School of Law
UNIVERSITY OF GEORGIA

UNIVERSITY OF GEORGIA
SCHOOL OF LAW

RESEARCH PAPER SERIES

Paper No. 2018-35 September 2018

BIAS IN, BIAS OUT

128 YALE L.J. __ (FORTHCOMING 2019).

SANDRA G. MAYSON
Assistant Professor of Law
University of Georgia School of Law
smayson@uga.edu

BIAS IN, BIAS OUT

Sandra G. Mayson[*]

ABSTRACT

Police, prosecutors, judges, and other criminal justice actors increasingly use algorithmic risk assessment to estimate the likelihood that a person will commit future crime. As many scholars have noted, these algorithms tend to have disparate racial impact. In response, critics advocate three strategies of resistance: (1) the exclusion of input factors that correlate closely with race, (2) adjustments to algorithmic design to equalize predictions across racial lines, and (3) rejection of algorithmic methods altogether.

This Article's central claim is that these strategies are at best superficial and at worst counterproductive, because the source of racial inequality in risk assessment lies neither in the input data, nor in a particular algorithm, nor in algorithmic methodology. The deep problem is the nature of prediction itself. All prediction looks to the past to make guesses about future events. In a racially stratified world, any method of prediction will project the inequalities of the past into the future. This is as true of the subjective prediction that has long pervaded criminal justice as of the algorithmic tools now replacing it. What algorithmic risk assessment has done is reveal the inequality inherent in all prediction, forcing us to confront a much larger problem than the challenges of a new technology. Algorithms shed new light on an old problem.

Ultimately, the Article contends, redressing racial disparity in prediction will require more fundamental changes in the way the criminal justice system conceives of and responds to risk. The Article argues that criminal law and policy should, first, more clearly delineate the risks that matter, and, second, acknowledge that some kinds of risk may be beyond our ability to measure without racial distortion—in which case they cannot justify state coercion. To the extent that we can reliably assess risk, on the other hand, criminal system actors should strive to respond to risk with support rather than restraint whenever possible. Counterintuitively, algorithmic risk assessment could be a valuable tool in a system that targets the risky for support.

* Assistant Professor of Law, University of Georgia School of Law. I am grateful for extremely helpful input from David Ball, Mehrsa Baradaran, Solon Barocas, Stephanie Bornstein, Kiel Brennan-Marquez, Bennett Capers, Nathan Chapman, Andrea Dennis, Sue Ferrere, Sean Hill, Mark Houldin, Gerry Leonard, Kay Levine, Anna Roberts, Hannah Sassaman, Andrew Selbst, Tim Schnacke, Megan Stevenson, and Stephanie Wykstra, as well as for thoughtful comments from fellow participants in the 2017 Southeastern Junior / Senior Faculty Workshop, CrimFest 2017 & 2018, and the 2017 and 2018 UGA / Emory Faculty Workshops.

TABLE OF CONTENTS

INTRODUCTION

"There's software across the country used to predict future crime. And it's biased against blacks."[1] So proclaimed an exposé by news outlet ProPublica in the summer of 2016. The story focused on a particular algorithmic tool, the COMPAS, but its ambition, and effect, was to stir alarm about the ascendance of algorithmic crime prediction overall.

The ProPublica story, *Machine Bias,* was emblematic of broader trends. The age of algorithms is upon us. Automated prediction programs now make decisions that affect every aspect of our lives. Soon they will drive our cars, but in the meantime they shape advertising, credit lending, hiring, policing – just about any governmental or commercial activity that has some predictive component. There is reason for this shift. Algorithmic prediction is profoundly more efficient, and often more accurate, than human judgment. It eliminates the irrational biases that contort so much of our decision-making. On the other hand, it has become abundantly clear that machines can discriminate.[2] Algorithmic prediction has the potential to perpetuate or amplify social inequality, all while maintaining the veneer of high-tech objectivity.

Nowhere is the concern with algorithmic bias more acute than in criminal justice. Over the last five years, criminal justice risk assessment has been spreading rapidly. In this context, "risk assessment" is shorthand for the actuarial measurement of some defined risk, usually the risk that the person assessed will commit future crime.[3] The concern with future crime is not new; police, judges, prosecutors, and probation and parole officers have long been tasked with making subjective determinations of dangerousness. The shift is from subjective to actuarial assessment.[4] With the rise of big data and bipartisan ambitions to be smart on crime, algorithmic risk assessment has taken the criminal justice system by storm. It is the lynchpin of the bail reform

[1] Julia Angwin *et al.*, *Machine Bias*, PROPUBLICA.COM (May 23, 2016), www.propublica.org /article/machine-bias-risk-assessments-in-criminal-sentencing.

[2] *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); SAFIYA NOBLE, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017).

[3] Most risk assessment tools, however, do not actually measure the likelihood of future crime commission, but instead the likelihood of future *arrest*, which is a poor proxy. *See infra* Part II.B.1.

[4] Parole boards have used risk assessment instruments since the 1920s, *see* BERNARD HARCOURT, AGAINST PREDICTION 7-18 (2007), but actuarial tools were hardly known in other parts of the criminal justice system until the last few years.

movement,[5] the cutting edge of policing,[6] and increasingly used in charging,[7] sentencing,[8] and to allocate supervision resources.[9] This development has sparked profound concern about the racial impact of risk assessment.[10] Given that algorithmic crime prediction tends to rely on factors heavily correlated with race, it appears poised to entrench the inexcusable racial disparity so characteristic of our justice system, and to dignify the cultural trope of black criminality with the gloss of science.

Thankfully, we have reached a moment in which the prospect of exacerbating racial disparity in criminal justice is widely understood to be unacceptable. And so, in this context as elsewhere, the prospect of algorithmic discrimination has generated calls for interventions to the predictive process to ensure racial equity. This raises the difficult question of what equality looks like. The challenge is that there are many possible metrics of racial equity in statistical prediction, and some of them are mutually exclusive.[11] The law provides no useful guidance about which to prioritize.[12] In the void it leaves, data scientists are exploring different statistical measures of equality and technical methods to achieve them.[13] Legal scholars have begun to weigh in.[14] Beyond the ivory tower, this debate is happening in courts,[15] city council chambers,[16] and community meetings.[17] The stakes are real. Criminal justice institutions must decide whether to adopt risk

[5] *See, e.g.*, Sheila Dewan, *Judges Replacing Conjecture With Formula for Bail*, N.Y. TIMES (July 26, 2015); Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490 (2018); Megan T. Stevenson, *Assessing Risk Assessment in Action*, __ MINN. L. REV. (forthcoming 2018).

[6] *See, e.g.*, Selbst, *supra* note 2 at 113 ("…[P]redictive policing [is] a popular and growing method for police departments to prevent or solve crimes."); Letter from Jonathan Wroblewski, Director of the Office of Policy Legislation to Hon. Patti Saris, Chair of the U.S. Sentencing Comm'n 2 (July 29, 2014) [hereinafter DOJ Letter to U.S.S.C] (noting that "Predictive Policing—the use of algorithms that combine historical and up-to-the-minute crime information—is spreading").

[7] *See, e.g.*, Andrew Guthrie Ferguson, *Predictive Prosecution*, 51 WAKE FOREST L. REV. 705 (2016) (explaining "predictive prosecution" and exploring its "promise and perils").

[8] *See, e.g.*, Erin Collins, *Punishing Risk*, __ GEO. L. J. __ (forthcoming 2019); Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583 (2018).

[9] Issue Brief, Pew Ctr. on the States, Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders 2 (2011), www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2011/PewRiskAssessmentbriefpdf.pdf (describing growing use of risk assessment to allocate supervision resources).

[10] *See infra* Part I.A.

[11] *See infra* Part I.B.

[12] Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT. R. 237, 237 (2015); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, __DUKE L.J.__ (forthcoming 2019).

[13] *See infra* Part I.C.

[14] Huq, *supra* note 13.

[15] *E.g.* State v. Loomis, 881 N.W.2d 749 (Wis. 2016).

[16] *E.g.* Philadelphia City Council Special Committee on Criminal Justice Reform, Interim Report Fall 2016: A Shift from Re-Entry to Pre-Entry 12, http://phlcouncil.com/wp-content/uploads/2016/11/SCFall2016InterimReport.pdf ("During prior public hearings, members of the Special Committee raised concerns that the data used in a risk assessment tool's calculations may be inherently biased, because of the decades of disparate impact and racial imbalance within the criminal justice system.").

[17] *E.g.*, Chris Palmer & Claudia Irizarry-Aponte, *Dozens of Speakers at Hearing Assail Pa. Plan to Use Algorithm in Sentencing*, PHILLY.COM (June 6, 2018), http://www.philly.com/philly/news/crime/philadelphia-pennsylvania-algorithm-sentencing-public-hearing-20180606.html.

assessment tools and if so, what measure of equality to demand that those tools fulfill. They are making these decisions as I write.[18]

Among racial justice advocates engaged in the debate, a few common themes have emerged.[19] The first is a demand that race, and factors that correlate heavily with race, be excluded as input variables for prediction. The second is a call for "algorithmic affirmative action" to equalize adverse predictions across racial lines. To the extent that scholars have grappled with the necessity of prioritizing a particular equality measure, they have mostly urged stakeholders to demand equality in the false-positive and false-negative rates for each racial group, or in the overall rate of adverse predictions across groups ("statistical parity"). Aziz Huq offers a more abstract prescription, proposing that we should design each algorithm to ensure that it imposes no net burden on communities of color, which might require some algorithms to set different thresholds for risk classes by race.[20] Lastly, critics argue that, if algorithmic risk assessment cannot be made meaningfully race-neutral, the criminal justice system must reject algorithmic methods altogether.

This Article contends that these demands are at best superficial and at worst counter-productive, because they ignore the real source of the problem: the nature of prediction itself. All prediction functions like a mirror. Its premise is that we can learn from the past because, absent intervention, the future will repeat it. Individual traits that correlated with crime commission in the past will correlate with crime commission in future. So what any predictive analysis does is hold a mirror to the past. It distills patterns in past data and interprets them as projections about the future. Algorithmic prediction produces a precise reflection of digital data. Subjective prediction produces a cloudy reflection of anecdotal data. But the nature of the analysis is the same. To predict the future under status quo conditions is simply to project history forward.

Given the nature of prediction, a racially unequal past will necessarily produce racially unequal outputs. To adapt a computer science idiom, "bias in, bias out."[21] Specifically: If the thing that we undertake to predict—say arrest—happened more frequently to black people than white in the past data, a predictive analysis will project it more frequently for black people than white in the future. The predicted event, called the target variable, is thus the key to racial disparity in prediction.

The strategies for racial equity that currently dominate the conversation amount to distorting the predictive mirror or tossing it out. Consider input data. If the thing we have undertaken to predict happens more frequently to people of color, an accurate algorithm will predict it more

[18] *Id.*; *see also* Phase 1 Reports, Pennsylvania Commission on Sentencing, Risk Assessment, http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment (last visited June 30, 2018) (collecting information relating to Commission's project to develop risk assessment tool with public input).

[19] *See infra* Part III.

[20] Huq, *supra* note 13.

[21] The computer science idiom is "garbage in, garbage out," which refers to the fact that algorithmic prediction is only as good as the data on which the algorithm is trained.

frequently for people of color. Limiting input data cannot eliminate the disparity except by crippling the predictive tool. The same is true of algorithmic affirmative action to equalize outputs. Some calls for such interventions are motivated by the well-founded belief that, because of racially disparate law enforcement patterns, the standard target variable, arrest, embeds racial distortion vis-à-vis the event we actually want to avoid, presumably serious crime. But unless we know actual offending rates (which we generally do not), reconfiguring the data or algorithm to reflect a statistical scenario we prefer merely distorts the predictive mirror, so it neither reflects the data nor any demonstrable reality. Along similar lines, calls to equalize adverse predictions across racial lines require an algorithm to forsake the statistical risk assessment of individuals in favor of risk sorting within racial groups. And wholesale rejection of algorithmic methods rejects the predictive mirror directly.

The Article's normative claim is that neither distorting the predictive mirror nor tossing it out is the right path forward. If the image in the predictive mirror is jarring, the answer is not to bend it to our liking. That does not solve the problem. Nor does rejecting algorithmic methods, because there is every reason to expect that subjective prediction entails an equal degree of racial inequality. To reject algorithms in favor of judicial risk assessment is to discard the precise mirror for the cloudy one. It does not eliminate disparity. It merely turns a blind eye.

What actuarial risk assessment has done, in other words, is reveal the racial inequality inherent in *all* prediction in a racially unequal world, forcing us to confront a much deeper problem than the dangers of a new technology. In making the mechanics of prediction transparent, algorithmic methods have exposed the disparities endemic to all criminal justice risk assessment, subjective and actuarial alike. Tweaking an algorithm or its input data, or even rejecting actuarial methods, will not redress the racial disparities in crime- or arrest-risk in a racially stratified world.

The inequality exposed by algorithmic risk assessment should instead galvanize a more fundamental rethinking of the way in which the criminal justice system understands and responds to risk.[22] To start, we should be more thoughtful about what we want to learn from the past, and more honest about what we can. If the risk that really matters is the risk of serious crime but we have no access to data that fairly represent the incidence of it, there is no basis for predicting that event at all. Nor is it acceptable to resort to predicting some other event, like "any arrest," that happens to be easier to measure. This lesson has profound implications for all forms of criminal justice risk assessment, both actuarial and subjective.

