ("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.").

"antidiscrimination regimes are generally operationalized ... simply by excluding from the data available ... any information on membership in legally suspect classes or, in some cases, the most obvious proxies for such group membership.'[12]

We then review the most common predictive algorithms in the criminal justice system and their inputs. Surveying the field, we find that all commonly-used predictive algorithms exclude race as an input. The 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 non-racial algorithmic inputs that may be correlated with race. At least some commonly-used predictive algorithms purposely exclude non-race 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 all possible non-race inputs.

In the second part of the paper, we develop a statistical framework that formalizes the mainstream legal position that the use of both race and non-race correlates is problematic and potentially unconstitutional under the Equal Protection Clause. Consistent with 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 non-race 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 non-race correlates can generate unwarranted racial disparities.

Building on this statistical framework, we discuss three potential solutions that can eliminate the direct and proxy effects of race in predictive algorithms. The first potential 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 are able to achieve full race-neutrality.

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, but without requiring that race and any race-correlates be excluded. As a result, both solutions are implementable even if all non-race inputs are racial proxies. Importantly, both algorithms achieve race-neutrality by considering or using race in the estimation step, running counter to the intuitive but statistically incorrect and overly formalistic anti-classification principle that the use of race in any form would violate the Equal Protection Clause.[13] 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

[12] Anya Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, IOWA L. Rev. forthcoming, at 6-7.

[13] *See, e.g.*, Starr, *supra* note X, at 870 (concluding that "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 X, at 14 ("Colorblindness would simply prohibit the use of race as an input variable for prediction. Colorblindness would also prohibit the intentional use of race proxies.")

the primary principles underlying the Equal Protection doctrine. Our algorithms are consistent with the anti-classification principle, as they use race solely to 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 anti-subordination principle, as they use race precisely to avoid inflicting harm on disadvantaged groups.[14]

In the final part of the paper, 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 disparities, undermining the objective of equal treatment.[15] 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 they used our proposed statistical models instead of the more commonly-used predictive algorithms.

Our paper links two important literatures: a legal literature on the constitutionality of predictive algorithms under anti-discrimination law,[16] and a social science literature on algorithmic fairness.[17] In our reading, the legal literature has adopted an overly formalistic interpretation of the principles of equal treatment,

[14] This anti-subordination principle is most closely linked to Owen Fiss' article *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107 (1976). 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 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*, 17 Issues in Legal Scholarship 1, 1 (2003).

[15] 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, in a forthcoming paper, Aziz Huq notes that "to the extent that race is thought to be already highly correlated with socioeconomic characteristics related to criminogenic and victimization distributions, it might 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 will either to close to, or even a good approximation of, racial distributions in the population." Aziz Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L. J. (forthcoming).

[16] *See, e.g.,* Starr, *supra* note X, at X; Dawinder Sidhu, *Moneyball Sentencing*, 56 B.C. L. Rev. 671, 694 (2015); *see also* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 698 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm. & Mary L. Rev. 857, 904 (2017).

[17] *See, e.g.,* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AEJ: Policy 206 (2011); Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, & Ashesh Rambachan, *Advances in Big Data Research in Economics: Algorithmic Fairness*, 108 AEA Papers and Proceedings 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."); Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, & Richard Zemel, *Fairness Through Awareness*, ITCS (2011); Toon Calders & Indre Zliobaite, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, in Discrimination and Privacy in the Information Society at 12 (2013); Moritz Hardt, Eric Price, & Nathan Srebro, *Equality of Opportunity in Supervised Learning* (2016).

leading to the misguided conclusion that the use of 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,[18] but has largely ignored the implications of such use under the law.[19]

The contribution of our paper is to challenge the mainstream legal position that the use of a protected characteristic always violates the Equal Protection Clause, 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, one that embraces 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.

## II. Predictive Algorithms and the Equal Protection Clause

In this section, 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 non-protected characteristics can essentially "proxy" for these protected characteristics because of their correlation with those characteristics. We 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 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.[20] This positive correlation between race and pretrial misconduct means that predictive algorithms will assign a higher risk

[18] *See, e.g.,* Hardt et al., *supra* note X, 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* Indre Zliobaite, Faisal Kamiran, & Toon Calders, *Handling Conditional Discrimination*, in 11th IEEE International Conference on Data Mining at 1 (2011) ("discrimination may occur even if the sensitive information is not directly used in the model").

[19] 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 anti-discrimination law in the context of lending).

[20] *See infra* Section VI.

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.[21] 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 that classification, whether to burden or benefit such groups, violate the Constitution's "immunity from inequality of legal protection."[22] While not a blanket ban on the use of racial classifications, the Equal Protection Clause does subject judicial review of such classifications to strict scrutiny.[23] Under strict scrutiny, a policy with a racial classification must represent a compelling government interest and must be narrowly tailored to achieve that interest.[24] 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."[25] 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.[26]

[21] A second, less discussed, theory for which the use of race and gender in risk assessments may be constitutionally problematic is if the government policy in question is "motivated by a racially discriminatory purpose," most often applied with respect to facially neutral laws. *Washington v. Davis*, 426 U.S. 229, 240 (1976) (stating the "basic equal protection principle 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."). However, 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." *Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252, 264-65 (1977). 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, 281-82 (1987).

As legal scholars have noted, this strand of the Equal Protection doctrine would likely be a poor basis for any challenge of a risk assessment instrument because it would be difficult to show that an algorithm was specifically designed with a racially discriminatory motive. *See* Huq, *supra* note X, at 30 (citing *Personnel Admr. v. Feeney*, 442 U.S. 256, 279 (1979) (noting that"[w]ithout 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 have been included, as the doctrine requires."); *see also* Sidhu, *supra* note X, 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 reasons. Rather, these instruments are clearly used to control crime. As a result, facially neutral risk-assessments would likely survive a constitutional attack" (citations omitted)). Similar arguments have been made in the context of predictive algorithms and Title VII. *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 698 (2016) ("Except for masking, discriminatory data mining is by stipulation unintentional.").

[22] *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.").

[23] *See Parents Involved in Cmty. Sch. 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").

[24] *See, e.g., Adarand Constructors v. Pena*, 515 U.S. 200, 234 (1995) ("Federal racial classifications, like those of a State, must serve a compelling government interest, and must be narrowly tailored to further that interest.").

[25] *Richmond v. J. A. Croson Co.*, 488 U.S. 469, 493 (1989) (plurality opinion).

[26] Strict scrutiny is not "strict in theory, but fatal in fact." *Adarand Constructors, Inc. v. Pena*, *supra* note X, at 237. For example, one of the earliest examples was a federal racial classification underlying a curfew applicable only to persons of Japanese

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,"[27] 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 individual talents and capacities."[28] 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.'"[29] While there are numerous examples of gender-based classifications that have been invalidated, some have been upheld.[30]

To date, there is no legal precedent on how these anti-classification principles have been 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.[31] 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 demographic and socioeconomic status."[32] Starr specifically argues that using demographic and socioeconomic characteris-

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 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" in *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 (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." *Id.* 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 Index" (PAI). 136 S. Ct. 2198, 2205-07 (2016).

[27] *United States v. Virginia*, 518 U.S. 515, 531 (1996) (internal quotation marks omitted) (holding that the exclusively male admissions policy of the Virginia Military Institute (VMI) at the time violated the Equal Protection clause).

[28] *Id.* at 532 (citing *Kirchberg v. Feenstra*, 450 U.S. 455, 462-463 (1981); *Stanton v. Stanton*, 421 U.S. 7 (1975).

[29] *Id.* at 533 (alteration in original) (citations omitted). For other cases that applied intermediate scrutiny to gender classifications, *see Miss. Univ. for Women v. Hogan*, 458 U.S. 718 (1982); *Craig v. Boren*, 429 U.S. 190 (1976).

[30] 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 old-age insurance benefits for female wage earners relative to otherwise similarly situated male wage earners. 430 U.S. 313, 317 (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.")

[31] Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 Stan. L. Rev. 803, 806 (2014).

[32] Starr, *supra* note X, at 806. Aziz Huq calls this assertion a "dubious proposition" and "not a fair reading of current law." Huq, *supra* note X, at 7-8. Richard Primus has also noted, "many practices that do involve government actors' identifying people by race

tics 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."[33] One of Starr's main concerns is 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 an algorithmic input concern race/ethnicity.[34] For example, Starr claims that there "appears to be a general consensus that using race would be unconstitutional."[35] 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 anti-classification

are not always subject to strict scrutiny." Richard Primus, *Equal Protection and Disparate Impact*, 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).

[33] *Id.*; *see also* Sonja B. Starr, *The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong*, 27 Federal Sentencing Reporter 229, 230 (2015) ("When the government instructs judges to consider risk scores based on factors 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.").

[34] 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. Similarly, in a recent article, 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 State Journal of Criminal Law 583, 592 (2018). For example, John Monahan has argued with respect to gender, "that 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 Virginia 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 difficulty 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 prevention." Monahan, *supra* note X, 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 X, at 823-29. Starr specifically questions the notion that "actuarial fairness" or relatedly statistical discrimination, are permissible under the Constitution. *Id.* at 825-26 (citing to 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 also* Sidhu, *supra* note X, 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.*, Sidhu, *supra* note X, at 700-701 ("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 risk assessment tools is unconstitutional because it is equivalent to "punishing a person for his poverty." Starr, *supra note X*, at X; Bearden v. Georgia, 461 U.S. 660, 666-67 (1983) (quoting Williams v. Illinois, 399 U.S. 235, 260 (1970) (Harlan, J., concurring in the result)).

[35] Starr, *supra* note X, at 812.

cases "should put to rest any suggestion that these traits [referring to race and religion] are constitutionally appropriate in risk-assessment."[36] Sharing these views, Christopher Slogobin raises similar equal protection issues with risk assessment in the juvenile context. As he notes, "use of race, ethnicity...as risk factors should require a compelling justification, rather than merely a rational one...but because [these] factors are considered highly suspect classifications, traditional Fourteenth Amendment analysis would also require the government to show that their use as risk factors is crucial to the achievement of that objective. Such a showing is unlikely, given the less-than-robust-correlation between these characteristics and risk, as well as the larger number of other risk factors available to the government. In any event, most courts have accepted the proposition that race may not be considered in determining dangerousness."[37] An important point is that 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 decision-makers on the basis of race.[38] Numerous courts and sentencing commissions have, for example, proclaimed that a "defendant's race or nationality may play no adverse role in the administration of justice, including at sentencing."[39]

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."[40] This provision embodies Congress' position that it is inappropriate "to afford preferential treatment to defendants of a particular race...."[41] However, such a provision was made with respect to decisions made by human judgment alone and is related to concerns about unwarranted sentencing disparities,[42] 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

[36] Sidhu, *supra* note X, at 699.

[37] Christopher Slobogin, *Risk Assessment and Risk Management in Juvenile Justice*, Crim. Just., Winter 2013, at 13-14.

[38] *See generally* Carissa Byrne Hessick, *Race and Gender as Explicit Sentencing Factors*, 14 J. Gender Race & Just. 127 (2010) for an in-depth history of the use of race and gender in sentencing. For example in *United States v. Kaba*, 480 F.3d 152, 156 (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." (citing Leung, 40 F.3d at 586). 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 his national origin, and not because he trafficked in drugs that emanated from a source country.").

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

[40] *See* 28 U.S.C. §994(d) (2012).

[41] S. Rep. No. 98-225, at 171 (1983).

[42] *See* S. Rep. No. 98-225, at 38 (1983) (Senate Report on precursor to federal Sentencing Reform Act of 1984) ("[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 service in prison.").

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."[43]

The second notable example is the American Law Institute's (ALI) 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 professional training, experience, and judgment of the persons making predictions."[44]

However, according to the reporter's note in the March 2011 draft of Model Penal Code on Sentencing, "the consideration of race and ethnicity is disapproved...and raises serious constitutional concerns...."[45]

With that said, 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*.[46] Applying strict scrutiny, Oleson argues that protecting the public from crime is a compelling state interest,[47] and that inclusion of race in predicting risk is narrowly tailored given studies showing that race is highly correlated with recidivism.[48] 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.[49] As he argues, "race 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)."[50] Similarly, Judge Richard Kopf has argued that "a sentencing system based upon a robust actuarial data set consisting of *all* factors [including age, race, or gender] statistically correlated with risk would arguably pass constitutional muster, even under strict scrutiny."[51] 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.

[43] Dawinder Sidhu, *Moneyball Sentencing*, 56 B.C. L. Rev. 671, 694 (2015).

[44] Model Penal Code: Sentencing §6B.09 cmt. a at 53, 55 (Tentative Draft No. 2, 2011).

[45] Model Penal Code: Sentencing §6B.09 reporter's note.

[46] *See* J.C. Oleson, *Risk in Sentencing: Constitutionally Suspect Variables and Evidence-Based Sentencing*, 64 SMU L. Rev. 1329, 1385-86, 1377.

[47] *Id.* at 1385.

[48] *Id.* at 1350 (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).

[49] *Id.* at 1337, citing Joan Petersilia & Susan Turner, *Guideline-Based Justice: Prediction and Racial Minorities*, 9 CRIME & JUST. 151, 174 (1987) (noting that omitting race-correlated factors reduces accuracy of recidivism prediction by five to twelve percentage points).

[50] *Id.* at 1386.

[51] Judge 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 Federal Sentencing Reporter 207, 213 (2015).

*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 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 an input ..., its use in sentencing criminal defendants would almost certainly constitute an Equal Protection violation."[52]

We also view it as highly likely that courts considering this issue in the years to come may reject the use of protected characteristics, in particular, race and perhaps even 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."[53] Citing the concerns raised by scholars like Starr and Sidhu, 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."[54] Case law also suggests that even if statistical differences between black or white individuals, or male and female individuals, are "actuarially fair," a court would still likely apply heightened scrutiny in assessing the permissibility of using such traits.[55]

[52] Kehl, Danielle, Guo, Priscilla, Kessler, Samuel. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing at 24 (July 2017). Responsive Communities, available at https://cyber.harvard.edu/publications/2017/07/Algorithms.

[53] Brief at 19.

[54] Brief at 19 (although arguing that the petition for a writ of certiorari should be denied in this case).

[55] 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. *See, e.g.,* Phelps 1972; Arrow 1973. While the Supreme Court has never explicitly addressed the constitutionality of statistical discrimination on the basis of race, it has suggested that strict scrutiny would likely apply to most policies that rely on this type of rationale. *See* Huq, *supra* note X, at 27-28 (stating "the Court has not been clear on whether such statistical discrimination triggers constitutional concerns....All that can safely be said is that at 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 with any clarity."); Strauss, *supra* note X, 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"); *Cf.* Starr, *supra* note X, at 827 (claiming that the "Supreme Court has squarely rejected statistical discrimination – that is, the use of group tendencies as a proxy for individual characteristics – as a permissible justification for otherwise constitutionally forbidden discrimination").

One of the few 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 cited that its decision was in the best interests of the child because "it is inevitable that Melanie [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 the "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 'the 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 433-35. 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. Memphis*, 373 U.S. 526, 536 (1963), suggesting that statistical evidence showing a true relationship between race and risk may yield a different conclusion. In a more recent case, the Court in *Johnson v. California* considered an unwritten California prison policy

### B. Proxy Effects of Protected Characteristics

The second legal concern on the use of protected characteristics is that seemingly "fair" algorithmic inputs such as criminal history can proxy for protected characteristics such as race. In this scenario, the use of these seemingly fair 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. 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 code in predictive algorithms is "almost tantamount to using race."[56]

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.[57] We show formally in Section 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, and thus inputs such as zip code of residence, education, and employment status should be excluded from predictive algorithms,[58] although whether any particular algorithmic input is actually correlated with race is an empirical question that may differ across contexts.[59]

*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

that racially segregation inmates for up to 60 days upon arrival. 543 U.S. 499, 502 (2005). The asserted rationale for the policy was that 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 "prisons 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. Most recently, in *Buck v. Davis*, an ineffective assistance of counsel case where the defense attorney introduced statistical evidence that the defendant was statistically more likely to act violently because he is black, the Court stated that "it 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).

[56] Cathy O'Neil, *The Ethical Data Scientist*, SLATE (Feb. 4, 2016), https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html.

[57] 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. *See, e.g.,* Kim, *supra* note X, at 904. Computer scientists have also highlighted the importance of proxy effects, labeling this problem "redundant codings," 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.,* Cynthia Dwork et al., *Fairness Through Awareness*, 3 Proc. Innovations Theoretical Computer Sci. Conf. 214, app. at 226 (2012). Economists have also noted the potential importance of proxy effects in predictive algorithms, in particular how such proxy effects could generate unwarranted disparities. *See* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AEJ: Policy 206 (2011).

[58] *See, e.g.,* Starr, *supra* note X, at 838 ("socioeconomic 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.").

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

correlated with race will "exacerbate the unacceptable racial disparities in our criminal justice system."[60] For instance, Larry Krasner, the current District Attorney in Philadelphia, has argued that "there is a real danger that the components going into the risk assessment are proxies for race and for socioeconomic status."[61] 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. Despite the fact that current charge and prior criminal history are routinely used,[62] 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.[63] For example, prior arrests may not just reflect 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.[64]

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.[65] 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....strict scrutiny is only triggered if the individuals challenging the law can show that it was also adopted with a racially discriminatory intent. If not, rational basis review applies, a highly deferential standard."[66]

Aside from potential constitutional constraints, many use 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 ill-defined. For example, Cathy O'Neil, author of *Weapons of Math Destruction*,

[60] Bernard Harcourt, *Risk as a Proxy for Race*, 27 Federal Sentencing Report 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." *Id.*

[61] Anna Orso, *Can Philly's new technology predict recidivism without being racist?*, BillyPenn (Sep. 25, 2017), https://billypenn.com/2017/09/25/can-phillys-new-technology-predict-recidivism-without-being-racist/.

[62] *See, e.g.,* Attorney General Holder Remarks, *supra* note X ("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.").

[63] Bernard E. Harcourt, Risk as a Proxy for Race, CRIMINOLOGY & PUB. POL'Y, (forthcoming) (manuscript at 2), available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1677654; *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).

[64] *See, e.g.,* Stephen Goldsmith and Chris Bousquet, *The Right Way to Regulate Algorithms*, City Lab (Mar. 20, 2018) ("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."), https://www.citylab.com/equity/2018/03/the-right-way-to-regulate-algorithms/555998/.

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

[66] Berkman Klein Report, *supra* note X, at 24 (citing *Personnel Administrator v. Feeney*, 442 U.S. 256 (1979) (holding that in order to find discriminatory intent, a state legislature has to have acted "because of," not "in spite of," the effects of a statute in relatively disadvantaging members of a particular minority group).

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."[67]

One possible dividing line is that correlated inputs should be excluded if the reason for the correlation is because of past discrimination or racial animus.[68] 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:

> "Difficult problems are also 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, before the algorithm was asked to do its predictive work. There is a risk here that algorithms might perpetuate discrimination, and extend its reach, by using factors that are genuinely predictive, but that are products of unequal treatment. It might make discrimination into a kind of self-fulfilling prophecy."[69]

But commonly-used inputs in many predictive algorithms do not seem to follow this principle. For example, there appears to be a plethora of empirical evidence suggesting that lengthier prior criminal histories among black men could be due to discriminatory policing.[70] As a result, criminal history is consistently highly correlated with race,[71] yet is nearly universally embraced by legal scholars and policymakers and is almost

[67] Cathy O'Neil, *When Not to Trust the Algorithm*, Harvard Business Review (Oct. 6, 2016), available at https://hbr.org/ideacast/2016/10/when-not-to-trust-the-algorithm.html.

[68] *See, e.g.,* Prince & Schwarcz, *supra* note X, at 35 ("These points are most salient in the criminal justice system, where it is widely understood that African Americans systematically fare less well than whites due to broad economic inequalities and the innumerable prejudices that are baked into the criminal justice system. For this reason, race is often highly predictive of seemingly neutral objectives, like maximizing arrest probabilities, minimizing reported crimes, or limiting recidivism. AIs that are programmed around these objectives will consequently inevitably seek to capture this predictive power of race by relying on proxies for that characteristic. This reality, of course, ultimately validates past prejudice and discrimination.")

[69] Cass Sunstein, *Algorithms, Correcting Biases*, Social Research (2018) at 8. Similarly, Tafari Mbadiwe notes that the "root of the problem" is that "Racism may well be a significant factor in the higher arrest and conviction rates among black people to begin with. And because of this, racial proxies are a double-edged sword: They can bolster algorithmic accuracy, but only at the cost of validating and perpetuating the vicious cycle in which our justice system's propensity to disproportionately arrest and incarcerate black people fuels the disproportionate arrest and incarceration of black people. In this light, the fairness paradox is cold comfort, since it doesn't absolve us of the charge that including racial proxies amounts – in effect, if not necessarily intent – to judging people by the color of their skin." Tafari Mbadiwe, *Algorithmic Injustice*, The New Atlantis, https://www.thenewatlantis.com/publications/algorithmic-injustice.

[70] Cite empirical literature here.

[71] Bernard Harcourt, *Risk as a Proxy for Race*, *supra* note X, 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.").

always used in risk assessment instruments.[72] In fact, criminal history is often portrayed as the counterpoint to protected characteristics such as race in terms of both legal and ethical permissibility. For example, Berk and Hyatt claim that "the 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.'[73] And as summarized by Mark Moore, the consensus view appears to be that "some 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.'[74]

*Summary*: Based on our review, we see the mainstream position as discouraging the use of proxies in predictive algorithms. However, the Equal Protection doctrine has far less of a bite here than it does with respect to direct use of protected characteristics. As a result, deciding which proxies are permissible and which are not is often an ad hoc process.[75] Specifically, the arguments in favor of or against certain inputs often rely on ad hoc normative judgments of what is morally troubling and what is not.[76] At the extreme, some have suggested that all racial proxies be excluded from predictive algorithms.[77] But this position, while principled, is untenable in practice because excluding any correlated variable would likely mean that algorithms are totally unusable because every possible input is likely correlated with race. As some have

[72] *See, e.g.,* Sonja B. Starr, *The New Profiling: Why Punishing Based on Poverty and Identity Is Unconstitutional and Wrong*, 27 Federal Sentencing Reporter 229, 231 (2015) ("In contrast to gender and socioeconomic variables, some other risk factors in the instruments are constitutionally permissible considerations. These include criminal history as well as some demographic classifications, such as age, that do not trigger special constitutional scrutiny.").

[73] Berk and Hyatt, *supra* note X, at 226 (citing Carissa Byrne Hessick & F. Andrew Hessick, *Recognizing Constitutional Rights at Sentencing*, 99 Calif. L. Rev. 47-94 (2011); *Almendarez-Torres v. United States*, 523 U.S. 224 (1998).

[74] 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 et al. eds., 1986).

[75] A related debate is what non-race 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* (2010), *available* at http://islandia.law.yale.edu/ayers/ayresincludedvariablebias.pdf. Similarly, Jung et al. note that as "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 effect of discrimination on 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* (2018).

[76] For example, Slobogin claims that non-race 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, *Principles*, *supra* note X, 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 low, but the reason for the correlation is discrimination, does that mean the input should nevertheless be included? Conversely, if the correlation is high, but the reason for the correlation is warranted, why should the input be excluded?

[77] *See, e.g.,* Prince & Schwarcz, *supra* note X, at 36 ("It is completely reasonable, of course, to assume that race is only predictive of facially-neutral goals because of past discrimination or current biases."); *see also* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, Journal of Institutional and Theoretical Economics (2018) (noting that even seemingly "socially acceptable" inputs may themselves 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").

noted, "[i]f you want to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[78] We return to this question in our empirical results below.

### C. Trade-Off Between Fairness and Accuracy

We conclude this section 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 'fair algorithms' that explicitly seek to eliminate the capacity of any facially-neutral considerations to proxy for the prohibited characteristic through omitted variable bias."[79]

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,[80] choosing to exclude protected characteristics comes at the cost of predictive accuracy.[81] Removing correlated inputs that serve as proxies for protected characteristics also comes with a loss in accuracy.[82] Berk and Hyatt, for example, note 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."[83]

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."[84] In fact, the degree to which an input enhances an algorithm's accuracy may be a factor that

[78] Nadya Labi, *Misfortune Teller*, ATLANTIC, https://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846/.