If the data do fairly represent the incidence of serious crime, on the other hand, the place to redress racial disparity is not in the measurement of risk, but in the response to it. Risk assessment must reflect the past; it need not dictate the future. The default response to risk could be supportive rather than coercive. In the long term, a supportive response to risk would help to

[22] *See infra* Part IV.

redress the conditions that produce risk in the first place. In the short term, it would mitigate the disparate racial impact of prediction. Counterintuitively, algorithmic assessment could play a valuable role in a system that targets the risky for support rather than restraint.

The Article makes three core contributions. The first is explanatory. Thus far, the computer science and statistical literature on algorithmic fairness and the legal literature on criminal justice risk assessment have largely evolved on separate tracks.[23] Part I offers the most comprehensive and accessible taxonomy to date of potential measures of equality in prediction, synthesizing recent work in computer science with legal equality constructs. The Article's second contribution is the descriptive analysis of practical and conceptual problems with strategies to redress predictive inequality that are aimed at algorithmic methods *per se,* given that all prediction replicates the past. The Article's third contribution is the normative argument that meaningful change will require a more fundamental rethinking of the role of risk in criminal justice.

This Article is about criminal justice risk assessment, but it is also a window onto the broader conversation about algorithmic fairness, which is itself a microcosm of perennial debates about the nature of equality. Through a focused case study, the Article aims to contribute to the larger literatures on algorithmic fairness and on competing conceptions of equality in law. The Article's conclusion draws out some of the larger connections.

Two caveats are in order. First, the article focuses on racial disparity in prediction, severed from the messy realities of implementation. Megan Stevenson has shown that the vagaries of implementation may affect the treatment of justice-involved people more than a risk assessment algorithm itself.[24] Still, risk assessment tools are meant to guide decision-making. To the extent they do, disparities in classification will translate into disparities in outcomes. For that reason and for purposes of clarity, this Article focuses on disparities in classification alone. The second caveat is that this Article speaks of race in the crass terminology of "black" and "white." This language reduces a deeply fraught and complex social phenomenon to an artificial binary. The Article uses this language in part of necessity, to explain competing metrics of equality with as much clarity as possible, and in part in recognition that the criminal justice system itself tends to deploy this reductive schema. Whether the Article is warranted in taking this approach, the reader may judge.

The Article proceeds in four parts. Part I chronicles the recent scholarly and public debate over risk assessment and racial inequality, using the ProPublica saga and a stylized example to illustrate why race-neutral prediction is impossible. It concludes with a comprehensive taxonomy of the most important potential metrics of predictive equality. Part II lays out the

[23] A handful of seminal articles, however, have helped to bridge the gap. *See generally* Selbst & Barocas, *supra* note 2; Selbst, *supra* note 2; Huq, *supra* note 13; Kroll *et al*, *Accountable Algorithms*, 165 UNIV. PA. L. REV. 633 (2017).

[24] Stevenson, *supra* note 5.

Article's central conception of prediction as a mirror. For clarity of analysis, it draws an important distinction between two possible sources of racial disparity in prediction: racial distortions in the data vis-à-vis underlying crime rates, and a difference in underlying crime rates by race. Accounting for both, Part III explains why the prescriptions for racial equity that currently dominate the public and scholarly debate will not solve the problem. Part IV argues for a broader rethinking of the role of risk in criminal justice. The Conclusion draws out implications for other predictive arenas.

I. THE IMPOSSIBILITY OF RACE-NEUTRALITY

A. The Risk Assessment-and-Race Debate

Just a few years ago criminal justice risk assessment was an esoteric topic. Today it is fodder for The Daily Show,[25] of interest to major mainstream media,[26] and the subject of a vibrant and growing body of scholarship.[27] That literature offers an introduction to risk assessment that need not be repeated here. But it is important to define some key terms. As used in this Article, "criminal justice risk assessment" refers to the actuarial assessment of the likelihood of some future event, usually arrest for crime. The term encompasses two kinds of risk assessment tools: the more basic and more prevalent checklist instruments, and the more sophisticated machine-learned algorithms that represent the future.[28]

As the use of criminal justice risk assessment has spread, concern over its potential racial impact has exploded. The watershed year was 2014. A journalist asked whether Chicago's new predictive policing strategy was "racist,"[29] legal scholar Sonja Starr argued that the Constitution prohibits the use of race, gender, or income-correlated variables in risk assessment tools

[25] *Disrupting the Legal System with Robots*, THE DAILY SHOW (March 7, 2018), https://youtu.be/VkizYljxcD8.

[26] *E.g.* Angwin *et al*., *supra* note 1; Anna Maria Barry-Jester *et al*., *Should Prison Sentences Be Based on Crimes That Haven't Been Committed Yet?*, FIVETHIRTYEIGHT (Aug. 4, 2015), fivethirtyeight.com/features/prison-reform-risk-assessment (including simulations demonstrating risk assessment outcomes and disparate racial impact); Dewan, *supra* note 5.

[27] *See, e.g.*, Collins, *supra* note 8; Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017); Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231 (2015); Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT. R. 237 (2015); Huq, *supra* note 13; John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, __ WASH. L. REV. __ (forthcoming 2018); Mayson, *supra* note 5; Anne Milgram et al., *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making*, 27 FED. SENT. R. 216 (2015); Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671 (2015); Slobogin, *supra* note 8; Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014); Stevenson, *supra* note 5.

[28] For a brief explanation of the difference, *see* Mayson, *supra* note 5, at 509-11, n.97; *see also, generally*, Richard Berk and Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT. R. 222 (2015).

[29] Matt Stroud, *The Minority Report: Chicago's New Police Computer Predicts Crimes, But Is It Racist?*, THEVERGE (Feb. 19, 2014), https://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist.

used at sentencing,[30] and the Department of Justice flagged both "the promise and danger of data analytics in sentencing and corrections policy."[31] Then-Attorney General Eric Holder warned that risk assessment tools might "exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.[32] The following year, Bernard Harcourt expanded on the problem.[33] The nation's long history of social and economic oppression of African Americans—including criminal laws and law enforcement that targeted black men—has produced higher rates of arrest, prosecution, conviction and incarceration among black Americans than white. The result is that criminal history now correlates with race.[34] Any form of risk assessment that relies on criminal history will therefore have a disparate impact on black communities, and on black men in particular.[35] Media, advocacy organizations, and other scholars echoed the concern.[36] In 2016, the ProPublica exposé supercharged the debate.[37]

Communities, scholars and policymakers are now highly focused on the potential racial effects of criminal justice risk assessment. Grassroots advocacy groups have launched campaigns to demand racial equality as new risk assessment tools are implemented, including a major national campaign urging jurisdictions to reject such tools altogether in the pretrial context.[38]

---

[30] Starr, *supra* note 28. She also noted that the use of such instruments "is likely to further concentrate mass incarceration's racial impact," because many factors included in the tools correlate with race. *Id.* at 838; *see also* Sonja B. Starr, *Sentencing, by the Numbers*, N.Y. TIMES (Aug. 10, 2014).

[31] DOJ Letter to U.S.S.C., *supra* note 6, at 1-8 (cautioning that "the use of risk assessment at sentencing raises constitutional questions because of the use of group-based classifications and suspect characteristics in the analytics").

[32] Eric Holder, United States Att'y Gen., Remarks at the Nat'l Ass'n of Crim. Defense Law. 57th Ann. Meeting, (Aug. 1, 2014), www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th.

[33] Harcourt, *supra* note 28, at 237 (arguing that heavy reliance on criminal history information for purposes of risk assessment "will unquestionably aggravate the already intolerable racial imbalance in our prison populations").

[34] *See, e.g.*, Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race and Recividism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 683-84; 704-06 (2016) (concluding that criminal history correlates with race in their dataset); Frank McIntyre & Shima Baradaran, *Race, Prediction, and Pretrial Detention*, 10 J. EMPIRICAL LEGAL STUD. 741, 759 (2013).

[35] Harcourt, *supra* note 28, at 240 ("[T]he continuously increasing racial disproportionality in the prison population necessarily entails that the prediction instruments, focused as they are on prior criminality, are going to hit hardest the African American communities.").

[36] Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 260 (2015) (discussing challenges including constitutional doctrine relating to racial classifications); Melissa Hamilton, *Back to the Future: The Influence of Criminal History on Risk Assessments*, 20 BERKELEY J. CRIM. L. 75, 78 (2015) (exploring concerns with the use of criminal history in risk assessment, including "the potential that criminal history is an unfortunate proxy for race and social disadvantage"); Anna Maria Barry-Jester *et al.*, *The New Science of Sentencing*, MARSHALL PROJECT (Aug. 4, 2015), www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing; Barry-Jester *et al.*, *supra* note 27 (including simulations demonstrating risk assessment outcomes and disparate racial impact); Anna Orso, *Can Philly's New Technology Predict Recidivism Without Being Racist?*, BILLYPENN (Sept. 25, 2017), https://billypenn.com/2017/09/25/can-phillys-new-technology-predict-recidivism-without-being-racist; *Risk Assessment or Race Assessment?*, SENTENCING PROJECT (July 23, 2015), http://www.sentencingproject.org/news/race-justice-news-risk-assessment-or-race-assessment.

[37] Angwin *et al.*, *supra* note 1.

[38] In August of 2018, the Leadership Conference on Civil and Human Rights and 115 other advocacy groups released The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of

Legal scholars[39] and policy organizations[40] are increasingly attentive to the possible racial impact of criminal justice risk assessment, as are computer scientists and econometricians who write about criminal justice.[41] Aziz Huq has laid out the history of racial oppression in criminal justice that makes the concern so acute, as well as the inadequacy of current legal doctrine to address it.[42]

Notwithstanding the growing interest, the debate remains hampered by ambiguous terms.[43] To some people, to say that a decision procedure is "biased" is to say that it is statistically unsound.[44] A risk assessment algorithm is racially biased in this sense if it systematically over- or understates the average risk of one racial group relative to another.[45] Others, however, view a judgment procedure as "biased" if it produces differential effects across racial groups that present a moral concern, even if the judgments themselves are not systematically less accurate for one group than the other.[46] "Discrimination" also carries ambiguity; it can mean any "act of

---

Civil Rights Concerns, available at https://leadershipconferenceedfund.org/pretrial-risk-assessment. *See also, e.g.*, Predictive Policing, Media Mobilizing Project, https://mediamobilizing.org/predictive-policing.

[39] *See, e.g.*, Megan T. Stevenson & Sandra G. Mayson, *Bail and Pretrial Detention*, *in* REFORMING CRIMINAL JUSTICE: A REPORT OF THE ACADEMY FOR JUSTICE, BRIDGING THE GAP BETWEEN SCHOLARSHIP AND REFORM, 34-39 (2017), http://academyforjustice.org/volume3/; Anupam Chander, *The Racist Algorithm*, 115 MICH. L. REV. 1023 (2017); Eaglin, *supra* note 28, at 94-99 (discussing how risk assessment might "compromise[e] equality"); Mayson, *supra* note 5, at 494-96; Selbst, *supra* note 2; *see also* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 863–64 (2017) (exploring racial effects of algorithmic prediction in employment context).

[40] *See, e.g.*, Harvard Law School Criminal Justice Policy Program, MOVING BEYOND MONEY: A PRIMER ON BAIL REFORM (2016), cjpp.law.harvard.edu/assets/FINAL-Primer-on-Bail-Reform.pdf.

[41] *See, e.g.*, Jon Kleinberg *et al.*, *Human Decisions and Machine Predictions* (Nat'l Bureau of Econ. Res., Working Paper No. 23180, Feb. 2017), http://www.nber.org/papers/w23180; Stevenson, *supra* note 5.

[42] Huq, *supra* note 13.

[43] *Accord* Selbst, *supra* note 2, at 123 (noting that "[t]he words 'discrimination,' 'fairness', and 'bias' evoke a family of related concepts").

[44] In econometrics, "bias" describes any systematic deviation of a statistical calculation from the true value of the thing calculated. *E.g.* BRUCE E. HANSEN, ECONOMETRICS § 4.2 (digital ed. 2017), www.ssc.wisc.edu/~bhansen/econometrics/ EconometricsKindle.pdf ("An estimator [calculation technique] with the property that its expectation [the average of the values it produces over many iterations] equals the parameter it is estimating [true value of the thing it is estimating] is called unbiased."); *see also Bias*, MERRIAM-WEBSTER ONLINE ("d(1): deviation of the expected value of a statistical estimate from the quantity it estimates; (2) systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others").

[45] William Dieterich *et al.*, COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 1, 2-3, 8-13 (Technical Report, Northpointe Inc., July 2016), university.pretrial.org /viewdocument/compas-risk-scales-demonstrating-a (asserting that a predictive instrument is biased only if a given score, or classification, means a different likelihood of the predicted outcome for members of one racial group than members of the other); *see also* Anthony W. Flores *et al.*, *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks."*, 80 FED. PROBATION 38, 38, 40 (2016) (arguing that "well-established and accepted standards exist to test for bias in risk assessment"); *see also* Skeem & Lowenkamp, *supra* note 34, 685 (asserting that if "a given score [has] the same meaning regardless of group membership," the instrument is "unbiased").