[79] Prince & Schwarcz, *supra* note X, at 10.

[80] *See, e.g.,* Paul Gendreau et al., *A Meta-Analysis of the Predictors of Adult Offender Recidivism: What Works!* 34 CRIMINOLOGY 57, 576 (1996).

[81] *See, e.g.,* Pati 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 ("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 and Rhonda Cupp, *Gender and Risk Assessment: The Empirical Status of the LSI-R for Women*, 23 Journal of Contemporary Criminal Justice 363 (2007) (arguing for separate risk assessment instruments for men and women given different pathways to crime for men and women).

[82] *See, e.g.,* Toon Calders & Indre Zliobaite, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, in Discrimination and Privacy in the Information Society at 12 (2013) ("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 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."))

[83] Richard Berk and Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 Federal Sentencing Reporter 222, 227 (2015).

[84] Oleson, *supra* note X, at 1352.

is considered in a legal inquiry under the Equal Protection Clause.[85] 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.[86] As a result, some legal scholars have argued that exclusion of race and racial proxies would

[85] In a string of recent 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 and race, 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.

Take, for example, two recent state cases, *Malenchik v. State* and *State v. Loomis*. In *Malenchik v. State*, a 2010 case decided by the Supreme Court of Indiana, the defendant was sentenced to six years in prison (two years suspended) after pleading guilty to receiving stolen property and admitting to being a habitual offender. *Malenchik v. State*, 928 N.E.2d 564, 566 (Ind. 2010). Prior to sentencing, the county probation department prepared a pre-sentence investigation report. As 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." The scores from both LSI-R and SASSI were referenced two times by the judge at sentencing, who noted, among other things, that "[Y]our LSIR score is high. Your SASSI score is high with a high probability of substance dependence disorder." 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' 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. *Id.* at 574. However, the court rejected this argument, noting that Indiana's law required such factors to be included in the pre-sentence 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.* at 574. 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, Ph.D. & Kristin Bechtel, M.S., *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*, 71 Fed. Probation 25, 27-29 (Dec. 2007)).

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.3d. 749, 754 (2016). Prior to sentencing, a probation officer prepared a presentence investigation report (PIR), 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.

Loomis was subsequently sentenced to six years in prison and five years of extended supervision. *Id.* at 756. The defendant filed a motion for post-conviction relief requesting a new sentencing hearing. *Id.* at 756. 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. Notably, Loomis did not bring an Equal Protection claim regarding the use of gender.

As to this claim on the use of gender in the COMPAS algorithm, the court noted 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 women will misclassify both genders." *Id.* at 766. 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. *Id.* at 766-67. 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 the algorithmic risk assessment did not violate the defendant's due process rights. *Id.* at 770-71.

[86] 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 compelling interests....*If*, instead,...race or ethnicity was not a significant corre-

"compromis[e] the ability of the government to achieve its compelling interest (preventing crime)."[87] 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.

## III. Predictive Algorithms in the Criminal Justice System

In this section, 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.[88] 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. The algorithm has subsequently been adopted by over 60 police departments across the country, including by departments in Kansas, Washington, and South Carolina. 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. 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."[89]

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.[90] Using data on arrestees from Chicago, the algorithm predicts the probability that individuals

late...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 (2014). Even Starr claims that if there is a "marginal gain in predictive accuracy" from adding characteristics like race and gender, her "constitutional objections...would be alleviated."Starr, *The New Profiling*, *supra* note X, at 232 (citing a few studies that purport to show that including demographic and socioeconomic factors does not significantly increase predictive accuracy).

[87] Oleson, *supra* note X, at 1386.

[88] For a general overview of risk assessments in the criminal justice system, see Brandon L. Garrett & John Monahan, *Judging Risk*, 108 Cal. L. Rev. 101, 112 (forthcoming).

[89] *See* http://www.predpol.com/how-predictive-policing-works/.

[90] *See Inside the Algorithm That Tries to Predict Gun Violence in Chicago*, N.Y. TIMES (June 13, 2017), https://www.nytimes.com/2017/06/13/upshot/what-an-algorithm-reveals-about-life-on-chicagos-high-risk-list.html; *see also* Jessica Saunders, Priscilla Hunt & John S. Hollywood, Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot, 12 J. EXPERIMENTAL CRIMINOLOGY 347 (2016).

will be involved in a shooting and ranks individuals on a risk scale of zero to 500.[91] 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.[92] SSL explicitly excludes race and gender as algorithmic inputs.[93]

*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 Foundation, to predict the risk of pretrial misconduct. The PSA has been rapidly adopted by at least 40 jurisdictions to date, including Charlotte, Chicago, and Phoenix, 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.

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. In contrast, the PSA explicitly excludes inputs such as race, gender, education, socioeconomic status, and neighborhood of residence.[94]

In creating the PSA, Arnold Ventures wanted to create an objective and fair pretrial decision tool, which it interpreted as "meaning that they should not contain factors that would lead defendants to be treated differently because of their race, gender, or socioeconomic status. To 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."[95] Citing research that shows that race and gender are not the best predictors of pretrial risk,[96] 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."[97]

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

[91] *See id.*

[92] *See* https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np.

[93] *Id.*

[94] *See* Laura and John Arnold Foundation, Public Safety Assessment: Risk Factors and Formula, available at https://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf.

[95] Anne Milgram, Alexander M. Holsinger, Marie VanNostrand, and Matthew W. Alsdorf. *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making*, 27 Federal Sentencing Reporter 216, 220 (2015).

[96] *Id.* (citing K. Bechtel, C.T. Lowenkamp, and A.M. Holsinger, *Identifying the Predictors of Pretrial Failure: A Meta-Analysis*, 75(2) Fed. Probation 78-87 (2011); M. VanNostrand and G. Keebler, *Pretrial Risk Assessment in the Federal Court*, 73 Fed. Probation 3-29 (2009)

[97] *Id.*

Department of Criminal Justice Services in 2003. 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.[98] These factors are then converted into a risk level, which is used as an input into the Praxis decision-making tool that provides recommendations for release and detention, as well as the appropriate terms of pretrial supervision.[99] Factors like race and gender are not used as predictive inputs.

*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 1994, known as the Nonviolent Risk Assessment (NVRA). The risk assessment tool was mandated by the Virginia General Assembly, with the goal of diverting 25 percent of nonviolent offenders to alternative sanctions in lieu of incarceration by identifying low-risk individuals.[100] In 2012, the Virginia Sentencing Commission re-validated 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.[101] The Commission includes 11 factors to predict recidivism, including gender, age, marital status, employment status, current offense information, prior record, and prior juvenile incarceration.[102] The Commission also found that race was highly predictive of recidivism.[103] However, it chose to exclude race from the risk assessment because 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."[104] Interestingly, the Commission noted that by excluding race, the "procedure inevitably led to the loss of some predictive efficiency."[105]

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. While not yet enacted, the tool proposed by the PA Sentencing Commission has considered including factors such as age, gender, and the number and types of prior convictions.[106] Importantly, the PA

[98] *See* Race and Gender Neutral Pretrial Risk Assessment, Release Recommendations, and Supervision: VPRAI and Praxis Revised - Luminosity 2016, available at https://university.pretrial.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=7ebee4a7-4bde-62f5-c031-6a3df7a4bc13&forceDialog=0.

[99] *Id.*

[100] *See* Brian J. Ostrom et al., Nat'l Ctr. For State Courts, Offender Risk Assessment in Virginia 17, 9 (2002).

[101] *See* John Monahan, Anne L. Metz, & Brandon L. Garrett, *Judicial Appraisals of Risk Assessment in Sentencing*, 36 Behav Sci Law 565, 567 (2018).

[102] Ostrom, *supra* note X, at 11.

[103] *See* Brian J. Ostrom et al., Nat'l Ctr. For State Courts, Offender Risk Assessment in Virginia 17, 27-28 (2002).

[104] *Id.*

[105] *Id.* at 28, fn 10.

[106] *See* Pennsylvania Commission on Sentencing, Risk Assessment Project II, Interim Report 2, available at

Sentencing Commission purposely excluded the use of race in its risk assessment tool.[107]

*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.[108] 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. 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 attitudes.[109] 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.[110]

A second commonly-used risk assessment instrument is the Level of Service Inventory Revised (LSI-R). Developed in 1995, 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."[111] The LSI-R uses 54 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."[112] These factors then

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.

[107] *See* https://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html Interestingly, the PA 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 Commission Report, fn 8 at 12.

[108] 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, specifically that blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. *See* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Although this allegation was challenged by both Northpointe and various academics, who noted that the algorithm exhibited similar rates of recidivism among white and black offenders who received the same score (e.g. 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. *See* Corbett-Davies, E. Pierson, A. Feller, S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear," Washington Post, 17 October 2016, available at www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas; *see also* J. Kleinberg, S. Mullainathan, M. Raghavan, *Inherent trade-offs in the fair determination of risk scores*, available at https://arxiv.org/abs/1609.05807v2(2016).

[109] *See* https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html.

[110] *See, e.g.* Julia Dressel and Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 Science Advances 1, 1 (2018) ("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.").

[111] Christoper T. Lowenkamp & Edward J. Latessa, *Validating the Level of Service Inventory Revised in Ohio's Community Based Correctional Facilities* at 5, available at http://www.uc.edu/CCJR/Reports/ProjectReports/OHIOCBCFLSI-R.pdf. 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).

[112] Malenchik v. Indiana, 928 N.E.2d 564, X (Ind. 2010); *see also* Alexander M. Holsinger, Christopher T. Lowenkamp, and

generate a risk prediction of each offender's likelihood of recidivism. Gender and race/ethnicity are not included among the various risk factors.[113]

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.[114] Designed to predict the risk of future offending, the most current iteration (1991) of the SFS includes factors such as prior convictions, incarcerations, age at commencement of current offense, recent commitment free period, no parole revocation, and custody status.[115] 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 the SFS even though they weakened predictive accuracy.[116] Characteristics such as age, employment, education, residential status, and family characteristics were initially included but eventually discarded because they were deemed "heavily correlated with race," with the U.S. Parole Commission deciding that their use would be "unjust."[117]

*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.[118] 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. 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.[119] Using these inputs, MacArthur researchers placed patients into one of five risk categories using a "classification tree" methodology.[120] The MacArthur researchers explicitly excluded race from the algorithm "to avoid any possible misinterpretation of our risk assessment procedures as a form of 'racial profiling."' The researchers also note that "The revised models without race differed only trivially in accuracy from the original ones that included race."[121]

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 decision-making in the county's child welfare system. The AFST includes factors such as criminal

Edward J. Latessa, *Ethnicity, gender, and the Level of Service Inventory-Revised*, 31 J. Crim. Just. 309, 310 (2003) (describing the LSI-R).

[113] Holsinger et al. at 312-13.

[114] *See* Tonry, *supra* note X, at 168.

[115] *Id.* at 168 (Table 1).

[116] *Id.* at 172.

[117] *Id.*; *see also* Peter B. Hoffman, *Screening for Risk: A Revisited Salient Factor Score (SFS 81)*, 11 J. Crim. Justice 539 (1983).

[118] *See, e.g.,* Paul S. Appelbaum et al., *Violence and Delusions: Data from the MacArthur Violence Risk Assessment Study*, 157 Am. J. Psychiatry 566 (2000).

[119] *See* John Monahan et al., Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence 134 (2001) for a detailed description of method, at 412.

[120] Monahan, *supra* note X, at 412.

[121] Monahan, *supra* note X, at 119 n.1.

history in the predictive algorithm, but race is explicitly excluded as an input. Government reports justified the exclusion of race by explaining, “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.”[122]

### B. Algorithmic Inputs and the Direct and Proxy Effects of Race

Table 1 summarizes the most commonly-used predictive algorithms used in the criminal justice system. For each predictive algorithm, we list the setting, whether race is excluded as an input, whether any non-race correlates are excluded as inputs, and examples of any excluded non-race correlates.

Table 1: Predictive Algorithms in the Criminal Justice System

| Algorithm | Setting | Excludes Race | Excludes Any Racial Proxies | Examples of Excluded Racial Proxies |
|---|---|---|---|---|
| 1. PredPol | Policing | Yes | Yes | SES |
| 2. SSL | Policing | Yes | No | |
| 3. PSA | Pretrial | Yes | Yes | Education, SES, Neighborhood |
| 4. VPRAI | Pretrial | Yes | No | |
| 5. VA NVRA | Sentencing | Yes | No | |
| 6. COMPAS | Sentencing & Parole | Yes | No | |
| 7. LSI-R | Sentencing & Parole | Yes | No | |
| 8. SFS | Parole | Yes | Yes | Education, Employment |

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

Based on our review, none of the most commonly-used predictive algorithms in the criminal justice system directly use race as an input.[123] 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.[124] The decision to exclude race as an algorithmic input, despite no settled legal precedent on the issue, is likely because the “explicit use of race, ethnicity,

[122] *See* ALLEGHENY COUNTY PREDICTIVE RISK MODELING TOOL IMPLEMENTATION: PROCESS EVALUATION at 7 (Jan. 2018); Dare, Tim, and Eileen Gambrill 2017 Ethical Analysis: Predictive Risk Models at Call Screening for Allegheny County. In Vaithianathan, Rhema, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney 2017 Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. Centre for Social Data Analytics, Auckland University of Technology.

[123] 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.

[124] *See* Michael Tonry, *Legal and Ethical Issues in the Prediction of Recidivism*, 26 Federal Sentencing Reporter 167, 169 (2014). Despite what he perceives as “toothless” legal constraints, Tonry notes that “race, ethnicity, and religion are not to my knowledge anywhere used as an explicit factor in prediction instruments or in sentencing or parole policies” because “explicit use of race, ethnicity, or religion ... is widely regarded as unseemly, and so the issue is unlikely to arise.” Id. at 170; see also Luis Daniel, The Dangers of Evidence-Based Sentencing,GOVLAB BLOG (Oct. 31, 2014), http://thegovlab.org/the-dangers-of-evidence-based-sentencing/ (noting that “[o]verwhelmingly, states do not include race in the risk assessments since there seems to be a general consensus that doing so would be unconstitutional.”); *See, e.g.,* Bjerk and Hyatt, at 227 (“actuarial methods need not include race as a predictor, and to the best of our knowledge, most do not”); *see also* 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”).

or religion...is widely regarded as unseemly."[125] As some have argued, the exclusion of race from predictive algorithms in the criminal justice system "suggests that companies responsible for creating algorithms perceive either legal or reputational costs to using race as a parameter."[126]

Predictive algorithms in the criminal justice system are much more varied in how they deal with non-race correlates and the proxy effects of race. Three of the predictive algorithms we reviewed exclude at least some non-race correlates that seem likely to generate proxy effects, while the remaining five do not explicitly exclude these non-race correlates. PredPol, for example, uses "[n]o demographic, ethnic or socio-economic information....This eliminates the possibility for privacy or civil rights violations."[127] 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." Based on this stance, the PSA excludes both race and non-race correlates such as education, socioeconomic status, and neighborhood of residence.[128] 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."[129] The remaining predictive algorithms use many input factors that are likely to generate racial proxy effects, however. 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.

There is also considerable variation in which non-race correlates are considered problematic, with no clear principle guiding the choice of these non-race correlates. While some of the above algorithms exclude factors that are correlated with race out of a view that these proxy effects are unfair, they also universally include characteristics related to the current offense or defendant's criminal history. We view the universal inclusion of these characteristics 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 predictive algorithm may lead to predictions that are also racially biased. Under the mainstream view of proxy effects, this results in 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 non-race correlates and proxy effects, sometimes excluding inputs deemed to be correlated with race yet also retaining others that are likely also correlated with race, in particular, current offense and criminal history.

[125] Tonry, *supra* note X, at 170.

[126] Huq, *supra* note X, at 36. But as he notes, "current law does not address the precise question whether the availability of race as a potential *input* into the deliberative process that results in state action violates the Equal Protection Clause on anticlassification grounds." *Id.* at 36.

[127] *See supra* note X.

[128] *See supra* note X.

[129] *See supra* note X.

## IV. A Statistical Framework for Predictive Algorithms

In this section, we provide a simple statistical framework that formalizes the mainstream legal consensus outlined in Section II. We then use this framework to formalize 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 non-race 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) non-correlates of protected characteristics. This simple categorization will allow us to both formalize the mainstream legal consensus described in Section II 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 $\mathbf{X}^{\mathbf{Protected}}$. 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.[130]

*Correlates of Protected Characteristics*: The second set of potential algorithmic inputs consists of correlates of protected characteristics, denoted by $\mathbf{X}^{\mathbf{Correlated}}$. 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 non-race correlates may include the zip code of residence, education levels, and employment status,[131] although whether an algorithmic input is actually correlated with race is an empirical question that may vary across contexts.

*Non-Correlates 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 $\mathbf{X}^{\mathbf{Uncorrelated}}$. For simplicity, we assume that $\mathbf{X}^{\mathbf{Uncorrelated}}$ are also uncorrelated with $\mathbf{X}^{\mathbf{Correlated}}$, but all of our results are easily extended to allow for some correlation between $\mathbf{X}^{\mathbf{Uncorrelated}}$ and $\mathbf{X}^{\mathbf{Correlated}}$. In the context of race and the criminal justice system, these uncorrelated characteristics may include the current charge and criminal history, but as above, whether an algorithmic input is actually uncorrelated with race is an empirical question that may vary across contexts.

[130] *See infra* Section VII.A.

[131] *See, e.g.,* Starr, *supra* note X, at 838 ("socioeconomic 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.").

### B. Benchmark Statistical Model

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

$$Y_i = \beta_0 + \beta_\mathbf{1} \cdot \mathbf{X_i^{Uncorrelated}} + \beta_\mathbf{2} \cdot \mathbf{X_i^{Correlated}} + \beta_\mathbf{3} \cdot \mathbf{X_i^{Protected}} + \epsilon_\mathbf{i} \tag{1}$$

where, for simplicity, we assume that each set of input characteristics enters linearly and additively. We consider extensions to this framework in Section VII. $Y_i$ is the observed outcome for individual $i$, $\mathbf{X_i^{Protected}}$ includes all protected characteristics, $\mathbf{X_i^{Correlated}}$ includes all correlated input characteristics, $\mathbf{X_i^{Uncorrelated}}$ includes all uncorrelated input characteristics, and $\epsilon_i$ is an error term that is mean zero and uncorrelated with the potential algorithmic inputs.

In our statistical framework, $\beta_1$, $\beta_2$, and $\beta_3$ represent the true 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. 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, $\mathbf{X_i^{Correlated}}$ is the set of potential inputs that is correlated with the set of protected characteristics $\mathbf{X_i^{Protected}}$. We assume that the relationship between $\mathbf{X_i^{Correlated}}$ and $\mathbf{X_i^{Protected}}$ is equal to:

$$\mathbf{X_i^{Protected}} = \alpha_\mathbf{0} + \alpha_\mathbf{Corr} \cdot \mathbf{X_i^{Correlated}} + \mathbf{v_i} \tag{2}$$

where $\alpha_\mathbf{Corr}$ represents the true relationship between $\mathbf{X_i^{Correlated}}$ and $\mathbf{X_i^{Protected}}$.

### 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. This position is broad in defining all proxy effects as unwarranted, but we view this choice as most consistent with the mainstream legal consensus,[132] 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.[133] We will also illustrate

[132] *See, e.g.* Sandra G. Mayson, *Bias In, Bias Out*, 128 Yale L.J. forthcoming at 3 ("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.").

[133] 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." As noted by Altenburger and Ho, "such classification can be highly contested." Altenburger & Ho, *supra* note X, at X. We share the view of Altenburger and Ho that a "commonsense classification of

these ideas by means of simple examples, showing exactly how both direct use of race and indirect use of non-race 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:

$$Y_i = \beta_0 + \beta_{\mathbf{1}} \cdot \mathbf{X_i^{Uncorrelated}} + \beta_{\mathbf{2}} \cdot \mathbf{X_i^{Correlated}} + \beta_{\mathbf{3}} \cdot \mathbf{X_i^{Protected}} + \epsilon_{\mathbf{i}} \tag{3}$$

With all inputs included in the regression, the estimated coefficients yield unbiased (in the statistical sense) estimates of $\beta_1$, $\beta_2$, and $\beta_3$.

We can then form the following prediction:

$$\hat{Y}_i^{Direct} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathbf{X_i^{Uncorrelated}} + \hat{\beta}_{\mathbf{2}} \cdot \mathbf{X_i^{Correlated}} + \hat{\beta}_{\mathbf{3}} \cdot \mathbf{X_i^{Protected}} \tag{4}$$

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.

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 $\mathbf{X_i^{Protected}}$ is an indicator variable that is equal to one if an individual is black and equal to zero 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.

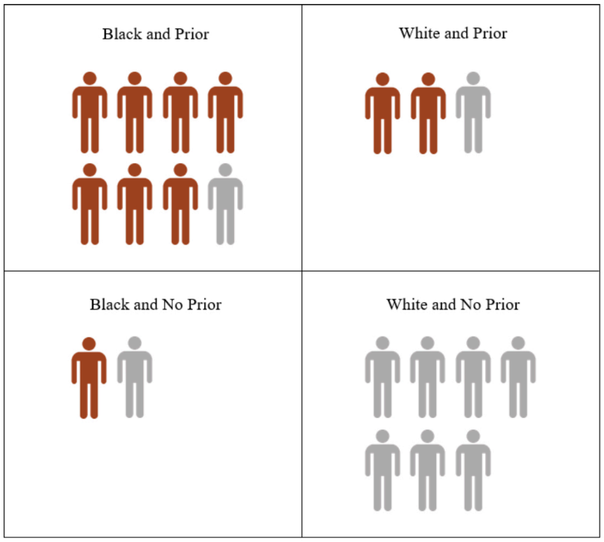
To provide a concrete illustrative example of these direct effects, suppose that there are 100 total individuals (50 black and 50 white), with the distribution of characteristics as follows in Table 2.

In the hypothetical example illustrated in Table 2, eight out of every 10 black individuals have a prior criminal history and three out of every 10 white individuals have a prior criminal history. 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 red, while individuals who will appear at all court appearances are denoted in 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.

Mapping the example in Table 2 to our statistical framework, $Y_i$ is an indicator variable for FTA, $\mathbf{X_i^{Protected}}$ is an indicator equal to 1 if an individual is black and equal to 0 if an individual is white, and $\mathbf{X_i^{Correlated}}$ is an indicator equal to 1 if an individual has a prior criminal history and equal to zero if an individual does not have a prior criminal history. We first estimate Equation (3), where we control for

'socially acceptable' does not necessarily imply statistical independence. Many predictions that may superficially seem "socially acceptable" are in fact highly correlated with race." Altenburger & Ho, *supra* note X, at X.

Table 2: Hypothetical Example to Illustrate the Direct and Proxy Effects of Race



**Note:** This table presents hypothetical relationships between FTA, prior criminal history, and race. Individuals who FTA if released are denoted in red, while individuals who will appear at all court appearances are denoted in gray. See the text for additional details.

race through $\mathbf{X}_\mathbf{i}^{\mathbf{Protected}}$ and prior criminal history through $\mathbf{X}_\mathbf{i}^{\mathbf{Correlated}}$. 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 |
| $R^2$ | 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 the text for additional details.

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, 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 blacks will be treated differently than otherwise identical whites is at the heart of the mainstream argument that including race in predictive algorithms would constitute a violation of the Equal Protection Clause.[134] To address this legal concern, most if not all predictive algorithms therefore exclude race as an input.[135]

The use of direct effects can result in large racial gaps in predicted risk. If predictions were truly race-neutral, the average predicted risk for whites is 0.37 and the average predicted risk for blacks is 0.64. When direct effects are used to predict risk, the average predicted risk for whites is 0.20 and the average predicted risk for blacks is 0.80. Thus, when direct effects are incorporated into algorithmic predictions, blacks are disadvantaged relative to whites.