[46] *E.g.* Kim, *supra* note 40, at 866 ("Classification bias occurs when employers rely on classification schemes, such as data algorithms, to sort or score workers in ways that worsen inequality or disadvantage along the lines of race, sex, or other protected characteristics."). These two uses of the word "bias" correspond to the notions of irrational versus rational (or statistical) discrimination. Deborah Hellman, *What Makes Genetic Discrimination Exceptional?*, 29 AM. J.L. & MED. 77, 83-86

making or perceiving a difference" or only an *unjustified* act of making or perceiving a difference.[47] Along similar lines, Jennifer Skeem and Christopher Lowenkamp have contested Bernard Harcourt's claim that criminal history serves as a "proxy" for race in risk assessment—but in fact they just define "proxy" differently than he does.[48] These ambiguous terms can obscure the questions at stake, which are already complex enough.

B. The Problem of Equality Tradeoffs

The central complication is that there is no single measure of racial equality in risk assessment. There are many possible measures. In most circumstances, moreover, it is impossible to achieve racial equality by every measure at once.

The ProPublica saga illustrates the problem. ProPublica concluded that the COMPAS was "biased against blacks" by analyzing data from a county where the COMPAS was used to assess the likelihood that a pretrial defendant would be rearrested if s/he remained at liberty.[49] The ProPublica researchers compared the COMPAS' risk classifications with defendants' actual outcomes—whether each defendant was rearrested or not. The company that owns the COMPAS, Northpointe, responded with indignation: ProPublica's own data showed that the COMPAS was demonstrably race-neutral![50]

The fascinating thing was that both ProPublica and Northpointe were right; they were just emphasizing different metrics of equality. The fact that led Northpointe to claim race-neutrality was that black and white defendants classified as high-risk by the COMPAS were in fact rearrested at equal rates. A high-risk classification meant the same chance of rearrest for a black

(2003); Jeffrey S. Morrow, *Insuring Fairness: The Popular Creation of Genetic Antidiscrimination*, 98 GEO. L.J. 215, 230-32 (2009); Anya E.R. Prince, *Insurance Risk Classification in an Era of Genomics: Is A Rational Discrimination Policy Rational?*, 96 NEB. L. REV. 624, 630-34; 641-42 (2018). Frederick Schauer offers a similar analysis of the ambiguity of the terms "stereotype" and "prejudice." FREDERICK SCHAUER, PROFILES, PROBABILITIES AND STEREOTYPES 7, 13-17 (2003) (noting that these terms may refer to a generalization that is irrelevant or statistically unsound or to a generalization that is both relevant and statistically sound but deployed in a morally objectionable way).

[47] *Discrimination*, MERRIAM-WEBSTER ONLINE (definitions 1(a)) and 3(b)). Note the "discrimination" can also be used in a technical legal sense, to mean only such differential treatment or impact as would incur liability pursuant to anti-discrimination law.

[48] Skeem & Lowenkamp, *supra* note 34, at 698-700 (assessing whether criminal history functioned as a proxy for race in the federal Post Conviction Risk Assessment tool (PCRA) and concluding that it did not). Skeem and Lowenkamp define a "proxy" to mean a variable that merely stands in for another and has no independent predictive value; in this sense, criminal history is not a proxy for race. Even after subtracting the predictive value of race from the predictive value of criminal history, as it were, criminal history retains additional—independent—predictive value. (It is unclear from their analysis whether they find criminal history to function as a mediator or a moderator of race for purposes of the PCRA, but the analysis better supports the latter conclusion.) Harcourt calls criminal history a "proxy" for race in the more modest sense that it correlates with race (even if it also has independent predictive value), such that relying on it will have disparate impact across racial lines.

[49] Angwin *et al.*, *supra* note 1.

[50] Dieterich *et al.*, *supra* note 45; *see also* Flores *et al.*, *supra* note 45 (reporting results of independent study of same data and concluding that COMPAS was equally predictive for white and black defendants).

defendant as for a white (approximately 60% on the any-arrest-risk scale and 20% on the violent-arrest-risk scale, over a period of two years).[51] This metric of equality is sometimes called "predictive parity." The fact that led ProPublica to claim racial bias was something more subtle: A black defendant who would *never* be rearrested was much more likely to be classified as high-risk (45%) than a white defendant would never be rearrested (23.5%).[52] In statistical terms, the false-positive rate was much higher for the black defendants than the white.[53] Meanwhile, a white defendant who ultimately *would* be rearrested was more likely to be deemed low-risk (48%) than a black defendant who ultimately would be arrested (28%).[54] The false-negative rate was much greater for white defendants than black. ProPublica saw these racial differences in the COMPAS's error rates as a serious injustice.

The racial disparity in error rates was not, however, the result of nefarious distortion in the COMPAS algorithm itself.[55] It was a mathematical result of the divergent rates of arrest between the black and white defendants in the underlying dataset. Because the rate of arrest was higher among the black defendants, the black defendants, on average, had higher arrest-risk profiles. When the average risk is higher for one group than another, a greater proportion of that group will be predicted to be arrested, and a greater proportion of the group will also be *mistakenly* predicted to be arrested. This is true no matter how carefully designed the algorithm, so long as it is also striving to have equal predictive accuracy for each racial group.

To see this aspect of prediction more clearly, consider a stylized hypothetical. The image below depicts two groups of ten arrestees each—gray and black—who are subject to risk assessment (Figure 1). Say that the algorithm in question predicts rearrest within a year. For clarity, presume that it makes binary decisions: For each person, it predicts either rearrest or no rearrest. An arrest prediction is a "positive." If it is correct, it is a "true positive," and if it is incorrect it is a "false positive." A no-arrest prediction is a "negative." The shadowed figures represent the people who will ultimately be arrested. Note that the groups have different base rates of arrest: A greater proportion of the gray group will actually be rearrested (2/10) than the black (1/10). The boxes, finally, represent the algorithm. It predicts arrest for the people within the boxes.

[51] These rates were calculated on the basis of outcomes over a two-year period. Dieterich *et al.*, *supra* note 45, at 12. If anything, the rate of rearrest was higher for black defendants in each risk category. In other words, the risk classifications were more "generous" to black defendants than white. *See* Flores *et al.*, *supra* note 45, at 41-42; *id.* at 43 ("A given COMPAS score translates into roughly the same likelihood of recidivism, whether a defendant is Black or White.").

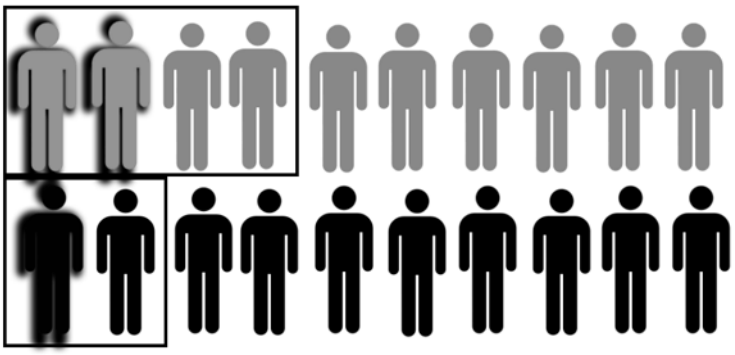
[52] Angwin *et al.*, *supra* note 1.

[53] Whether or not the statistical concepts of "false positives" and "false negatives" are applicable in the context of risk assessment is debatable, and discussed below. *See infra* Part III.C.2.

[54] Angwin *et al.*, *supra* note 1.

[55] Again, there is controversy in the literature over whether the language of "prediction" and "error rates" is appropriate to the risk assessment context. The debate is discussed more fully below. *See infra* Part III.C.2.

Figure 1: Groups with Different Base Rates of Arrest; Predictive Parity



This algorithm produces forecasts that are equal across the two groups in one sense: A positive forecast is equally accurate for each group. For both the black and gray groups, 50% of those forecast for rearrest (figures in the boxes) are indeed rearrested (shadowed figures). When the algorithm is deployed prospectively, a positive prediction for any individual will mean a 50% chance or rearrest, regardless of whether the person is gray or black. This is to say that the algorithm achieves "predictive parity," the equality metric that Northpointe emphasized.

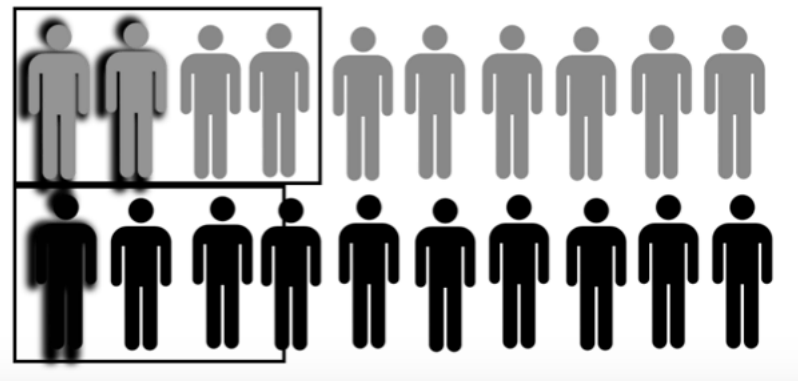
On the other hand, the algorithm produces unequal results in other ways. Consider the rate of false predictions among those who will *not* be arrested—the false-positive rate. Of the eight gray people who will not be arrested (the un-shadowed gray figures), two are mistakenly forecast for arrest. Of the nine black people will not be rearrested (the un-shadowed black figures), only one is mistakenly forecast for arrest. The false-positive rate is much higher for the gray group (25%) than for the black (11%). This is the form of inequality that ProPublica discovered in the COMPAS data. And as in the ProPublica study, this algorithm produces unequal results in another sense as well: Twice as many gray people as black are forecast for rearrest. The algorithm has a much greater overall impact on the group with the higher base rate. In the terminology favored by data scientists, the tool does not achieve "statistical parity." The table below records these three metrics.

	Gray	Black	
% of Rearrest Forecasts That Are Correct	50	50	Predictive Parity
% of No-Arrests Falsely Forecast for Arrest	25	11	Disparate False-Positive Rates
% of Group Forecast for Rearrest	40	20	Statistical Disparity

It is possible to modify the algorithm to equalize the false-positive rates for the two groups, but at a cost. The figure below represents one possible modification. For both the black and the gray groups, now, 25% of

the *non-arrestees* (un-shadowed figures) are mistakenly forecast for arrest. That is, the false-positive rate is 25% for each color group. The total number of people forecast for arrest is also much closer across groups. But notice the effect on the accuracy of the arrest forecasts themselves (the boxes). For the gray group, a prediction of arrest is still 50% likely to be true. But it is only about 30% likely to be true for a black person. When the algorithm is deployed prospectively, an arrest forecast will mean something different depending on whether the person is gray or black.

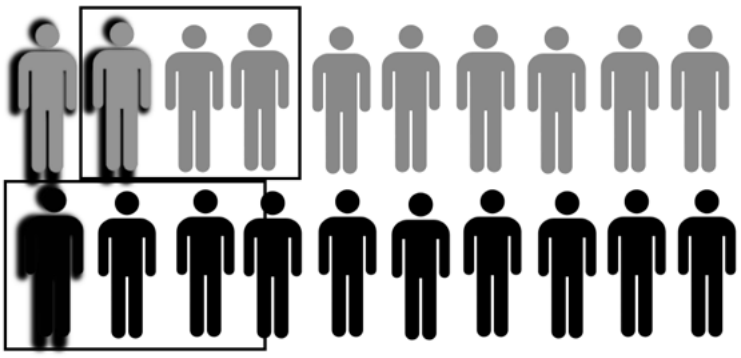
Figure 2: Groups with Different Base Rates of Arrest; Parity in False Positive Rates



	Gray	Black	
% of Rearrest Forecasts That Are Correct	50	31	Disparate Predictive Accuracy
% of No-Arrests Falsely Forecast for Arrest	25	25	Parity in False Positive Rates
% of Group Forecast for Rearrest	40	30.3	Closer to Statistical Parity

It is simple enough to recover predictive parity by altering the gray group for whom arrest is forecast, as depicted below. But that will introduce a new disparity. Now, among those who *are* rearrested (the shadowed figures), the algorithm correctly predicts arrest for 100% of the black arrestees, but "misses" 50% of the gray arrestees. There is now a dramatic disparity in false-negative rates.

Figure 3: Groups with Different Base Rates of Arrest; Parity in False-Positive Rates and Predictive Parity



	Gray	Black	
% of Rearrest Forecasts That Are Correct	33	31	~ Predictive Parity
% of No-Arrests Falsely Forecast for Arrest	25	25	Parity in False Pos. Rates
% of Group Forecast for Rearrest	30	30.3	~ Statistical Parity
% of Arrests Missed	50	0	Disparate False-Neg. Rates

What this example illustrates is that, if the base rate of the predicted outcome differs across racial groups, it is impossible to achieve (1) predictive parity, (2) parity in false-positive rates, and (3) parity in false-negative rates at the same time (unless prediction is perfect, which it never is). Computer scientists have provided mathematical proofs of this fact.[56] When base rates differ, we must choose to prioritize one of these metrics at the expense of another. Race-neutrality is not attainable.