*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:

$$Y_i = \gamma_0 + \gamma_\mathbf{1} \cdot \mathbf{X_i^{Uncorrelated}} + \gamma_\mathbf{2} \cdot \mathbf{X_i^{Correlated}} + \epsilon_\mathbf{i} \tag{5}$$

We can then form the following prediction:

$$\hat{Y}_i^{Proxy} = \hat{\gamma}_0 + \hat{\gamma}_1 \cdot \mathbf{X_i^{Uncorrelated}} + \hat{\gamma}_\mathbf{2} \cdot \mathbf{X_i^{Correlated}} \tag{6}$$

where $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the estimated relationship between each set of inputs and the outcome of interest.

The coefficient $\hat{\gamma}_2$ estimated in Equation (5), is, in general, <u>not</u> identical to the estimated coefficient $\hat{\beta}_2$ estimated in Equation (3). Recall that we have assumed that $\mathbf{X_i^{Correlated}}$ is correlated with $\mathbf{X_i^{Protected}}$, and that $\mathbf{X_i^{Protected}}$ is predictive of the outcome such that $\beta_\mathbf{3} \neq \mathbf{0}$. 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. We can illustrate these proxy effects by substituting the expression for $\mathbf{X_i^{Protected}}$ from Equation (2) into Equation (1). Doing so yields the following expression:

$$Y_i = (\beta_0 + \beta_\mathbf{3}\alpha_\mathbf{0}) + \beta_\mathbf{1} \cdot \mathbf{X_i^{Uncorrelated}} + (\beta_\mathbf{2} + \beta_\mathbf{3}\alpha_\mathbf{Corr}) \cdot \mathbf{X_i^{Correlated}} + (\epsilon_\mathbf{i} + \beta_\mathbf{3}\mathbf{v_i}) \tag{7}$$

The standard OVB formula shows us that $\hat{\gamma}_2$ is not a consistent estimate of $\beta_2$, but rather the expression $(\beta_\mathbf{2} + \beta_\mathbf{3}\alpha_\mathbf{Corr})$. Intuitively, $\beta_\mathbf{2}$ includes the portion of the correlated characteristics that is orthogonal to (or uncorrelated with) protected characteristics, and $\beta_\mathbf{3}\alpha_\mathbf{Corr}$ includes the portion of the correlated characteristics that is purely a proxy for protected characteristics. One can think of $\beta_\mathbf{2}$ as capturing predictive

[134] *Cf.* Oleson, *supra* note X, at 1386; Kopf, *supra* note X, at 213 for scholars who argue that direct effects of race should be included because they increase predictive accuracy, a compelling government interest.

[135] *See supra* Section III.

variation in the correlated characteristics *within* a protected class, and $\beta_{\mathbf{3}}\alpha_{\mathbf{Corr}}$ 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, $\beta_{\mathbf{3}}\alpha_{\mathbf{Corr}}$. 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.

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.

Table 4 presents estimates from a series of OLS regressions of FTA on potential algorithmic 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 $\beta_{\mathbf{2}} + \beta_{\mathbf{3}}\alpha_{\mathbf{Corr}}$, where $\beta_{\mathbf{3}}\alpha_{\mathbf{Corr}}$ 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 true predictive relationship between that input and the probability of FTA, $\beta_{\mathbf{2}}$.

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 true predictive relationship is only 54.1 percentage points, meaning that the proxy effects of race add 16.7 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 true 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 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 truly race-neutral, the average predicted risk for whites is 0.37 and the average predicted risk for blacks is 0.64. When proxy effects are used to predict risk (even when direct effects are excluded), the average predicted risk for whites is 0.32 and the average predicted risk for blacks is 0.67. Thus, proxy effects in algorithmic predictions also disadvantage blacks relative to whites.

*Summary*: We have demonstrated theoretically 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.

Table 4: Hypothetical Example of Proxy Effects

| | Prob of FTA | | |
|---|---|---|---|
| | Proxy Effects | No Proxy Effects | Difference (1) - (2) |
| | (1) | (2) | (3) |
| Prior Criminal History | $0.707^{***}$ | $0.541^{***}$ | $0.167^{***}$ |
| | (0.072) | (0.077) | (0.059) |
| Black | | $0.330^{***}$ | |
| | | (0.076) | |
| Constant | $0.111^{**}$ | 0.038 | |
| | (0.054) | (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 the text for additional details.

## V. Legal and Statistical Solutions to Ensuring Race Neutrality

In this section, we discuss three potential solutions that can eliminate the direct and proxy effects of race and non-race correlates in predictive algorithms. The first legal solution follows both current practice and the mainstream legal consensus by excluding both race and all non-race 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. Our first recommended solution instead purges all algorithmic inputs of the proxy effects of race in the estimation of the predictive algorithm, then uses these “colorblind” inputs to predict outcomes. Our second solution instead uses only whites in the estimation of the predictive algorithm, then uses these “colorblind” estimates to predict outcomes for both whites and blacks.

### A. Current Legal 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, $\mathbf{X_i^{Protected}}$, *and* excludes using any correlated inputs, $\mathbf{X_i^{Correlated}}$. 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 $\mathbf{X_i^{Uncorrelated}}$. We believe that this solution most intuitively follows from legal definitions of fairness given our survey of risk assessment tools as reviewed in Section III, which generally explicitly exclude race and often factors that are correlated with race out of a view that their inclusion would be illegal, unethical, and/or unjust. Many even 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 "an effective risk assessment must be gender and race neutral....[t]he more risk factors you have, the less likely you'll be able to eliminate gender and racial bias."[136]

*Estimation of the Algorithm*: To illustrate how this algorithm would form predictions, we first estimate the following statistical relationship, using only uncorrelated inputs:

$$Y_i = \delta_0 + \delta_{\mathbf{1}} \cdot \mathbf{X_i^{Uncorrelated}} + \epsilon_{\mathbf{i}} \tag{8}$$

We can then form the following predictions:

$$\hat{Y}_i^{ExcludingInputs} = \hat{\delta}_0 + \hat{\delta}_1 \cdot \mathbf{X_i^{Uncorrelated}} \tag{9}$$

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 $\mathbf{X_i^{Uncorrelated}}$ 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 $\mathbf{X_i^{Protected}}$ and/or $\mathbf{X_i^{Correlated}}$ 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.

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. Perhaps because of the likely impossibility of finding uncorrelated inputs, most if not all predictive algorithms today fail to meet the standard of race-neutrality. Recall that some existing risk assessment instruments in the criminal justice system attempt to eliminate the direct and proxy effects of race by excluding race and race correlates, but as we will argue below, they fail to do so because many, if not all, of the remaining inputs are highly correlated with race. As a result, there is no guarantee that the estimates from these predictive algorithms rely only on $\mathbf{X_i^{Uncorrelated}}$ and are truly race-neutral.

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

We now turn to our first statistical solution, which we call 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,

[136] See https://www.wired.com/story/bail-reform-tech-justice/.

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 correlated inputs should be included or excluded. 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 legal solution. 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 legal solution would be impossible to implement.

As we will formally 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 non-race inputs, and then ignores individual race information in the prediction step. In theory, using race in the estimation step may run counter to the intuitive but statistically incorrect legal mainstream position that the use of a protected characteristic always violates the Equal Protection Clause. To the best of our knowledge, this algorithm is not used in practice today, likely because of the formal legal prohibition on the use of protected characteristics and the statistically incorrect arguments supported by the mainstream legal position.

*Estimation of the Algorithm*: To construct our colorblinding-inputs model, we follow the approach developed by Pope and Sydnor (2011) which utilizes only the predictive power from input variables that is *orthogonal* to (or uncorrelated with) protected characteristics. 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, our model is estimated in two steps. In the first step, we estimate the benchmark statistical case from Equation (1) that includes the full set of input characteristics:

$$Y_i = \beta_0 + \beta_\mathbf{1} \cdot \mathbf{X_i^{Uncorrelated}} + \beta_\mathbf{2} \cdot \mathbf{X_i^{Correlated}} + \beta_\mathbf{3} \cdot \mathbf{X_i^{Protected}} + \epsilon_\mathbf{i} \tag{10}$$

where, as discussed previously, the estimates from this model yield the unbiased 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 $\mathbf{X_i^{Protected}}$. Thus, this first estimation step ensures that we eliminate all proxy effects from including $\mathbf{X_i^{Correlated}}$. Intuitively, we ensure that the estimated relationship between our outcome of interest and $\mathbf{X_i^{Correlated}}$ is unbiased by only keeping the predictive power from $\mathbf{X_i^{Correlated}}$ that is orthogonal to (or uncorrelated with) $\mathbf{X_i^{Protected}}$. As a result, we are able to colorblind $\mathbf{X_i^{Correlated}}$.

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 <u>not</u> $\hat{\beta}_3$, to form risk predictions. Specifically, we form predictions that use the average $\mathbf{X_i^{Protected}}$ across all individuals, $\overline{\mathbf{X}}^\mathbf{Protected}$, but the actual input values of $\mathbf{X_i^{Uncorrelated}}$ and $\mathbf{X_i^{Correlated}}$. We therefore form the following prediction:

$$\hat{Y}_i^{ColorblindingInputs} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \mathbf{X_i^{Uncorrelated}} + \hat{\beta}_\mathbf{2} \cdot \mathbf{X_i^{Correlated}} + \hat{\beta}_\mathbf{3} \cdot \overline{\mathbf{X}}^\mathbf{Protected} \tag{11}$$

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

To provide a concrete example, suppose again that $Y_i$ is an indicator variable for FTA, $\mathbf{X_i^{Protected}}$ is an indicator equal to 1 if an individual is black, and $\mathbf{X_i^{Correlated}}$ is an indicator equal to 1 if an individual has a prior criminal history. Return to the hypothetical distribution of individuals from Table 2. In the first step, we estimate predictions 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 step, rather than use the real 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, which is 50 percent by construction (see Table 2). 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.

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

We now introduce a second statistical solution, which we call 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. The difference is that the minorities-as-whites algorithm uses only the predictive power within whites, not both whites and minorities, thereby ensuring that the algorithm treats minorities exactly the same way it treats whites. That is, we use only whites in the estimation of the predictive algorithm, then rely on the resulting "colorblind" predictive relationships to predict outcomes for both whites *and* non-whites. By focusing only on whites in the estimation step, there is 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 non-whites 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 whites alone can eliminate any proxy effects.

A key feature of the minorities-as-whites algorithm is estimating predictive relationships only on the white population, which requires the consideration of race in the estimation step. To the best of our knowledge, this algorithm is also not used in practice today, likely because of the perceived legal prohibition on the use or consideration of protected characteristics.

*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 non-white individuals.

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

$$Y_i^W = \beta_0^W + \beta_\mathbf{1}^\mathbf{W} \cdot \mathbf{X_i^{Uncorrelated}} + \beta_\mathbf{2}^\mathbf{W} \cdot \mathbf{X_i^{Correlated}} + \beta_\mathbf{3}^\mathbf{W} \cdot \mathbf{X_i^{Protected}} + \epsilon_\mathbf{i}^\mathbf{W} \tag{12}$$

where the estimates from this model yield the unbiased coefficients for whites $\hat{\beta^W}_1$, $\hat{\beta^W}_2$, and $\hat{\beta^W}_3$.

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

$$\hat{Y}_i^{MinoritiesasWhites} = \hat{\beta}_0^W + \hat{\beta}_1^W \cdot \mathbf{X_i^{Uncorrelated}} + \hat{\beta}_\mathbf{2}^\mathbf{W} \cdot \mathbf{X_i^{Correlated}} + \hat{\beta}_\mathbf{3}^\mathbf{W} \cdot \mathbf{X_i^{Protected}} \tag{13}$$

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, suppose again that $Y_i$ is an indicator variable for FTA, $\mathbf{X_i^{Protected}}$ is an indicator equal to 1 if an individual is black, and $\mathbf{X_i^{Correlated}}$ is an indicator equal to 1 if an individual has a prior criminal history. Return again to the hypothetical distribution of individuals from Table 2. In the first 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 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 briefly 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 legal excluding-inputs solution is that both our statistical proposals explicitly require the consideration and use of race in the estimation process precisely in order to achieve a race-neutral prediction. In contrast, the excluding-inputs algorithm prohibits the use of race and all race-correlates.