C. Charting Predictive Equality

The reality is even more complex than this stylized example, because there are many additional possible metrics of inter-group equality. This subsection briefly charts the most important metrics, synthesizing the recent computer science literature on algorithmic fairness with the familiar legal concepts of disparate treatment and disparate impact. This taxonomy does not analyze legal liability. The goal, rather, is to organize the possible conceptual

[56] Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, BIG DATA (2017), https://arxiv.org/abs/1703.00056; Jon Kleinberg *et al.*, *Inherent Trade-Offs in the Fair Determination of Risk Scores* 4 (2017), arxiv.org/abs/1609.05807v2; *see also* Huq, *supra* note 13 (explaining this "impossibility result"); Richard A. Berk *et al*., *Forecasting Domestic Violence: A Machine Learning Approach To Help Inform Arraignment Decisions*, 13 J. EMPIRICAL LEGAL STUD. 94 (2016) (illustrating the impossibility using real arraignment data and a machine-learned algorithm that forecasts new arrest for a domestic violence offense within a period of twenty-one months).

measures of inter-group equality in a format accessible to both lawyers and statisticians. Those readers already immersed in the field may wish to skip directly to Part II.

U.S. law divides racially unequal action into two major categories: disparate treatment and disparate impact.[57] Neither triggers legal liability if the differential treatment or impact is adequately justified, but for purposes of this discussion we will ignore second-order questions of justification. Conceptions of equality in risk assessment can be classified as either disparate treatment or disparate impact metrics. Disparate-treatment metrics relate to the algorithmic process itself. Disparate-impact metrics relate to its outputs. This division also aligns loosely with the distinction between "individual" and "group" equality metrics, although that distinction is not a clean one.[58]

1. Disparate Treatment (Input-Equality) Metrics

Although disparate treatment is a contested concept, in current doctrine the term refers to any intentional differential treatment on the basis of race.[59] A prohibition on disparate treatment regulates the decision-making process itself. In the algorithmic context, the relevant process is the formula by which an algorithm produces a risk assessment (or forecast) for each individual. There are two possible metrics that can be understood as prohibitions on disparate treatment.

The first is **colorblindness.** 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 rationale for colorblindness is that if race can affect one's risk score, there will be some set of people with otherwise identical risk prognoses who receive different risk scores on the

[57] There are two primary vehicles for asserting discrimination claims in U.S. law: the Equal Protection Clause of the federal Constitution, and federal and state statutes that prohibit discrimination on various grounds, including on the basis of race. A discrimination claim pursuant to the Equal Protection Clause must allege and prove disparate treatment to succeed; a showing of disparate impact alone will not suffice. Anti-discrimination statutes also permit disparate-treatment claims, and some permit disparate-impact claims as well. As Richard Primus explains, although there are technical differences in the constitutional and statutory disparate treatment frameworks, analysis of a disparate treatment claim pursuant to either is fundamentally the same. *See* Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1354-62 (2010).

[58] Much recent work in algorithmic fairness has categorized measures of output-equality as either "group fairness" or "individual fairness" metrics. This dichotomy, however, can be misleading. Almost every possible measure of "group fairness" can be phrased using the word "individual" (*i.e.* predictive parity requires that, for any individual, a given risk score communicates the same average risk regardless of race). Conversely, any "individual fairness" metric can be phrased using the word "group" (*i.e.*, individual-risk equality requires that the group of people who present any given degree of risk all receive the same risk score). The difference is that "individual-fairness" metrics relate to how the algorithm arrives at its output in each individual case, whereas "group-fairness" metrics relate to the distribution of outputs and/or their accuracy across specified groups.

[59] *E.g.* Washington v. Davis, 426 U.S. 229, 241 (1976) (holding that differential treatment of people of different races violates the Equal Protection Clause only if motivated by "discriminatory racial purpose").

basis of race, and this is disparate treatment.[60] A mandate of colorblindness would align with anti-classification conceptions of equality in law.[61]

The second process-equality metric, which I call **individual-risk equality**, would prohibit the algorithm from assigning different scores, on the basis of race, to two individuals who present the same statistical risk. Put conversely, it would require the algorithm to treat individuals who present the same statistical risk in the same way. Individual-risk equality might seem synonymous with colorblindness, but it is not. Whereas colorblindness prohibits consideration of race in the calculation of risk, individual-risk equality kicks in later in the logic of risk assessment: Once an individual's statistical risk has been calculated, it prohibits the algorithm from considering race in deciding how to classify that risk—what risk score the person will receive. If a white person who poses an 8% chance of rearrest for violent crime is classified as "high-risk," or as a "6" on a six-point risk scale, a black person who poses the same risk must also be so classified, and vice-versa. Any two people who present the same risk must receive the same score (or classification, or forecast). This notion of equality prohibits different "cut points" by race.[62]

Both colorblindness and individual-risk equality reflect the Aristotelian notion that similarly-situated individuals should be treated alike. They just reflect slightly different judgments about which individuals are similarly situated for purposes of risk assessment. Individual-risk equality presumes that two individuals are similarly situated if they present the same statistical risk, calculated with as much precision as possible. Colorblindness presumes that two individuals are similarly situated if they present the same statistical risk, calculated without reference to race. If race moderates the predictive value of other factors, the two can be mutually incompatible.[63]

2. Disparate Impact (Output-Equality) Metrics

Disparate impact refers to differential effects of some decision-making process on members of one racial group.[64] It relates to the fairness of decision-making outputs. There are many different ways to compare algorithmic outputs across racial groups, because there are many different ways to measure the "output" of a predictive algorithm. Because these are inherently statistical concepts, in order to evaluate an algorithm by any one of these measures it is necessary to have a sizable number of the algorithm's

[60] In practice, "people with otherwise identical risk prognoses" will include people who have precisely the same observable risk traits, excluding race. But it may also include two people who each have different traits, but who nonetheless present equivalent statistical risk according to our best method of estimation.

[61] *See* Balkin & Siegel, *supra* note 130, 10-11 (2003) (explaining the distinction between these two approaches to equality law).

[62] Cut points are the statistical risk thresholds set for different risk classes – for instance, the classes of "high-risk," "moderate-risk," and "low-risk." *See, e.g.*, Eaglin, *supra* note 28.

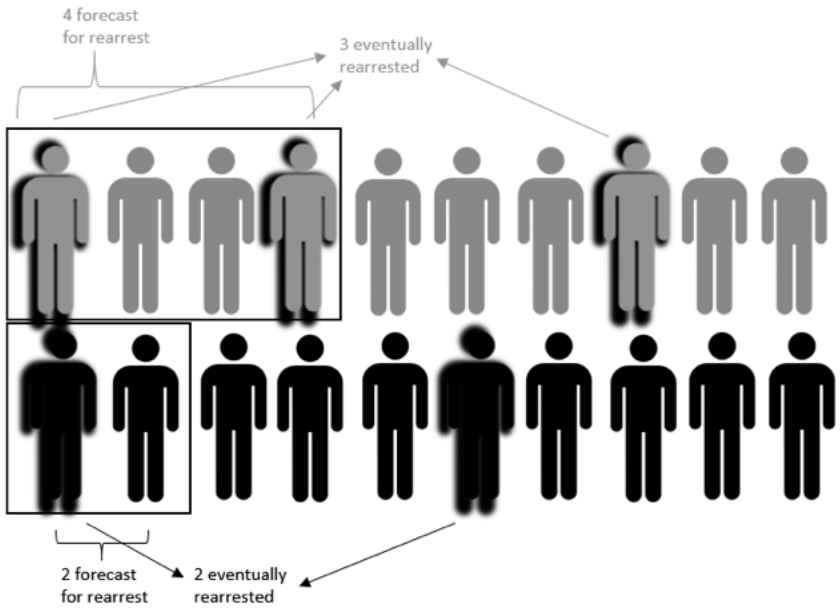
[63] For a fuller explanation of this possibility, *see infra* Part III.A.

[64] *See, e.g.*, Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(2) (2012); Griggs v. Duke Power Co., 401 U.S. 424, 430-31 (1971).

predictions for members of each racial group, and in most cases to know how many were ultimately correct. Output-equality metrics align with anti-subordination conceptions of equality.[65]

The following schema presents a core set of potential output-equality metrics. The figure below once again depicts two groups, gray and black, with different base rates of the outcome in question—say arrest for violent crime. Assume once again that the algorithm makes binary arrest / no-arrest forecasts. Once again, the shadowed figures will ultimately be arrested and the boxes represent those forecast for arrest.

Figure 4: Groups with Different Base Rates of Arrest, Again



Statistical Parity

Statistical parity requires that the same percentage of each group be forecast for arrest. That is, it requires parity in the total population impact of the prediction at issue. This is the simplest measure of inter-group equality. It is also the one that dominates disparate impact law. Federal EEOC guidance, for examples, provides that too great a divergence from statistical parity is prima facie evidence of "adverse impact."[66] In our example, the algorithm does not come close to achieving statistical parity: 40% of the gray group but only 20% of the black group are forecast for arrest (figures in the

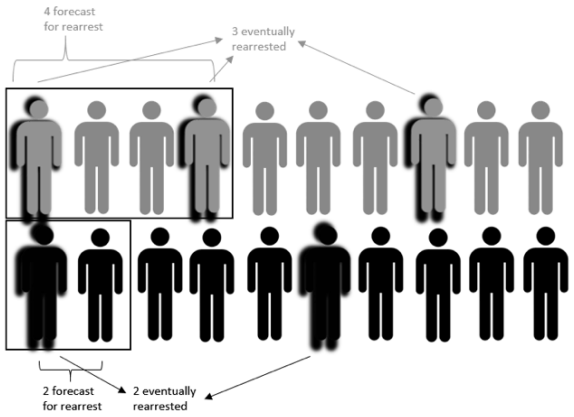
[65] Huq, *supra* note 2.

[66] The "four-fifths rule" provides that "[a] selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact." EEOC, Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2017).

boxes).[67] Statistical parity is sometimes called "demographic parity." Related metrics in the computer science literature include the Calders-Verwer (CV) gap[68] and the p-% rule.[69]

Predictive Parity

Predictive parity, the metric that Northpointe emphasized in the ProPublica / Northpointe debate,[70] measures the algorithm's rate of accuracy among those who receive the same forecast. If the algorithm's arrest forecasts are correct at an equal rate for each group, the algorithm achieves parity in *positive predictive value.* If the no-arrest forecasts are correct at an equal rate for each group, it achieves parity in *negative predictive value*. And if both are true, it achieves overall *predictive parity*. Statisticians and computer scientists have also referred to this metric of equality as "calibration within groups"[71] and "conditional use accuracy equality."[72] In our example, the algorithm achieves parity in positive predictive value only. For both the black and gray groups, 50% of those forecast for rearrest are indeed rearrested (the shadowed figures in the boxes).



[67] Note that the concept of population impact requires a definition of the relevant population. For purposes of comparing across racial groups, we might be interested in what proportion of defendants (for each group) are forecast for rearrest, or what proportion of the total group population in the county, or what proportion of some subgroup of defendants. We might, for instance, want to ensure that, among the subgroup of defendants with equivalent criminal histories, the percentage forecast for future arrest is the same for each racial group. Scholars call this "conditional statistical parity." Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness* (2017), https://arxiv.org/pdf/1701.08230, at 2.

[68] Kamishima *et al*. *Fairness-aware Learning through Regularization Approach*, JOINT EUROPEAN CONFERENCE ON MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES (Peter A. Flach, Tijl De Bie & Nello Cristianini eds., 2012).

[69] Zafar *et al*., *Fairness Constraints: Mechanisms for Fair Classification*, 54 PROC. 20TH INT'L CONF. ARTIFICIAL INTELLIGENCE & STAT. 952 (2017), http://proceedings.mlr.press/v54/zafar17a /zafar17a.pdf.

[70] For each racial group, the same percentage of the COMPAS's predictions were correct. This was true for each classification group—both for those deemed high-risk and for those deemed low-risk.

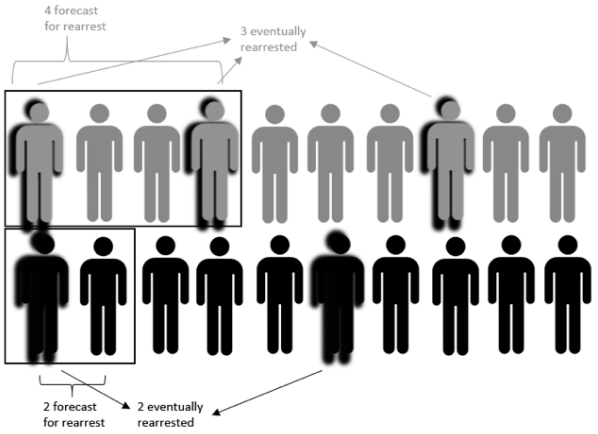
[71] Kleinberg *et al., supra* note 56.

[72] Berk *et al., Fairness in Criminal Justice Risk Assessments: The State of the Art*, https://arxiv. org/pdf/1703.09207.pdf, at 10.

Equal False-Positive / True-Negative Rates (Equal Specificity)

ProPublica, on the other hand, argued that equality requires parity in false-positive rates. The false-positive rate and its inverse, the true-negative rate, measure the algorithm's accuracy among those people who are "true negatives"—those who are *not* ultimately rearrested. The false-positive rate is the proportion of such people who are nonetheless forecast for arrest—the law-abiders mistakenly projected to commit crime. In our model, it is twice as high for the gray group as for the black. Of the seven gray people who will *not* be rearrested (un-shadowed gray figures), two are mistakenly forecast for arrest (29%), whereas of the eight black people who will *not* be rearrested (un-shadowed black figures), only one is mistakenly forecast for arrest (12%). The proportion of non-arrestees who are *correctly* predicted is the true-negative rate (or the algorithm's "specificity").[73] Statisticians and computer scientists have referred to equal specificity as "balance for the negative class" and as "predictive equality."

There is disagreement about whether this statistical vocabulary for forecasting errors is appropriate to risk assessment, because most risk assessment tools do not actually predict outcomes; they only assess the probability of a future event. If an event assessed as likely does not transpire, it does not render the initial probabilistic assessment "false."[74] Nonetheless, the binary language of true versus false prediction is an extremely helpful heuristic to explain where the costs of uncertainty fall.



Equal False-Negative / True-Positive Rates (Equal Sensitivity)

Whereas specificity measures the algorithm's accuracy among the true negatives (people who are *not* ultimately arrested), sensitivity measures the algorithm's accuracy among the true positives—people who *are* ultimately arrested. The proportion of this group correctly forecast for arrest is the true-positive rate; the proportion mistakenly forecast for no-arrest is the

[73] In our model this is the percentage of un-shadowed figures correctly left outside the box (five of seven gray (71%) and the seven of eight black (88%)).

[74] For further discussion of this point, *see infra* note 153 and accompanying text.

false-negative rate. The false-negative rate, in other words, is the percentage of future arrests that an algorithm "misses."

Our algorithm does not achieve equal sensitivity. For the gray group, two of the three people actually rearrested (the shadowed figures) are correctly predicted, so the true-positive rate is 2/3 (67%), and the false-negative rate is 1/3 (33%). For the black group, one of the two people actually rearrested (shadowed figures) is correctly predicted and one is not, so both the true-positive and false-negative rates are 1/2 (50%).

Computer scientists have referred to parity in true-positive rates as "balance for the positive class"[75] and as "equal opportunity," because it means that a true positive will have an equal chance of being correctly predicted regardless of group membership.[76] The happy language of "equal opportunity" is inapt in the criminal justice context, where a "positive" typically means rearrest; it makes more sense in assessment contexts where the "positive" outcome predicted is something good, like succeeding on the job or repaying a loan.

A related metric would demand parity in both sensitivity and specificity. In the technical literature, scholars have called this "balance for both classes," "equalized odds," "conditional procedure accuracy equality," and "equality of opportunity."

Equal Rate of Correct Classification

It is also possible to conceive of equality as parity in the rate of correction classification overall, or the percentage of each group correctly predicted. In our model, 70% of the gray figures are correctly classified (two actual rearrests—shadowed figures—in the box, and five no-rearrests outside the box). Of the black group, 80% are correctly classified (one actual rearrest in the box and seven no-arrests outside it).[77] Richard Berk and colleagues call parity in the rate of correct classification "overall parity."[78]

Equal Cost Ratios (Ratio of False Positives to False Negatives)

A last possible metric of equality in terms of error rates is parity in the ratio of false positives to false negatives, sometimes called the "cost ratio." The ratio matters because one kind of error may be worse than the other. Incorrectly predicting future arrest may be worse than incorrectly predicting no future arrest, or vice versa. Any algorithm will produce *some* ratio of false positives to false negatives. If stakeholders care what this ratio is, the algorithm can and should be designed accordingly. In the development of a predictive algorithm for a pilot project in Philadelphia, for instance, stakeholders determined that missing a new arrest for domestic violence (DV)

[75] Kleinberg *et al.*, *supra* note 56, at 4.
[76] Hardt *et al.*, *Equality of Opportunity in Supervised Learning* (2017), arxiv.org/abs/1610.02413.
[77] Inversely, only 20% of the black group but 30% of the gray group are classified incorrectly.
[78] Berk *et al.*, *supra* note 72.

with was ten times worse that incorrectly predicting that outcome.[79] Richard Berk and his colleagues, who were building the algorithm, therefore designed it to accept ten false positives rather than produce an additional false negative. They designed it, in other words, to produce a false positive-to-negative ratio of 10:1. Berk and colleagues call parity in cost ratios "treatment fairness."

AUC Parity

There are also a number of measures that express an algorithm's overall performance at sorting people along a spectrum of risk that tool developers frequently use to assess, and to claim, "race-neutrality." The most prominent is equality in the "area under the receiver operating characteristic curve" (also referred to as the "area under the curve," "AUC," or area under the "ROC") for a given tool as applied to each racial group. The AUC conveys the probability that, for any two people selected at random in the data, the algorithm will correctly order them in terms of risk (that is, it will score the higher-risk person as higher risk than the other). Parity in AUC scores is yet another measure of equality in predictive accuracy.

Table 1 charts these output metrics, their values in the black/gray example, and terms for each in the statistics and computer science literature.[80]

Table 1: Disparate Impact (Output Equality) Metrics

Parity In…	Gray	Black	Stats. / Comp. Sci. Equality Terms
Population Impact: % of group predicted P	40%	20%	Statistical Parity, Demographic Parity (related: Conditional Statistical Parity)[81]
Inverse: % predicted N	60%	80%	

[79] Berk *et al.*, *supra* note 56; *see also* Grant T. Harris & Marnie E. Rice, *Bayes and Base Rates: What Is an Informative Prior for Actuarial Violence Risk Assessment?*, 31 BEHAV. SCI. & L. 103, 106 (2013) (opining that "it can be reasonable for public policy to operate on the basis that a miss (*e.g.*, failing to detain a violent recidivist beforehand) is twice as costly as a false alarm (*e.g.*, detaining a violent offender who would not commit yet another violent offense)); Melissa Hamilton, *Adventures in Risk*, 47 Ariz. St. L.J. 1, 33 (noting that the contrary judgment is also reasonable).

[80] As did the text above, the table simplifies the relevant concepts in two ways: It (1) treats risk assessment as binary prediction, and (2) ignores the issue of whether the validation data will correspond to the population on which the tool is applied. This depends on the tool's "estimation accuracy," which is beyond the scope of this discussion. *See* Berk *et al.*, *supra* note 72, at 16.

[81] Berk *et al*, *supra* note 72 ("statistical parity"); Corbett-Davies *et al*., *supra* note 67 (same); Michael Feldman *et al.*; *Certifying and Removing Disparate Impact* (2015), arxiv.org/abs/1412.3756 (same); Fish *et al.*, *A Confidence-Based Approach for Balancing Fairness and Accuracy*, 2016 Proc. SIAM Int'l Conf. Data Mining 144 (2016), epubs.siam.org/doi/abs/10.1137/1.9781611974348.17 (same); Kamishima *et al.*, *Considerations on Fairness-aware Data Mining* Proc. 2012 IEEE International Conference on Data Mining Workshops 378 (2012), https://ieeexplore.ieee.org/abstract/document/6406465/ (same); Kleinberg *et al.*, *supra* note 56 (same); Richard Zemel *et al.*, *Learning Fair Representations*, 28 Proc. 30th Int'l Conf. Machine Learning (2013) (same); Hardt *et al.*, *supra*

Positive Predictive Accuracy: % of P predictions that are correct	50%	50%	Predictive Parity, Calibration within Groups, Conditional Use Accuracy Equality[82]
Negative Predictive Accuracy: % of N predictions that are correct	83%	88%	[Same terms as above]
True-Negative Rate (Specificity): % of Ns correctly predicted False-Positive Rate: % incorrectly predicted	71% 29%	88% 12%	Balance for the Negative Class, Predictive Equality[83]
True-Positive Rate (Sensitivity): % of Ps correctly predicted False-Negative Rate: % incorrectly predicted	67% 33%	50% 50%	Balance for the Positive Class, Equal Opportunity[84]
Both True Positive and True Negative Rates			Balance for Both Classes, Equalized Odds, Condt'l Procedure Accuracy Equality, Equality of Opportunity[85]
Overall Rate of Correct Classification: % of group correctly predicted	70% 30%	80% 20%	Overall Parity,[86] Overall Procedure Accuracy[87]

note 76 ("demographic parity"); Corbett-Davies *et al*., *supra* note 67 at 2 (using the term "conditional statistical parity" to refer to parity in population impact across smaller subgroups).

[82] Dieterich *et al.*, *supra* note 45 ("predictive parity"); Hardt *et al*., *supra* note 76 ("calibration within groups"); Kleinberg *et al*., *supra* note 56 ("calibration within groups"); Berk *et al*., *supra* note 72 ("conditional use accuracy equality").

[83] Kleinberg *et al.*, *supra* note 56 ("balance for the negative class"); Corbett-Davies *et al.*, *supra* note 67, at 2 ("predictive equality").

[84] Kleinberg *et al.*, *supra* note 56 ("balance for the positive class"); Hardt *et al.*, *supra* note 76 ("equal opportunity").

[85] Kleinberg *et al*. *supra* note 56 ("balance for both classes"); Hardt *et al.*, *supra* note 76 ("equalized odds"); Berk *et al*, *supra* note 72 ("conditional procedure accuracy equality"); Joseph *et al.*, *Fair Algorithms for Infinite and Contextual Bandits* (2017), https://arxiv.org/abs/1610.09559 ("equality of opportunity").

[86] Berk *et al*., *supra* note 72.

[87] Or "overall procedure accuracy." Berk *et al.*, *supra* note 72, at 13.

Inverse: % incorrectly predicted			
Distribution of Errors b/t FP & FN ("Cost Ratio")	2:1	1:1	Treatment Fairness[88]
Everything Above!			Total Fairness[89]

As explained in Section B above, several of these metrics will be mutually incompatible whenever base rates of the thing we have undertaken to predict diverge across racial groups. To achieve any one of these metrics, it will likely be necessary to sacrifice at least one of the others.

II. PREDICTION AS A MIRROR

A. The Premise of Prediction

There is a simple reason why it is impossible to achieve equality by every metric when base rates differ: Prediction functions like a mirror. The premise of prediction is that patterns observed in the past will repeat in the future. All that prediction does is identify such patterns and then offer them as projections about future events. It holds a mirror to the past, as the past is reflected in the data. If there is racial disparity in the data, there will be racial disparity in prediction too. It is possible to displace the disparity from one form to another, but impossible to eliminate it altogether. There can be no such thing as "race-neutral" prediction in a racially unequal world.

This fact about prediction is not unique to actuarial methods. Actuarial prediction reflects a particularly crystalline image of visible, quantified data, whereas subjective prediction reflects a foggy image of anecdotal data. But subjective and algorithmic prediction alike look to the past as a guide to the future, and thereby project past inequalities forward.

The deep problem, in other words, is not algorithmic methodology. Any form of prediction that relies on data about the past will produce racial disparity if the past data shows the very event we aspire to predict—the target variable—occurring with unequal frequency across racial groups. And if an algorithm's forecasts are correct at equal rates across racial lines, as were the COMPAS forecasts in Broward County,[90] any disparity in prediction is a reflection of disparity in the data. To understand and redress disparity in prediction, it is therefore necessary to understand how and when racial disparity arises in the data that we look to as a representation of *past* crime.

[88] *Id.*

[89] Berk *et al.*, *supra* note 72, at 15.

[90] That is, the algorithm achieved predictive parity. *See supra* notes 70-72 and accompanying text; Table 1.

B. Two Sources of Predictive Inequality

There are two possible sources of racial disparity in past-crime data: (1) racial distortion in the data vis-à-vis the underlying incidence of crime, or (2) an actual difference in the offending rate across racial lines for whatever category of crime a given tool predicts. The second of these possibilities evokes one of the most pernicious themes in racist ideology, the association of blackness with criminality.[91] Partly for that reason, it is essential to differentiate these two possible founts of predictive disparity. Some participants in the risk-assessment-and-race debate assume that any racial disparity in past-crime data reflects distortion;[92] others assume that it reflects differences in underlying crime rates.[93] So long as these conflicting assumptions go unstated, the debate cannot proceed.

Without confronting the two possible sources of disparity in prediction, moreover, it is impossible to remedy them, because each source of disparity demands a different response. Distortions in the data or risk assessment process can sometimes be corrected within that process itself. If correction is not possible—if the data cannot be made to reliably reflect the underlying incidence of crime—then it should not serve as the basis for risk assessment at all. If the data *does* reliably reflect the underlying incidence of crime, on the other hand, and predictive disparity flows from a difference in underlying crime rates, the disparity cannot be eliminated within the predictive process. But nor is the answer to jettison predictive tools. So long as the data reliably reflects the incidence of some event that is worth predicting, algorithmic risk assessment may have a valuable role to play.

It is thus imperative to acknowledge the two possible sources of disparity and strive to identify which is at issue in any given context. The remainder of this subsection explains the two possible sources of disparity in more depth.