A lack of understanding of the underlying statistical properties of direct and proxy effects in algorithms may lead a naive observer to conclude that both proposals are illegal because they run up against the widely accepted prohibition on the use or consideration of protected characteristics.[137] However, we argue that

[137] 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

under common conceptions of the Equal Protection Clause, both statistical solutions should be legally permissible. First, consider the anti-classification principle, which many argue drives our understanding of the Equal Protection doctrine. This anti-classification 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."[138] If we take the anti-classification principle as the dominant concern underlying the Equal Protection Clause, the strongest critique against our statistical approaches, which require the use of race in the first estimation step, may be that they constitute express racial classifications that trigger strict scrutiny. However, we believe that our approaches should not be subject to strict scrutiny given that the use/consideration of race is not meant to distinguish or treat individuals differently on the basis of membership in a particular racial group, but the exact opposite.[139] Even supposing that our approaches were treated as express racial classifications, we argue that they are narrowly tailored towards the aim of remedying and correcting for proxy effects and historical biases that can be "baked in" to an algorithm, a compelling state interest, such that they should withstand strict scrutiny.[140] Ultimately, we believe that our approaches do not violate the core tenet that underlies the anti-classification principle. Because our proposed solutions use race precisely to purge algorithmic predictions of any proxy effects that are due to membership in a particular group, our approaches are very much in line with the goal of treating citizens as individuals.

Second, and more broadly, we view our proposed solutions as consistent with an alternative conception of equal protection, the anti-subordination 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 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."[141] We view both our statistical

prohibited characteristics." Anya Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, IOWA L. Rev. forthcoming, at 57; *See also* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, available at https://arxiv.org/abs/1808.00023 ("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.")

[138] 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 Charles R. Lawrence III, *Two Views of the River: A Critique of the Liberal Defense of Affirmative Action*. 101 COLUM. L. REV. 928, 950 (2001) (associating the anti-classification 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"); *Missouri v. Jenkins*, 515 U.S. 70, 120021 (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...." (citations omitted)).

[139] Prince and Schwarcz note with respect to the colorblinding-inputs model, that "in a very real sense, the process explicitly discriminates with respect to membership in a legally protected group in order to prevent the effects of such discrimination from being felt by these individuals." Prince and Schwarcz, *supra* note X, at 58. In addition, there are other examples of race-conscious policies that have not been treated as express racial classifications. As Richard Primus has noted, "many practices that do involve government actors' identifying people by race are not always subject to strict scrutiny." Richard Primus, *Equal Protection and Disparate Impact*, *supra* note X, at 505.

[140] *See, e.g., Adarand v. Pena*, 515 U.S. 200, 227 (1995) (holding that 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..." must withstand strict scrutiny under the Equal Protection Clause).

[141] David A. Strauss, *"Group Rights and the Problem of Statistical Discrimination*, 17 Issues in Legal Scholarship 1, 1 (2003);

proposals as relying on race for a non-subordinating use, e.g., our statistical solutions use race precisely to avoid inflicting harm on disadvantaged groups, fully consistent with the anti-subordination principle.

### E. Racial Gaps 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 legal excluding-inputs algorithm, are not only implementable in practice but also better advance the widespread goal of 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. 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 mainstream legal consensus. Resulting racial disparities after elimination of direct and proxy effects, are thus by definition, not unwarranted. Specifically, our two statistical solutions would not ensure that predictions for protected classes and non-protected classes would be identical across all populations in a way that eliminates all racial disparity. In fact, 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 of failing to appear in court (see Table 4). If blacks, on average, are more likely to have a prior criminal history compared to whites, risk predictions may still be higher on average for blacks relative to whites. Some may argue that the overrepresentation of prior criminal records for blacks relative to whites 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. Unfortunately, we are not aware of any systematic approach of dealing with these measurement issues, either when dealing with algorithms or human decision-makers. For example, we are not aware of any government that attempts to correct for mismeasurement of say, prior conviction records for use in sentencing recommendations, by adjusting what it means to have a prior for black offenders versus white offenders.[142] The only real solution is to understand the possible sources of measurement error and find inputs that do not suffer from measurement error, a worthwhile goal when dealing with both algorithms and human decision-making.

*see also* Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N.Y.U. L. REV. 1003, 1007-1008 (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 X, 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.").

[142] In facts, as scholars have pointed out, some differences by protected characteristics can be justified under anti-discrimination law, such that "racial balance... is not legally mandated, and efforts to pursue that goal might themselves be struck down on constitutional grounds." Sunstein, *supra note X*, at 8.

## VI. Empirical Tests of Our Proposed Statistical Solutions

In this section, we present our main empirical results using information 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 nearly all commonly-used algorithms generate unwarranted racial gaps under the mainstream legal position 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 non-excessive bail is guaranteed by the Eighth Amendment to the U.S. Constitution, with almost all state constitutions granting similar rights to defendants. In New York, the state constitution states that "[e]xcessive bail shall not be required nor excessive fines imposed...."[143] New 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),"[w]hen 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." 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. If the defendant only has charges that are less than felony grade, the court must order recognizance or bail.[144] 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.[145]

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. The fingerprints are then sent to the Division of Criminal Justice Services (DCJS) in Albany to obtain the defendant's criminal history. During this time, 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." 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

[143] N.Y. CONST. art. I, §5.

[144] NY CPL §530.20.

[145] *See* NY CPL §510.30 ("With respect to any principal, the court must consider the kind and degree of control or restriction that is necessary to secure his court attendance when required."); *see also Matter of Sardino v. State Comm'n on Judicial Conduct*, 58 N.Y.2d 286, 289 (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.").

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.[146]

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 24 hours of arrest.[147] However, not all arrested individuals will be held in holding cells prior to arraignment. For individuals with no outstanding warrant at the time of arrest, the arresting police officer may use his or her discretion to issue a Desk Appearance Ticket (DAT) if the arrest charge is a violation, misdemeanor, or Class E felony.[148] This DAT allows the arrested individual to be released but requires them to return to court for a later pre-scheduled arraignment, with 28 percent of all misdemeanor arrests issued a DAT in 2016.[149] 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 percent).[150]

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.[151] The New York City arraignment court is open 365 days a year until 1AM each day. In contrast to some other jurisdictions, almost half of all case filings are disposed of at arraignment in New York City. For many misdemeanor defendants, for example, the case is often dismissed at arraignment or adjourned in contemplation of dismissal (ACD).[152] In 2013, for example, about 80 percent of first-time nonviolent misdemeanor youth had their cases resolved with an outright dismissal or ACD.[153]

For the cases that are not disposed of at arraignment, the assigned arraignment judge has a number of potential 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 percent of defendants in New York City are released ROR at arraignment such that no bail is set and no other court conditions are mandated.[154] Second, defendants may be required to post some sort of bail payment to secure release if they pose an appreciable risk of flight. In New York City, arraignment judges are required by law to set at least two forms of bail in these cases, which make up most of the remaining 30 percent of cases.[155] 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.[156] Cash bail requires the individual to pay the full bail amount upfront in order to secure release while insurance company bail bond requires an individual to deposit 10 percent 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

[146] *See* Criminal Court of the City of New York, 2013 Annual Report (hereinafter "NYC 2013 Annual Report"), at 18-20.

[147] *See* A More Just New York City: Independent Commission on New York City Criminal Justice and Incarceration Reform 2017 (hereinafter "The Lippman Report"), available at https://static1.squarespace.com/static/577d72ee2e69cfa9dd2b7a5e/t/595d48ab29687fec7526d338/1499285679244/Lippman+Commission+Report+FINAL+Singles.pdf.

[148] *Id.* DATs are not permitted for other types of felonies (e.g. Class A-D felonies).

[149] *Id.*

[150] *See* Lippman Report, *supra* note X, at X.

[151] *See* NYC 2013 Annual Report, *supra* note X, at 28.

[152] *See* Lippman Report, *supra* note X, at X.

[153] See Lippman Report, *supra* note X, at X.

[154] *See* Lippman Report, *supra* note X, at X.

[155] *See* Lippman Report, *supra* note X, at X (citing People ex rel. McManus v. Horn, 18 N.Y.3d 660 (2012)).

[156] NY CPL §520.10.

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 percent; and unsecured bonds that do not require upfront 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.[157] If the defendant is remanded or is unable to make the set bail, he or she is detained on Rikers Island or in county jails throughout the city until the adjudication of their case. For more serious crimes, the arraignment judge may require that the defendant is detained pending trial by denying bail altogether. 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. For example, in New York City, a class A felony, which includes murder, kidnapping, arson, and high-level drug possession and sale, almost always results in a denial of bail. These cases make up about 0.8 percent of all cases in New York City.[158] Finally, there are also about 1.5 percent of cases that are sent to a supervised release program as an alternative to pretrial detention.[159]

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 character, reputation, habits and mental condition; his employment and financial resources; ... his family ties and the length of his residence if any in the community; ... his criminal record, if any; ... his record of previous adjudication as a juvenile delinquent, or a youthful offender, if any; ... his previous record if any in responding to court appearances when required or with respect to flight to avoid criminal prosecution; ... where the principal is charged with a crime or crimes against a member or members of the same family or household...any violation by the principal of an order of protection...and the principal's history of use or possession of a firearm; ... the weight of evidence against him in the pending criminal action and any other factor indicating probability or improbability of conviction; and the sentence which may be or has been imposed upon conviction.[160]

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 resources" when setting bail,[161] many have noted that there is little evidence that judges consider individual ability to pay in practice.[162] 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.[163]

[157] In New York, there is a 3 percent surcharge on all cash bail if the defendant is convicted, which the government keeps. *See* https://www.nycourts.gov/courthelp/Criminal/bail.shtml.

[158] *See* Lippman Report, *supra* note X, at X.

[159] *See* Lippman Report, *supra* note X, at X.

[160] NY CPL §510.30.

[161] NY CPL §510.30 (a.2).

[162] *See* Lippman Report, *supra* note X, at X ("if 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.").

[163] *See* Emily Leslie and Nolan Pope, *The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from NYC*

*NYC Arraignment Judges*: The Criminal Court and its judges have the responsibility of conducting arraignments in New York City. Judges serving in the Criminal Court are appointed by the Mayor of the City of New York from a list of candidates selected by the Mayor's Advisory Committee on the Judiciary.[164] The selection committee is comprised of up to nineteen members: nine selected by the Mayor, four selected by the Chief Judge of the New York Court of Appeals, two from among the Presiding Justices of the Appellate Divisions of the Supreme Court for the First and Second Judicial Departments, and deans of the law schools in New York City.[165] Candidates for the Criminal Court must be New York City residents who were admitted to practice as attorneys in New York for at least 10 years prior to their potential selection. Candidates must also complete the Uniform Judicial Questionnaire to be considered as potential nominees.[166]

Criminal Court judges serve 10-year terms unless appointed to fill a vacancy, in which case the judge serves out the remainder of the vacated term.[167] Before beginning their term, newly appointed judges in any New York jurisdiction must attend a three-day judicial training institute at Pace Law School. Criminal Court judges must also attend a two- to three-week training program in Manhattan that consists of classroom instruction and shadowing arraignment judges before beginning their term.[168] For the one to two years of their first term, Criminal Court judges usually spend nearly all of their time presiding over arraignments. Criminal Court judges will then, in general, spend more of their time presiding over trials and sentencing decisions in future years of their first term. Criminal Court judges may then be reappointed after their term expires, but must step down after December 31st of the year in which they reach the age of 70.[169]

By statute, the New York City Criminal Court may only have 107 judges. However, due to the volume of arraignments and the other responsibilities of Criminal Court judges, judges from the Civil and Family Courts, as well as Supreme Court judges, may also preside over arraignments. For example, newly appointed Civil Court judges will often spend their first year or two of their term presiding over arraignments due to staffing needs.

*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 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. These bail funds include the Bronx Freedom Fund, the Brooklyn Community Bail Fund, and the Liberty Fund. 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.[170] 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

*Arraignment*, Journal of Law and Economics (2017), 60(3): 529-557.

[164] NY Constitution Article 6, §15(a).

[165] *See* "How to Become a Judge," New York City Bar 2012, available at www.nycbar.org/images/stories/pdfs/becomeajudge2012.pdf.

[166] NYC Criminal Court Act §22(1).

[167] NYC Criminal Court Act §22(2).

[168] *See* https://www.nytimes.com/2003/05/18/realestate/in-the-region-westchester-institute-for-state-judges-opens-at-pace-law-school.html; *see also* https://www.nytimes.com/2018/01/04/nyregion/judges-new-york.html.