1. Distortions in the Predictive Process

Distortions of the data or algorithm that exaggerate the riskiness of black men relative to other demographic groups will produce racial disparity in prediction. This kind of disparity is sometimes called "irrational" discrimination, because it has no statistical justification vis-à-vis the underlying reality. It corresponds to bias in the statistical sense. There are three primary mechanisms for such distortion in risk assessment: a proxy

[91] *See generally, e.g.*, RANDALL KENNEDY, RACE, CRIME AND THE LAW (1997); KATHERYN RUSSELL-BROWN, THE COLOR OF CRIME (1998).

[92] *See, e.g.*, testimony of Mark Houldin, Philadelphia Defender Association, hearing on the proposed Pennsylvania Risk Assessment Tool for Sentencing (June 13, 2018), available at http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument/testimony, at 8-9.

[93] *See, e.g.*, Houldin testimony, *supra* (citing research commissioned by the Pennsylvania Sentencing Commission as interpreting racial difference in arrest rates to reflect racial difference in commission rates).

target variable with racial skew, race-specific data flaws, and intentional distortion ("masking").

The most important form of distortion in the criminal justice context is the selection of a target variable that is a poor proxy for the actual event we wish to predict, and for which the base rate is racially skewed vis-à-vis the actual event of interest. To be more concrete: The goal of most criminal-justice risk assessment is to predict the commission of serious crime. And the tools mostly purport to predict the commission of future crime.[94] But that is not what they actually predict. They predict arrest, on the premise that this is the best available proxy for crime commission.[95] Most assess the likelihood of arrest for any offense at all (within a designated timespan). "Any arrest" is an extremely loose proxy for the commission of a serious crime, and in many jurisdictions there is likely to be a substantial racial skew between base rates of arrest and base rates of criminal offending.[96]

The choice to predict arrest therefore has the potential to introduce serious racial distortion in risk assessments vis-à-vis the risk that a person will actually *commit* crime. Arrest is an event largely contingent on the discretion of third parties—the police. And police have historically arrested black men with unjustified frequency, especially for drug- and low-level crimes.[97]

As between a black and white defendant who are equally likely to commit crime, the black defendant may be more likely to be arrested.[98]

---

[94] *See, e.g., Public Safety Assessment: Risk Factors and Formula*, www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf (purporting to predict "new criminal activity"); *Overview of the LSI-R*, https://www.mhs.com/MHS-Publicsafety?prodname=lsi-r (purporting to predict, inter alia, "recidivism").

[95] We do not have good data on criminal acts by particular individuals. Many crimes are not reported, many crimes that are reported are never prosecuted, many prosecutions are dropped, and even convictions do not necessarily establish with certainty what criminal act the convicted person committed. Whether arrest is actually the best proxy for commission of crime given this data problem is a difficult and contested question. *See* Anna Roberts, *Arrest as Guilt*, __ ALA. L. REV. __ (forthcoming).

[96] See, e.g., Kristian Lum, *Limitations of Mitigating Judicial Bias with Machine Learning*, 1 NATURE HUM. BEHAV. 1 (2017). Analogously, some algorithmic tools that purport to predict job success actually assess the likelihood of *getting hired* in status quo conditions. Brian Jacon, *et al.*, *Teacher Applicant Hiring and Teacher Performance: Evidence from DC Public Schools* 4 (Nat'l Bureau of Econ. Res., Working Paper No. 22054), http://www.nber.org/papers/w22054.

[97] *E.g.* Model Penal Code § 1.02(2) (proposed final draft 2017) (noting that racial disparities in sentencing that arise from racial skew in law enforcement "are largest for crimes at the low end of the seriousness scale—especially drug offenses" and collecting sources); David Huizinga *et al*. Report to the Office of Juvenile Justice and Delinquency Prevention, Disproportionate Minority Contact in the Juvenile Justice System: A Study of Differential Minority Arrest/Referral to Court in Three Cities (2007), www.ncjrs.gov/pdffiles1/ojjdp/grants/219743.pdf (study evaluating longitudinal data from three cities and finding substantial racial differences in police contact after controlling for differences in self-reported offending); Lauren Nichol Gase *et al*, *Understanding Racial and Ethnic Disparities in Arrest: The Role of Individual, Home, School, and Community Characteristics*, 8 RACE & SOC. PROBS. 296 (2016) (finding "that racial / ethnic differences in arrest were not explained by differences in individual-level delinquent behaviors," but were explained by "neighborhood racial composition"); Kristian Lum & William Isaac, *To Predict and Serve?*, 13 SIGNIFICANCE 14, 14-19 (2016).

[98] Preeti Chauhan *et al.*, *Trends in Arrests for Misdemeanor Charges in New York City, 1993-2016* 21, MISDEMEANOR JUSTICE PROJECT (2018), misdemeanorjustice.org/wp-content/uploads/2018/01/2018_01_24_MJP.Charges.FINAL_.pdf; Jeffrey Fagan & Tracey L. Meares, *Punishment,*

Conversely, the fact that a black defendant is more likely to be arrested may not mean s/he is more likely to commit crime. There is thus reason to think that tools that assess the likelihood of "any arrest" may be racially biased in the sense that a given score—which corresponds to some likelihood of arrest—will mean a different risk of crime *commission* for black versus white defendants. This kind of disparity is particularly troubling because it can be invisible. Without good data on true rates of offending in each group, it is impossible to tell whether there is racial skewing between arrest and crime rates.

The most direct solution to this problem is to choose a different target variable, one that better represents the event we want to predict without embedding racial skew. In practice, this can be extremely difficult. The complexities are discussed further in Part III.

Racial distortion can also result if the data is systematically less reliable for one racial group than another. This problem can arise if the data is either more limited or has greater systemic inaccuracy for one racial group.[99] In the criminal justice context, though, there is no indication that arrest data themselves are systematically more limited or less accurate for one racial group than another.

The last potential source of racial distortion in prediction is intentional manipulation of the data or algorithm to disadvantage one racial group, or what Solon Barocas and Andrew Selbst call "masking."[100] There is no evidence that this is a serious concern in the context of criminal justice risk assessment.[101] There are also ways to prevent it from becoming one. So long as the data on the basis of which tools are developed and validated is made public, as it should be, independent researchers can replicate the tool design and validation process and check for symptoms of racist manipulation.

In addition to these sources of distortion in predictions themselves, system actors can introduce racial distortion in responding to risk. A recent study by Megan Stevenson concludes that, when pretrial risk assessment was implemented in Kentucky, judges in rural and largely white counties responded to risk scores differently than judges in urban counties with a greater black population, with the result that the new process disproportionately benefitted white defendants.[102] In terms of actual outcomes, this potential source of disparity may be most important of all.

Deterrence and Social Control: The Paradox of Punishment in Minority Communities 6 OHIO ST. J. CRIM. L. 173, 178-80 (2008).

[99] An algorithm developed for maximum accuracy will conform to the majority data, and may be less accurate for members of the underrepresented group. Tool designers can ameliorate this problem by weighting the minority-group data more heavily, by developing separate algorithms for each racial group, or by endeavoring to include more data to equalize group representation in the dataset. Sukarna Barua *et al.*, *MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning*, 26 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENGINEERING 405 (2014).

[100] Barocas & Selbst, *supra* note 2, at 692-94. They call it "masking" because machine-learning technologies offer opportunities to intentionally distort an algorithm in ways that are difficult to detect.

[101] *Accord* Huq, *supra* note 13.

[102] *See* Stevenson, *supra* note 5.

Each of these mechanisms of distortion—a target variable with racial skew, race-specific data flaws, masking, and a race-skewed response to prediction—can be addressed in the risk assessment process. In theory each can be eliminated, although reality presents challenges. Of them, the target variable problem and the possibility of a race-skewed response seem by far the most significant sources of racial distortion in current practice.

2. Different Rates of Crime Commission

The second possible source of predictive disparity is a difference in the underlying incidence of crime. Even if every distortion is eliminated, prediction will still produce racial disparity if the rate of crime commission is unequal across racial groups in the relevant population, for whatever category of crime a given tool predicts.

This possibility arises because crime is the product of complex social and economic determinants that, in a race- and class-stratified society, may also correlate with demographic traits. Where that is so, the incidence of a given type of crime may vary among demographic groups. A number of recent studies have found, for instance, that contemporary white and Hispanic college students use illicit drugs at significantly higher rates than African-American and Asian students.[103] White men have committed the vast majority of mass shootings in the United States over the last thirty years.[104] Nationwide firearm homicide rates have been higher in recent decades in black communities than white, but the degree of disparity varies by state.[105] High-stakes financial crimes are disproportionately committed by white men working in financial services firms.[106]

In the Broward County data, as well as several other datasets used in recent risk assessment studies, arrest rates for offenses designated as "violent" were higher among the black population in the data than the

[103] *See, e.g.*, Sean Esteban McCabe et al., *Race/Ethnicity and Gender Differences in Drug Use and Abuse Among College Students*, 6 J. ETHN. SUBST ABUSE 75 (2007) (study providing "strong evidence from one university that Hispanic and White undergraduate students were at increased risk for drug use and abuse," and chronicling related literature).

[104] Number of Mass Shootings in the United States between 1982 and June 2018, by Mass Shooter's Race and Ethnicity, https://www.statista.com/statistics/476456/mass-shootings-in-the-us-by-shooter-s-race.

[105] *See, e.g.*, Alexia Cooper and Erica L. Smith, Homicide Trends in the United States, 1980-2008, 11 (Bureau of Justice Statistics 2011), https://www.bjs.gov/content/pub/pdf/htus8008.pdf#page=27; Michael Planty and Jennifer L. Truman, Firearms Violence 1993-2011, 5 (Bureau of Justice Statistics 2013) (showing rates of firearm victimization by race, and that most firearms crime is intra-racial), https://www.bjs.gov/content/pub/ pdf/fv9311.pdf; Corinne A. Riddell, Sam Harper, Magdalena Cerdá and Jay S. Kaufman, *Comparison of Rates of Firearm and Nonfirearm Homicide and Suicide in Black and White Non-Hispanic Men, by U.S. State*, 168 Ann. Internal Med. 712 (2018), http://annals.org/aim/fullarticle/2679556/comparison-rates-firearm-nonfirearm-homicide-suicide-black-white-non-hispanic.

[106] *See* Brian Clifton, Sam Lavigne, and Francis Tseng, Predicting Financial Crime: Augmenting the Predictive Policing Arsenal, The New Inquiry (April 26, 2017), https://thenewinquiry.com/white-collar-crime-risk-zones (offering new predictive technology "trained on incidents of financial malfeasance from 1964 to the present day, collected from the Financial Industry Regulatory Authority (FINRA)").

white.[107] Scholars Jennifer Skeem and Christopher Lowenkamp have opined that the disparity represents differential offending rates rather than differential enforcement.[108] This Article does not stake any position on whether that is so; I do not have the data or the expertise to judge.

The point is that *if* underlying offending rates do vary by race in the data on which a given algorithm is built, racial disparity in prediction is unavoidable. The reason, once again, is that prediction is a kind of mirror. If the black population in the relevant data is statistically riskier with respect to the designated crime category, risk assessment tools will recognize as much. If the mirror is modified to ignore this statistical fact, that very blindness will have disparate racial impact: In treating the black and white groups subject to assessment as statistically identical, the tools will "miss" more of the designated crimes committed by black individuals, which, because most crime is intra-racial, will disproportionately befall communities of color. No matter how the data or algorithm is altered, inequality in commission rates for the crime(s) we undertake to predict will produce inequality in prediction.

It is important, in considering this possibility, to recognize what any such difference in crime commission rates would and would not signify. Differential crime rates do not signify a difference across racial groups in individuals' innate "propensity" to crime.[109] What they signify are social and economic divides. Where the incidence of crimes of poverty and desperation varies by race, it is because society has segregated communities of color and starved them of resources and opportunity.[110] Where race and gender differences exist in the rate of high-stakes financial crime, it is because white men retain control of the levers of high-stakes finance. Crime rates are a manifestation of deeper forces; racial variance in crime rates, where it exists, manifests the enduring social and economic inequality that centuries of racial oppression have produced.

---

[107] Dieterich *et al.*, *supra* note 45; *see also* Flores *et al.*, *supra* note 45; Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments* (2017), crim.sas.upenn.edu/sites/default/files/Berk_FairJuvy_1.2.2018.pdf; Skeem & Lowenkamp, *supra* note 34, at 689-90.

[108] Skeem & Lowenkamp, *supra* note 34, at 689-90 (opining that arrest for a "violent offense" is a "valid criterion" free from racial skew in law enforcement); *see also* Alex R. Piquero et al., *A Systematic Review of Age, Sex, Ethnicity, and Race as Predictors of Violent Recidivism*, 59 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 5 (2015).

[109] The notion that differential crime rates signal a difference in innate criminal propensity has been a central justification for racist ideology and practices. *See generally, e.g.*, RANDALL KENNEDY, RACE, CRIME AND THE LAW (1997); KATHERYN RUSSELL-BROWN, THE COLOR OF CRIME (1998). Criminological literature has compounded the problem by occasionally describing differences in group statistical risk as a difference in "criminal propensity."

[110] *See, e.g.*, MEHRSA BARADARAN, THE COLOR OF MONEY: BLACK BANKS AND THE RACIAL WEALTH GAP (2017); KENNEDY, *SUPRA*; Model Penal Code § 1.02(2) (proposed final draft 2017) ("Serious crime rates, and victimization rates, are highest in America's most disadvantaged communities, which overwhelmingly are minority communities."); *id.* (citing sources on "the multiple causes of high crime rates in disadvantaged communities," along with research demonstrating that "the 'underclass' status of a community is associated with high crime rates among those who live there, regardless of race and ethnicity"). This is not to disclaim all individual responsibility for criminal acts. But individual responsibility for particular acts does not equate to group responsibility for group crime rates.

In sum: Figuring out the nature of the disparity in any predictive context is a necessary first step in redressing it. Disparities produced through distortion can be eliminated within a risk assessment system itself, at least in theory—and if the distortion cannot be corrected, the entire enterprise of risk assessment is compromised at its core. But disparities cannot be eliminated if they flow from underlying differences in the base rate of the very thing we want to predict. And rejecting risk assessment altogether, in those circumstances, may do more harm than good.

This is not to say that it will always be possible to disentangle distortion from differential crime rates. It sometimes may not be, as Part III discusses in more depth. That reality, too, is important to confront, because the question of how to proceed in such circumstances demands moral and policy judgment. Relatedly, acknowledging that crime rates vary across demographic groups for different crime categories helps to foreground the policy question of what kinds of crime we ought to predict.[111] The categories of "violent" or "serious" crime are themselves cultural constructs, and the way that stakeholders define them for purposes of risk assessment will have profound demographic implications.

These are the reasons that it is important to distinguish between distortion and differential offending rates as possible sources of racial disparity in prediction. Whichever the source, though, the three strategies most commonly advocated to redress predictive disparity are off the mark. Parts III explains why.

III. NO EASY FIXES

As the risk-assessment-and-race debate accelerates, critics have increasingly argued for three strategies to promote racial equity in prediction. The first is the exclusion of race and factors heavily correlated with race as input variables.[112] The second is "algorithmic affirmative action:" some intervention in the design of a predictive algorithm to equalize its outputs, by one or more of the metrics enumerated above.[113] In particular, advocates have urged intervention to ensure an equal rate of adverse predictions across racial groups ("statistical parity"), or equal error rates among those who have the same outcome in each racial group (parity in false-positive and false-negative rates). Aziz Huq has recently offered the more abstract prescription that each algorithm should be designed to ensure that its predictions impose no net

[111] *See, e.g.*, Timothy R. Schnacke, "Model" Bail Laws: Re-Drawing the Line between Pretrial Release and Detention (April 2017), www.clebp.org/images/04-18-2017_Model_Bail_Laws_CLEPB_.pdf (emphasizing the importance of defining the relevant risks in context of pretrial risk assessment).

[112] *E.g.*, Chander, *supra* note 35 (urging advocates to focus on "inputs and outputs" rather than algorithms themselves); Danielle Kehl, Priscilla Guo, & Samuel Kessler, *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing* (2017), dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf ("Critical issues also need to be addressed in the development phase of these algorithms, particularly with regard to the inputs and how they are used."); Huq, *supra* note 13, at 27 (discussing "the problem of distorting feature selection").

[113] *E.g.* Chander, *supra*, at 1039 (calling for "algorithmic affirmative action").

burden on communities of color.[114] The remainder of the discussion will use the term "algorithmic affirmative action" to refer to these proposals collectively, recognizing that this shorthand is reductive. Lastly, critics argue that, if algorithms cannot be made race-neutral, the criminal justice system should reject algorithmic methods altogether.[115]

This Part argues that all three strategies are misguided. Though well-intentioned, they have the potential to compromise the goal of racial equity rather than to further it.

A. Regulating Input Variables[116]

Input variables are often cited as the primary concern in the quest for racial equity in risk assessment.[117] It is an almost universal orthodoxy, in fact, that race must be excluded as an input to prediction.[118] Many people extend the principle to variables that correlate with race in a given locale, like zip code. The underlying concern is that the use of such factors will produce higher risk scores for black defendants and thereby compound historical racial oppression.

This focus on input variables, however, is not an effective path toward racial equity. The most basic reason is that excluding race and race proxies might actually hurt black defendants. In this context, as elsewhere, being blind to race can mean being blind to racism. As Justice Sotomayor replied to Justice Roberts, the "way to stop discriminating on the basis of race" is not to ignore race, but rather to apply law and develop policy "with eyes open to the unfortunate effects of centuries of racial discrimination."[119]

A simple example illustrates. In New Orleans, when I worked there as a public defender, the significance of arrest varied by race. If a black man had three arrests in his past, it suggested only that he had been living in New Orleans. Black men were arrested all the time for trivial things. If a white man had three past arrests, on the other hand, it suggested that he was really bad news! White men were hardly ever arrested; three past arrests indicated a highly unusual tendency to attract law-enforcement attention. A race-blind

---

[114] Huq, *supra* note 13, at 65.

[115] *E.g.* Leadership Council on Civil and Human Rights, *supra* note 38; John Ralphing, *Human Rights Watch Advises Against Using Profile-Based Risk Assessment in Bail Reform*, HUMAN RIGHTS WATCH (July 17, 2017), www.hrw.org/news/2017/07/17/human-rights-watch-advises-against-using-profile-based-risk-assessment-bail-reform.

[116] I explore this subject matter more comprehensively in a follow-on article, *Algorithmic Fairness and the Myth of Colorblindness* (manuscript on file with author).

[117] *See* sources cited in *supra* note 112.

[118] *See, e.g.*, Starr, *supra* note 28, at 811 ("There appears to be a general consensus that using race would be unconstitutional.").

[119] Justice Roberts, writing for the majority in *Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1*, declared that "[t]he way to stop discrimination on the basis of race is to stop discriminating on the basis of race." 551 U.S. 701, 748 (2007). Justice Sotomayor rejoined, seven years later, that "[t]he way to stop discrimination on the basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination." Schuette v. Coal. to Defend Affirmative Action, Integration & Immigrant Rights & Fight for Equal. By Any Means Necessary (BAMN), 134 S. Ct. 1623, 1676 (2014) (Sotomayor, J., dissenting).

algorithm would be blind to this difference. It would treat the two men as posing an identical risk. The algorithm could not consider the arrests in the context of disparate policing patterns and recognize that they were a much less significant indicator of risk for the black man than for the white.[120] It would perpetuate the historical inequality by overestimating the black man's relative riskiness and underestimating the relative riskiness of the white man.

A colorblind algorithm, in other words, might discriminate on the basis of race. In a shallow sense, the colorblind algorithm avoids racially disparate treatment. It treats two people with otherwise identical risk profiles exactly the same. In a deeper sense, though, the algorithm does engage in disparate treatment on the basis of race. In failing to recognize that the context of race powerfully affects the significance of past arrests, it inflates the black man's risk score and deflates the white man's relative to their true values.

In statistical terms, the problem is that, as a result of disparate law-enforcement practices, race might moderate the predictive value of certain variables (or the algorithm as a whole), such that the algorithm overestimates risk for black people relative to white.[121] A few risk assessment tool developers have encountered the problem in practice, discovering that variables like past arrests or misdemeanor convictions are less predictive for black people than white.[122] The usual response is to simply eliminate the problematic input variables from the model. But that solution has a price in accuracy.[123] The cost in accuracy might fall disproportionately on communities of color, as discussed at greater length below.[124]

The alternative is to allow an algorithm to assess the significance of risk factors *contingent on* race. If race does moderate the factors' predictive value as just described, this would lower risk scores, on average, for black defendants. It would achieve what a group of computer scientists have

[120] Michael Tracy makes an analogous argument for providing capital juries statistical information about how much more likely prosecutors are to seek the death penalty for black defendants. Michael Tracy, *Race As A Mitigating Factor in Death Penalty Sentencing*, 7 GEO. J.L. & MOD. CRITICAL RACE PERSP. 151, 159 (2015) (arguing that if jurors are aware of this disparity, a black defendant "may seem less deserving of a death sentence").

[121] This situation arises in every predictive context. In education testing, for instance, it is well established that the correlation between SAT scores and intelligence varies by race, and by circumstance. Harold Berlak, *Race and the Achievement Gap*, *in* CRITICAL SOCIAL ISSUES IN AMERICAN EDUCATION (H. Svi Shapiro & David E. Purpel eds., 3rd ed. 2005). A high score achieved by a student who benefited from the best possible primary education and extensive SAT preparation likely means less about her native intelligence than the same score achieved by a student who did not.

[122] Richard Berk and Marie Van Nostrand have each reported finding, in different datasets, that one-to-two past misdemeanor convictions were less predictive of future serious arrest for people of color than for white people. Berk, *supra* note 107; Christopher T. Lowenkamp, Marie VanNostrand & Alexander Holsinger, *Investigating the Impact of Pretrial Detention on Sentencing Outcomes*, ARNOLDFOUNDATION.ORG (2013), www.arnoldfoundation.org/wpcontent/uploads/2014/02/LJAF_ Report_statesentencing_FNL.pdf. The Pennsylvania Sentencing Commission recently rejected past arrests entirely as input variables because they had such different predictive significance across racial lines. Pa. Sentencing Commission, Risk Assessment Update: Arrest Scales (February 28, 2018 draft), *available at* www.hominid.psu.edu/ specialty_programs/pacs/publications-and-research/research-and-evaluation-reports/risk-assessment.

[123] The Pennsylvania Sentencing Commission, for instance, has elected to rely on past conviction rather than past arrest data despite the fact that it renders the model significantly less accurate overall.

[124] *See* Part IV.C.1.

dubbed "fairness through awareness."[125] And it would improve rather than compromise the accuracy of the tool. Under these circumstances, including race as an input variable would promote racial equity and accuracy at the same time.[126] This approach is not feasible for simple checklist tools, but it could be for the machine-learned programs that represent the future.

In fact, to achieve any specific form of output equality, it may be necessary to treat race as an input. To equalize false-positive rates across racial groups, for example, it will likely be necessary to have race-specific risk thresholds for each risk class—which is to say that the algorithm will treat people who pose the same risk differently on the basis of race.[127] The same is likely true for equalizing cost ratios across racial groups.[128] To achieve predictive parity, it may be necessary to manipulate the data to cancel out the effect of race on other observable variables,[129] or assess the predictive import of every input variable contingent on race. Algorithmic prediction thus offers a particularly clear lens on the conflict between anti-classification and anti-subordination conceptions of equality.[130]

Yet neither excluding race and related factors nor including them can equalize outcomes entirely if the thing we have undertaken to predict, the target variable, correlates with race itself. So long as the target variable correlates with race, regulating input data is futile. If the event we have undertaken to predict happens with greater frequency to people of color, a competent algorithm will predict it with greater frequency for people of color. Whatever input data is made available, the facts that correlate with the target variable—and therefore become the algorithm's predictors—will also correlate with race, because the target variable does. The only way to break the race correlation is by compromising the ability of the algorithm to predict the target variable at all. Excluding criminal history data, for instance, might dramatically reduce the disparate racial impact of predicting future arrest, but will also compromise its ability to predict future arrest in a dramatic way. To

[125] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, *in* PROC. 3RD INNOVATIONS IN THEORETICAL COMPUTER SCI. CONF. (2012).

[126] *Accord* Kim, *supra* note 40, at 918 ("If the goal is to reduce biased outcomes, then 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 and the risk of omitted variable bias.").

[127] Hardt *et al.*, *supra* note 76; Corbett-Davies *et al.*, *supra* note 67.

[128] Berk, *supra* note 107 at 13-14 (explaining that, to equalize cost-ratios across racial groups in a juvenile risk assessment context, it was necessary to undertake "separate forecasting exercises" for white and black juveniles, respectively, and that the machine-learned forecasting algorithms the data produced were different for each racial group).

[129] There are different ways to attempt this, and many risk assessment tool developers do. Marie VanNostrand, who has developed many of the checklist pretrial risk assessment tools in current use, simply searches for risk factors that are equally predictive across racial lines and discards those that are not. This approach is straightforward, but could have a steep cost in overall accuracy. *See* Richard Berk *et al.*, *Fairness in Criminal Justice Risk Assessments*, __ SOC. METHODS & RES. 1, 12-18 (2018).

[130] *See generally, e.g.*, Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003); Helen Norton, *The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality*, 52 WM. & MARY L. REV. 197, 206-215 (2010); Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 520-23 (2003). This theme is explored at greater length in Sandra G. Mayson, *Algorithmic Fairness and the Myth of Colorblindness* (work in progress).

eliminate racial disparity in the prediction of a racially disparate event is to cripple the predictive tool.