[169] NY Constitution Article 6, §25(b).

[170] *See* Lippman Report, *supra* note X, at X; *see also* https://www.courtinnovation.org/node/20042/more-info.

the program.[171] In 2017, New York's Criminal Justice Agency and the Mayor's Office of Criminal Justice then announced that it was redesigning its pretrial risk assessment tool.[172] The current tool classifies almost 50 percent of defendants as having a high risk of failing to appear, but only one in five of these high-risk individuals actually misses a required court appearance.[173] 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 hours.[174] 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.

### B. Data Description

This subsection 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.[175] These data contain information on a defendant's gender, race, date of birth, and county of arrest. The data also include extensive information that would also be 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 disparities but our tests can be easily extended to allow for other racial/ethnic comparisons. 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). Finally, we further limit the sample to the approximately 85 percent 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). The final sample thus contains 264,379 cases from 180,887 unique defendants.

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, has 2.0 prior misdemeanor convictions, 0.5

[171] *See* Lippman Report, supra note X, at X.

[172] *See* https://www.nycja.org/resources/details.php?id=1388.

[173] *See* Lippman Report, *supra* note X, at X.

[174] *See* https://www1.nyc.gov/office-of-the-mayor/news/226-18/mayor-de-blasio-launch-online-bail-new-york-city.

[175] These data exclude undocketed arrests as well as the substantial number of arrests for non-fingerprintable charges such as violations, infractions, and many misdemeanors (i.e. VTL 511s).

prior felony convictions, and 1.6 prior failures to appear. Fifty percent of released defendants also have a prior violent felony conviction, with 12 percent having a violent felony charge on the current case. Nineteen percent are charged with at least one drug charge, 6.0 percent with at least one DUI charge, 9.0 percent with at least one property charge, and 43.0 percent 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 percent of released defendants are released ROR at arraignment, with the remaining 18.0 percent 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 percent 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,"[176] which introduces proxy effects in forming predictions, generating arguably unwarranted racial disparities.

To demonstrate how proxy effects infiltrate commonly-used algorithms, we focus on one of the most prominent models in the pretrial context, the Arnold Ventures PSA. The PSA is designed to be both objective and fair, "not contain[ing] factors that would lead defendants to be treated differently because of their race, gender, or socioeconomic status."[177] For this reason, the PSA excludes factors that Arnold Ventures deem to be inconsistent with fairness under the law, including race, gender, socioeconomic status, and neighborhood, or what we call $\mathbf{X_i^{Protected}}$. 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 $\mathbf{X_i^{Correlated}}$ and $\mathbf{X_i^{Uncorrelated}}$. Implicit, however, in the PSA's mission statement of not treating individuals differently because of race, is the assumption that inputs like prior criminal history are not 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.

[176] Cathy O'Neil, *The Ethical Data Scientist*, SLATE (Feb. 4, 2016), https://slate.com/technology/2016/02/how-to-bring-better-ethics-to-data-science.html.

[177] Anne Milgram, Alexander M. Holsinger, Marie VanNostrand, and Matthew W. Alsdorf. *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making*, 27 Federal Sentencing Reporter 216, 220 (2015).

Table 5: Descriptive Statistics

| | All<br>Defendants | White<br>Defendants | Black<br>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 in the Arnold Ventures PSA are, in fact, $\mathbf{X_i^{Uncorrelated}}$ or $\mathbf{X_i^{Correlated}}$ in real-world data. We test whether the types of input variables used in the PSA are $\mathbf{X_i^{Uncorrelated}}$ or $\mathbf{X_i^{Correlated}}$ 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. 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. Again, if a potential input variable is uncorrelated with race (but correlated with pretrial misconduct), then it is $\mathbf{X_i^{Uncorrelated}}$ in our statistical framework. If, on the other hand, a variable is correlated with race, then it is $\mathbf{X_i^{Correlated}}$ in our statistical framework.

Table 6 presents the results from this first empirical test using our dataset on released black and released 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. 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, are less likely to have DUI and property charges, and more likely to be charged with a violent offense, again correlations that will lead to proxy effects if these inputs are included.

Table 6: Correlation Between Race and Algorithmic Inputs

| | *Dependent Variable*: Indicator for being Black | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 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 |
| $R^2$ | 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 – our 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 non-race input variables, but excluding defendant race. 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 | $-0.026^{***}$ | $-0.028^{***}$ | $0.002^{***}$ |
| | (0.001) | (0.001) | (0.000) |
| Violent Felony Charge | $-0.056^{***}$ | $-0.054^{***}$ | $-0.001^{***}$ |
| | (0.002) | (0.002) | (0.000) |
| Prior Misdemeanor Convictions | $-0.003^{***}$ | $-0.003^{***}$ | $0.000^{***}$ |
| | (0.000) | (0.000) | (0.000) |
| Prior Felony Convictions | $-0.010^{***}$ | $-0.008^{***}$ | $-0.001^{***}$ |
| | (0.001) | (0.001) | (0.000) |
| Prior Violent Felony Convictions | $0.007^{***}$ | $0.008^{***}$ | $-0.001^{***}$ |
| | (0.002) | (0.002) | (0.000) |
| Prior Failures to Appear | $0.024^{***}$ | $0.024^{***}$ | $-0.001^{***}$ |
| | (0.000) | (0.000) | (0.000) |
| County Income in 10,000s | $0.005^{***}$ | $0.004^{***}$ | $0.001^{***}$ |
| | (0.000) | (0.000) | (0.000) |
| Drug Charge | $-0.038^{***}$ | $-0.039^{***}$ | $0.001^{***}$ |
| | (0.002) | (0.002) | (0.000) |
| DUI Charge | $-0.064^{***}$ | $-0.072^{***}$ | $0.008^{***}$ |
| | (0.003) | (0.003) | (0.000) |
| Property Charge | $-0.013^{***}$ | $-0.015^{***}$ | $0.002^{***}$ |
| | (0.003) | (0.003) | (0.000) |
| Violent Charge | $-0.056^{***}$ | $-0.056^{***}$ | $-0.000$ |
| | (0.002) | (0.002) | (0.000) |
| Black | $0.035^{***}$ | | |
| | (0.002) | | |
| Constant | $0.179^{***}$ | $0.215^{***}$ | $-0.036^{***}$ |
| | (0.003) | (0.003) | (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 reports 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 currently non-controversial 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 will 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 legal 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 anti-discrimination 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 section 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 Blacks-as-Whites 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 non-race 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, precisely to eliminate non-race 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). The colorblinding-inputs model requires that we include race, just as in the benchmark model, when estimating the coefficients on all other input variables, as described previously in Section IV. Recall that including race allows us to construct race-orthogonal predictions that exclude both the direct and indirect effects of race, 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 both 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, but correlated non-race 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 reports 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 blacks-as-whites 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 blacks-as-whites 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 blacks-as-whites 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 blacks-as-whites 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 blacks-as-whites 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 Blacks-as-Whites Statistical Models

| | *Dependent Variable*: Failure to Appear | | | |
|---|---|---|---|---|
| | Benchmark Model | Excluding Race | Blacks as Whites | 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 reports 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 to construct risk predictions for every defendant in our sample under each algorithm. We then simulate different release policies, calculating the fraction of black (versus white) defendants among those released and the FTA rate 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. 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,[178] 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 black among released defendants if we were to make pretrial release decisions using each of the different predictive algorithms. The x-axis in Figure 1 varies the percent of all defendants that are released, ranging from 0 to 100 percent – what we call the "release threshold." The y-axis reports the fraction of released defendants that are black at each release threshold. 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 except for race, (3) our proposed colorblinding-inputs model and (4) our proposed blacks-as-whites model.

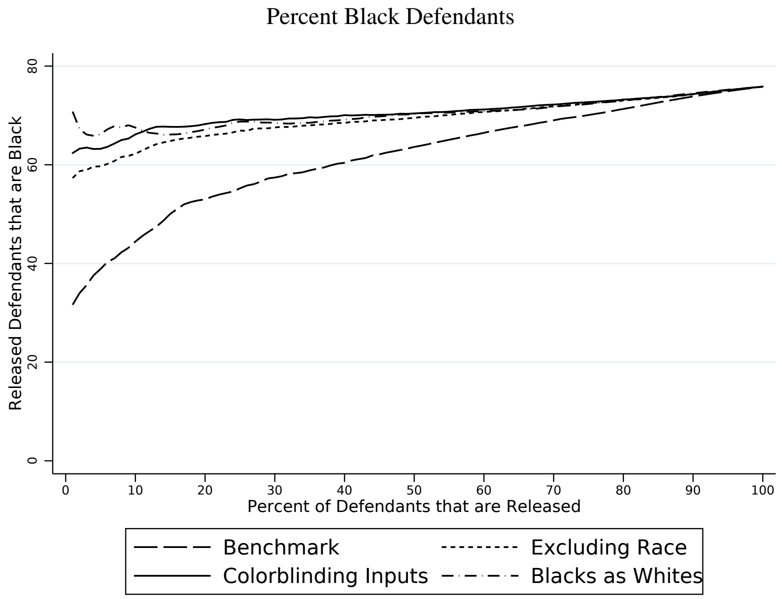
This figure reveals that the benchmark statistical model always results in the lowest share of black defendants among the released at each release threshold. This occurs because the benchmark statistical model uses race as a predictive input, giving rise to direct race effects. Given that being black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

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 over-detention of black defendants in the commonly-used algorithm. Similarly, our proposed blacks-as-whites 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.

[178] *See, e.g.,* Sunstein, *supra note X*, at 8 ("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.")

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 11/2008 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.

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 percent 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, commonly-used, colorblinding-inputs, and blacks-as-whites models, respectively. Column 5 reports the difference in share released that are black between the commonly-used model and the colorblinding-inputs model. Column 6 reports the difference in share released that are black between the commonly-used model and the blacks-as-whites model.

These results again show that our proposed colorblinding-inputs statistical model would significantly increase the fraction of blacks 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 percent of released defendants being black if the overall release rate was set at a threshold of 50 percent (column 1). The commonly-used model would increase the fraction of released defendants who are black to 69.5 percent (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 percent (column 3), almost a full percentage point increase compared to the commonly-used algorithm (column 5). If applied citywide, this model would release an additional 1,700 black defendants during our sample period compared to the typical algorithm used today if 50 percent of all defendants are released. In a similar nature, our proposed blacks-as-whites 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 1,500 black defendants during our sample period.

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 percent 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 sizeable increases in the number of released black defendants relative to the commonly-used algorithm.

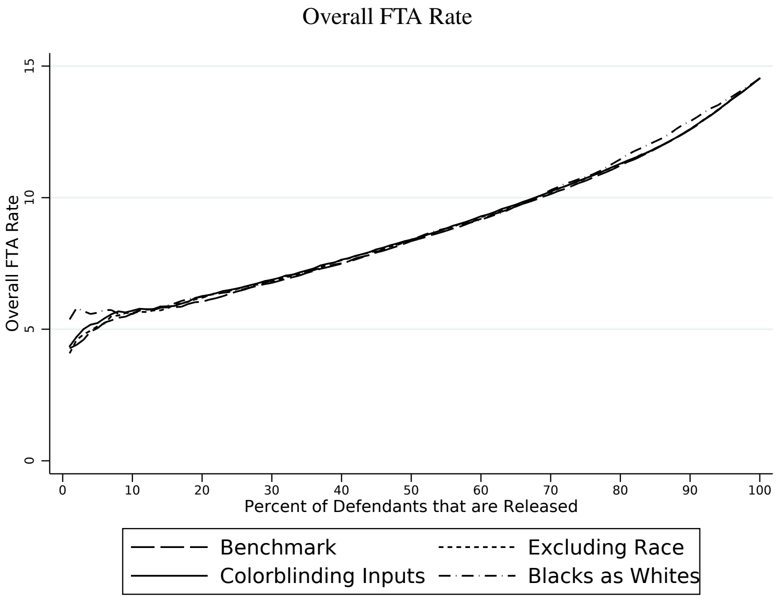
Recall that our measure of pretrial misconduct is failing to appear in court given that nonappearance at court is the only legitimate concern taken into account at the time of setting bail in New York City. However, our results are similar if we used our proposed algorithms to predict the risk of being 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 percent, our proposed colorblinding-inputs model would lead to the release of an additional 1,600 black defendants and our proposed blacks-as-whites algorithm would lead to the release of an additional 3,800 black defendants compared to the commonly-used algorithm.

Table 10: Simulations of Racial Disparities Under Different Predictive Algorithms

| | Share of Blacks Released | | | | | |
|---|---|---|---|---|---|---|
| | Benchmark Model | Excluding Race | Colorblinding Inputs | Blacks as Whites | Difference (2) - (3) | Difference (2) - (4) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| 50 Percent Release Rate | 63.61 | 69.49 | 70.40 | 70.29 | -0.92 | -0.81 |
| 70 Percent Release Rate | 69.01 | 71.83 | 72.20 | 71.75 | -0.37 | 0.08 |
| 90 Percent 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 blacks-as-whites 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 blacks-as-whites 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 amiss to not also illustrate that the choice of predictive algorithm comes with trade-offs in terms of accuracy, as mentioned in Section II.C. Recall that the benchmark statistical model, which uses all inputs, maximizes predictive accuracy. As we begin to eliminate both 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 paper, 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 percent of defendants are released, and algorithm A results in a 20 percent FTA rate among the released and algorithm B results in a 30 percent 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 blacks-as-whites 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 percent 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 model and the colorblinding-inputs model, and column 6 reports the difference in FTA rates between the commonly-used model and the blacks-as-whites 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 algorithm and our proposed algorithms are economically small. For example, if applied citywide in NYC, the colorblinding-inputs model would result in an additional 8 failures to appear during our sample period compared to the commonly-used algorithm if the city decided to release 50 percent of all defendants.

Table 11: Simulations of Accuracy Under Different Predictive Algorithms

| | FTA Rate Among Released Defendants | | | | | |
|---|---|---|---|---|---|---|
| | Benchmark Model | Excluding Race | Colorblinding Inputs | Blacks as Whites | Difference (2) - (3) | Difference (2) - (4) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| 50 Percent Release Rate | 8.35 | 8.38 | 8.40 | 8.42 | −0.02 | −0.04 |
| 70 Percent Release Rate | 10.13 | 10.20 | 10.23 | 10.29 | −0.03 | −0.09 |
| 90 Percent 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 blacks-as-whites 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 blacks-as-whites model. See the text for additional details on the specification and sample.

## VII. Extensions

In this section, we describe a number of potential extensions to our analysis and results. We begin by discussing the addition of protected characteristics such as gender. We then discuss non-linear prediction models, prediction models with non-race interactions, and prediction models with race interactions. We conclude by discussing how our proposed statistical models can be extended to other contexts.

### 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. 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* gender blinding-inputs algorithm could be estimated in a similar two-step procedure as described previously in Section V.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 females, using the two-step procedure described in Section V.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 non-white and non-female individuals.

### B. More Complicated Algorithms

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

*Non-Linear Prediction Models*: Our main proposed algorithms assume that there is a linear relationship between each predictive input and the outcome of interest, e.g. that 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 non-linearities in this relationship. A nice feature of both our proposed algorithms is that they can be easily adapted to allow for non-linearities. 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 non-linear model, such as a logit or probit model, and still be able to eliminate both direct and proxy effects of race.[179]

*Prediction Models with Non-Race Interactions*: Our main proposed algorithms also assume that there are no interaction effects between different non-race 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 five percentage point reduction in

[179] For a discussion of how to purge both proxy and direct effects of protected characteristics from a logit or probit model, see Pope and Sydnor (2011).

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. 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.

*Prediction Models with Race Interactions*: One might also want to allow interaction effects between a non-race predictive input and race itself. For example, one might want to generate predictions assuming that the relationship between age and risk differs by defendant race. If one were to fully interact each non-race input with race, the predictive algorithm would estimate separate risk predictions for white and black individuals. Such an approach is similar to that of Kleinberg et al. (2018) which fully utilizes the predictive power of all input factors, including protected characteristics. Under this approach, we would allow the coefficients, or "slopes," on the full set of input factors to differ by race. 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.

Here, however, we note that an important trade-off does exist between our ability to eliminate unwarranted racial disparities ("fairness") and predictive accuracy. In general, an unconstrained race-interacted model may result in either smaller or larger racial disparities. Racial disparities in predicted outcomes will, for example, increase if black defendants are statistically "riskier" than white defendants. In other words, an unconstrained race-interacted model may suffer from the same issue as the benchmark statistical model, where predictive accuracy is increased at the potential cost of racial fairness and legal permissibility.

This tension reflects a larger debate about the trade-offs between accuracy and fairness. 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...[because] the inclusion of such variables can increase both equity and efficiency."[180] In contrast, legal scholars have argued that 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."[181] Regardless of the trade-off that one decides to take, we do note that estimating a fully race-interacted model while reducing racial disparities as much as possible will often require some degree of ex post racial balancing. For example, Kleinberg et al. suggest that one could still achieve the desired racial composition by setting "a different threshold for the discriminated group,"[182] an approach that explicitly requires disparate treatment of individuals. In other words, 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."[183]

[180] *See* Kleinberg et al., *supra* note X, at 23.

[181] Barocas and Selbst, *supra* note X, at 721-722.

[182] Kleinberg et all, *supra* note X, at 24-25.

[183] Kleinberg et al, *supra* note X, at 23.

While we are in favor of this approach from a statistical perspective, we do have concerns about its legality given that it would require explicit racial balancing or fixing of a racial composition. Such an approach would likely run into a potential challenge given the Supreme Court's 2009 decision in *Ricci v. DeStefano*.[184] 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 eligible for promotion on the basis of the exam results. 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.[185] 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 of both Title VII and the Equal Protection Clause.[186] A five-person majority of the Court held that the City's race-based action violated the 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.[187] Thus, *Ricci* suggests that typically a decision-maker cannot engage in disparate treatment on the basis of a protected characteristic in order to avoid a disparate impact.

### C. Other Contexts

Our proposed statistical solutions can be easily applied to other contexts that face similar debates, including both 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 example, the Equal Credit Opportunity Act (ECOA) of 1974 prohibits discrimination on the basis of protected characteristics such as race, gender, or national origin.[188] 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.[189] 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.[190]

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."[191] These laws have also

[184] 557 U.S. 557 (2009).

[185] Id. at 562.

[186] Id. at 562-63.

[187] Id. at 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.

[188] See 15 U.S.C. §1691(a)(1) (2012).

[189] 12 C.F.R. §202.5 (2013).

[190] Id.

[191] Helen F. Ladd, *Evidence on Discrimination in Mortgage Lending*, 12 Journal of Economic Perspectives 41, 43 (1998).

been interpreted to prohibit the use of "redlining," or geographic discrimination using zip codes as proxies for the racial composition of neighborhoods.

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.[192] 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.[193] 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.[194] Some scholars have cautioned that lenders may even deliberately target certain racial or ethnic groups by using "facially-neutral proxy variables in its scoring models as stand-ins for characteristics like race."[195]

A legal excluding-inputs algorithm may prohibit credit scoring companies from using race and correlates of race from algorithms. But again, we note that this is likely to be impractical given that many, if not all, inputs are highly correlated with race. These non-race inputs are also likely to have substantial predictive power, even independent of their correlation within race.[196] 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 "to 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."[197]

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 prediction process as required under our proposals, there is no way of truly

[192] *See* Mikella Hurley and Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 Yale J. Law & Tech. 148, 182 (2016); *see also* Gillis and Spiess, *supra* note X, 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.").

[193] Hurley and Adebayo, *supra* note X, at 183.

[194] *See* Rob Berger, A Rare Glimpse Inside the FICO Credit Score Formula, DOUGHROLLER (Apr. 30, 2012),http://www.doughroller.net/credit/a-rare-glimpse-inside-the-fico-credit-score-formula. 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, https://www.myfico.com/credit-education/whats-not-in-your-credit-score/. 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* http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html. For example, machine-learning company ZestFinance uses public credit report data, but also proprietary and social network data to predict the likelihood that a borrower will repay their debts.

[195] Hurley and Adebayo, supra note X, at 191.

[196] 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), *available* at https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf (concluding that "[c]redit-based insurance scores appear to have little effect as a 'proxy' for membership in racial and ethnic groups in decisions related to insurance.)

[197] Gillis & Spiess, *supra* note X, at 469, 471.

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. 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 group is prohibited.[198] 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.[199] 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.[200]

Although many legal scholars have questioned whether Title VII is sufficient for dealing with the types of issues introduced by the use of algorithms,[201] 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."[202] Similarly, Barocas and 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 "under 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."[203]

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.[204] As Pauline 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 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,

[198] *See International Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977).

[199] *See Griggs v. Duke Power*, 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, 105 Stat. 1071, was enacted, which included a provision codifying the prohibition on disparate-impact discrimination.

[200] See 42 U.S.C. §2000e-2(K)(1)(A)(ii).

[201] *See, e.g.*, Barocas & Selbst, *supra* note 5, at 694 (concluding that Title VII is "not well equipped" to address data mining).

[202] Kim, supra note X, at 918.

[203] Barocas & Selbst, *supra* note X, at 695 and 719.

[204] Kim, supra note X, at 918.

> simply eliminating data on those characteristics cannot prevent models that are biased along these dimensions.

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."[205] 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."[206]

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 anti-discrimination law, our solutions would allow employers to retain some predictive power while eliminating direct and proxy effects from predictions. Our proposals therefore have the potential to reduce race or gender disparities in employment and hiring.

## VIII. Conclusion

In this paper, 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 towards adopting a formalistic view of algorithms that requires the exclusion of race and all non-race correlates in predictive algorithms, as nearly all potential algorithmic inputs are likely to be correlated with race. Our skepticism is supported in 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.

Our paper 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 blacks relative to whites 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.

We view our findings as requiring 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 interpre-

[205] Kim, *supra* note X, at 918.

[206] Barocas & Selbst, *supra* note X, at 692.

tation 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.

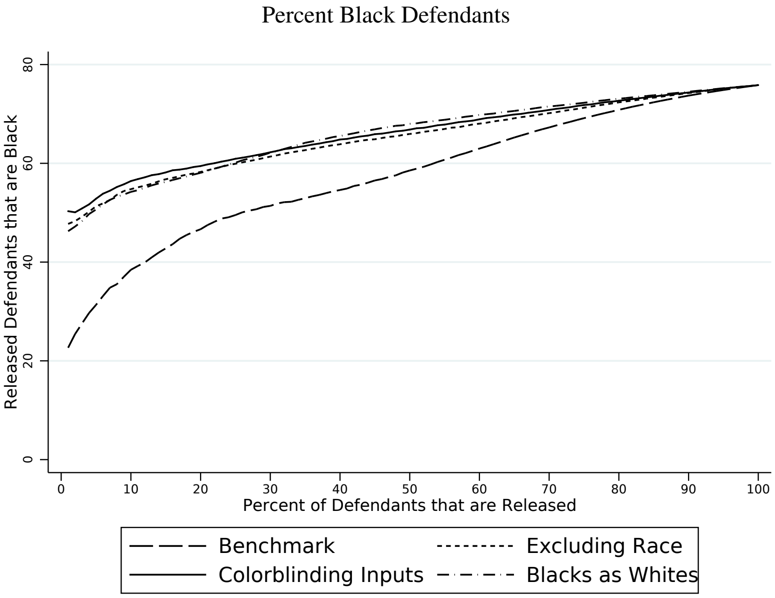
# A. Appendix

Table A1: Simulations of Racial Disparities Under Different Predictive Algorithms

| | Share of Blacks Released | | | | | |
|---|---|---|---|---|---|---|
| | Benchmark Model | Excluding Race | Colorblinding Inputs | Blacks as Whites | Difference (2) - (3) | Difference (2) - (4) |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| 50 Percent Release Rate | 58.55 | 65.95 | 66.83 | 68.00 | -0.88 | -2.05 |
| 70 Percent Release Rate | 67.26 | 70.16 | 70.82 | 71.51 | -0.66 | -1.35 |
| 90 Percent 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 blacks-as-whites 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 blacks-as-whites model. See the text for additional details on the specification and sample.

Figure A1: 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 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.