Some readers may feel that crippling the predictive tool is a good thing. If the tool predicts a race-skewed target variable like "any arrest," for instance, the tool has dubious value to begin with. In that situation, though, the better answer is to stop predicting the meaningless event entirely.[131] And if the target variable does *not* embed racial distortion, crippling the predictive tool can be counterproductive, because the loss in accuracy may inflict proportionally more "errors" on black communities than white.[132]

The larger point is that colorblindness is not a meaningful measure of equality. It can exacerbate rather than mitigate racial disparity in prediction.[133] Even if it does mitigate disparity in prediction, that improvement may come at a cost in accuracy that itself has racially disparate impact. So long as the target variable correlates with race, predictions will be racially uneven—or they will be so distorted as to be useless. In those circumstances, colorblindness is at best a superficial and at worst a counterproductive strategy for racial equity.[134]

B. Equalizing (Some) Outputs

Algorithmic affirmative action has similar shortcomings. Again, for purposes of this discussion "algorithmic affirmative action" refers to an intervention to produce statistical parity, equal false-positive or false-negative rates, or to equalize the "net burden" of prediction across racial lines. The stakes of such interventions depend on whether the disparity they seek to redress is a product of distortion in the data or of a difference in underlying crime rates by race. In either case, though, the interventions fall short.

1. Equalizing Outputs to Remedy Distortion

Consider, first, algorithmic affirmative action designed to remedy racial distortion in the data vis-à-vis the event we aspire to predict. In the context of criminal justice risk assessment, the gravest concern is that racial disparity in overall arrest rates reflects disparate law enforcement, rather than disparate rates of offending. If this is true, and what we assess is the likelihood of arrest, risk scores will overstate the riskiness of black men relative to the risk of actual crime commission. The goal of algorithmic

[131] *See infra* Part B.1.

[132] *See infra* Part B.2 and Appendix A.

[133] *Accord* Huq, *supra* note 13; Kroll *et al*, *Accountable Algorithms*, 165 UNIV. PA. L. REV. 633 (2017); Kim, *supra* note 40, at 867 ("if the goal is to discourage classification bias, then the law should not forbid the inclusion of race, sex, or other sensitive information as variables, but seek to preserve these variables, and perhaps even include them in some complex models.").

[134] David A. Strauss, *The Myth of Colorblindness* 1986 SUP. CT. REV. 99, 114 ("The one option that is not open is the ideal of colorblindness—treating race as if it were, like eye color, a wholly irrelevant characteristic. That is because it is not a wholly irrelevant characteristic. Race correlates with other things . . . .").

affirmative action is to adjust the data to cancel out this racial distortion in arrest rates.[135]

This strategy presumes that the scale of the distortion is known. If that is so, it should indeed be possible to cancel it out, although there are technical complexities. But it is hardly ever the case that the scale of the distortion is known. The reason we resort to arrest as a proxy for crime commission in the first place is that we cannot see crime commission directly.

Given that the scale of any distortion is usually unknown, the more direct solution to the problem is to simply avoid target variables that are likely to be racially skewed vis-à-vis the thing we really care about.[136] If arrest risk does not correspond to serious-crime risk, we should stop measuring it. It does not tell us what we want to know in any case. Risk assessment tools should predict something closer to the harm we actually want to avoid.

The challenge is to identify what we actually want to predict and avoid.[137] I have argued elsewhere that risk assessment tools should assess the risk of violent crime,[138] but the category is amorphous—does it include burglary? A bar fight? DUI?—and the judgment is contestable. Perhaps we should be equally concerned with the risk of financial crime.[139] The point is that the decision about what to predict is a momentous one, and should be made on the basis of law and considered policy judgment rather than what data is most readily available.[140]

Even resorting to more specific target variables may not solve the problem. Violent-crime arrest, for instance, remains an inexact proxy for violent crime itself. Police sometimes arrest the wrong person, and many violent crimes never lead to arrest at all. There could still be racial skew between the arrest rates and underlying offense rates. This might be so even if arrest rates track the incidence of reported crimes by race.[141] If white communities report domestic violence with less frequency when it happens, for example, violent-crime reports would embed racial skew vis-à-vis actual rates of offending, and arrest rates that track report rates would just carry the distortion forward.

---

[135] Berk, *supra* note 107 (considering data modifications along these lines); Sorelle A. Fiedler, *et al.*, *On the (Im)possibility of Fairness (*2017), arxiv.org/pdf/1609.07236 (raising a similar scenario with respect to SAT scores and college admissions algorithms designed to assess students' academic potential).

[136] SOCIETY FOR INDUSTRIAL ORGANIZATIONAL PSYCHOLOGY, PRINCIPLES FOR THE VALIDATION AND USE OF PERSONNEL SELECTION PROCEDURES 33 (4th ed. 2003), https://docplayer.net/58223-Principles-for-the-validation-and-use-of-personnel-selection-procedures-fourth-edition.html ("Confidence in the criterion measure is a prerequisite for an analysis of predictive bias.").

[137] Andrew Selbst characterizes this task as "defin[ing] the problem" for prediction. Selbst, *supra* note 2, at 131-33; *see also* Schnacke, *supra* note 111 at 109-14.

[138] Mayson, *supra* note 5, at 562.

[139] *See* Clifton et al, *supra* note 106.

[140] As Selbst notes in his discussion of predictive policing, "[u]sing data mining also tends to bias organizations toward questions that are easier for computers to understand." Selbst, *supra* note 2, at 132.

[141] The correspondence of arrest and crime-report rates by race is one fact that scholars sometimes cite as evidence that arrest rates lack racial skew vis-à-vis offending rates. *See* Skeem & Lowenkamp, *supra* note 34, at 690.

Stated in more general terms, one might object that we can never be confident that our target variable is free of racial distortion.[142] We must rely on the past to predict the future, but we see the past only hazily, through the splintered lens of data. We can never know how faithfully the data represent past reality because we have no direct access to past reality.

This is a profound objection, but it applies to more than algorithmic methods. It is an objection to prediction itself. All prediction presumes that we can read the past with enough reliability to make useful projections about the future. Perhaps in some contexts we cannot. Maybe the bottom line is that our past crime data is inadequate to serve as the basis for any prediction.[143] Or maybe the answer varies by crime category. But if this is the case, the answer is not to make the data reflect the past as we wish it had been. That merely distorts the mirror so that it neither reflects the data nor any demonstrable reality. The answer is simpler. If the past data does not reliably represent the events we want to avoid, we should stop consulting it as a guide to the future.

2. Equalizing Outputs in the Case of Differential Offending Rates

There are also problems with looking to algorithmic affirmative action to rectify predictive disparities that flow from differences in underlying rates of crime commission across racial lines. Calls to equalize false-positive and false-negative rates (the disparities that ProPublica identified) serve as a useful case study.[144] There is a practical argument against such interventions and a deeper conceptual one.

a) Practical Problems

The practical argument against intervention to equalize false-positive and -negative rates is that it is not likely to reduce the net burden of predictive regimes on communities of color. To begin with, it may not even be possible to equalize both error rates at once. An effort to equalize false-positive rates may widen the disparity in false-negative rates, or vice versa. Even if it is possible to equalize both error rates, moreover, the intervention is likely to have a substantial cost in accuracy, which means more incorrect predictions—or greater net cost—overall. And this greater net cost may fall disproportionately on black communities.

The first reason that increased error might fall disproportionately on black communities is that equalizing false-positive and false-negative rates does not mean equalizing the total number of errors for each racial group.

[142] As Selbst puts it, "it may be impossible to tell *when* the disparate impact truly reflects reality." Selbst, *supra* note 2, at 141-44; *see also* Barocas & Selbst, *supra* note 2, at 682.

[143] *Cf.* Barocas & Selbst, *supra* note 2; Grant T. Harris & Marnie E. Rice, *Bayes and Base Rates: What Is an Informative Prior for Actuarial Violence Risk Assessment?*, 31 BEHAV. SCI. & L. 103, Selbst, *supra* note 2.

[144] *See* Angwin, *supra* note 1; Huq, *supra* note 13.

Equalizing false-negative rates, rather, means equalizing the *proportion* of rearrests the algorithm misses for each racial group. If the algorithm misses fifty percent of rearrests for each racial group, and there are more rearrests among black defendants to begin with, the algorithm will miss more rearrests of black defendants than white. The difference in the absolute number of false negatives could overwhelm any benefit to communities of color that flows from equalized false-positive rates.[145] Appendix A illustrates this possibility with an example drawn from real data.

The second reason that increased error might disproportionately burden communities of color is that people of color might be overrepresented in the system. Even if the total error rate is lower for black defendants than white, a lower total error *rate* can translate into a much greater absolute number of errors if there are more black defendants in the system. Appendix A illustrates this possibility as well.

This is not to say that equalizing error rates will necessarily increase the net cost of prediction borne by black communities, just that it might. It depends on the underlying base rates and what the false-positive and false-negative rates are. As of yet, though, there is no basis to think that this metric is systematically more likely than any other to equalize the net burden of prediction. If prioritizing equality in error rates has too great a cost in accuracy, moreover, it will eliminate the utility of prediction.[146]

These practical arguments extend to algorithmic affirmative action to achieve statistical parity. Statistical parity requires that, for each racial group subject to assessment, the same proportion of the group must be classified as high-risk and presumptively detained. That will produce a lower false-positive rate for the high-base-rate group than the low-base-rate group. But it will produce a higher false-negative rate for the high-base-rate group and more false negatives for every false positive (that is, the cost ratio of false-positives-to-false-negatives will be low).[147] Depending on what the error rates are and the relative sizes of the black and white groups assessed, this could result in greater net costs for black communities. The same is true for efforts to equalize cost ratios for each racial group. In a recent study by Richard Berk, that form of algorithmic affirmative action increased the disparity in the rate of adverse predictions for each racial group, as well as the disparity in false-positive rates.[148]

---

[145] Equalizing false-positive rates will result in a lower total number of false-positives ("law-abiders" mistakenly forecast for rearrest) for the high-base-rate group than the low-base-rate group (because there are fewer "law-abiders" in the low-base-rate group to begin with).

[146] Sam Corbett-Davies and colleagues, analyzing the same Broward County data that ProPublica did, found that achieving parity in false-positive rates while still optimizing for public safety (and without detaining additional defendants) would result in a seven percent increase in violent crime. Furthermore, seventeen percent of those detained would be low-risk people for whom detention was unwarranted. Corbett-Davies *et al*, *supra* note 67.

[147] In Richard Berk's recent study of juvenile data, for instance, altering the algorithm to achieve statistical parity resulted in a lower false-positive rate for the black subset than the white (4% versus 9%) but a higher false-negative rate (50% versus 40%), and disparate cost ratios (5.25 to 1 versus 1 to 1.85). Berk, *supra* note 107.

[148] When Berk trained the algorithm only to optimize for overall accuracy, it forecast arrest for 17% of the white subgroup and 33% of the black subgroup (a difference of 16%); the false-positive

The point here is straightforward. The goal of algorithmic affirmative action is to reduce the net burden of crime prediction errors on black communities, but it is not likely to do so. If there is a difference in the base rate of the relevant crime across racial lines, distorting the statistical mirror to ignore it will just produce disparate rates of error, which might increase the net burden on the communities the intervention was intended to protect.

b) Conceptual Problems

The fact that its cost in accuracy might outweigh the benefit of algorithmic affirmative action suggests the deeper argument against it: In its essence, algorithmic affirmative action constitutes a rejection of actuarial risk assessment itself.

This argument begins with the very nature of equality. Equality is a formal concept. Peter Westen called it an "empty" one,[149] and many legal theorists find that a bridge too far. But there is widespread agreement that any equality demand—any mandate to treat like cases alike—will necessitate some substantive judgment about what makes two cases relevantly "like" for purposes of the action at hand.[150] Anti-discrimination laws, for instance, frequently require a claimant to show that she was treated differently than someone "similarly situated" in order to make out a prima facie case. To analyze such claims, judges must decide which traits are relevant. For purposes of an employment action, work experience and skill are probably relevant. Two people with equal skill and experience are therefore "similarly situated", and differential treatment of those two people might raise an inference of discrimination. A person's favorite ice cream flavor is likely not relevant; the fact that an employer treats two people differently despite their shared preference for mint chocolate chip does not signal any wrongdoing.

The question of what makes two people (or groups) relevantly "alike" for purposes of some action, moreover, is really a question about the permissible grounds for that action. To judge that skill and experience, but not ice-cream preferences, are relevant to employment decisions is to judge that skill and experience, but not ice-cream preferences, are good grounds on which to make an employment decision. To judge that two people are relevantly "alike" for purposes of a mortgage if they have equal credit scores is to judge that a credit score is a good basis for mortgage lending. Every

rates were 16% for the white subgroup and 28% for the black subgroup (a difference of 12%). When he altered it to equalize the cost ratios, it forecast arrest for 10% of the white subgroup and 29% of the black (a difference of 19%); the false-positive rates were 8% for the white subgroup and 22% for the black (a difference of 14%). Berk, *supra* note 128, at 8.

[149] Peter Westen, *The Empty Idea of Equality*, 95 HARV. L. REV. 537, 547 (1982) ("Equality is an empty vessel with no substantive moral content of its own.").

[150] *E.g.* H.L.A. HART, THE CONCEPT OF LAW 159 (Joseph Raz, & Penelope A. Bulloch eds., 3rd ed. 2012) ("[A]ny set of human beings will resemble each other in some respects and differ from each other in others and, until it is established what resemblances and differences are relevant, 'treat like cases alike' must remain an empty form."); Schauer, *supra* note 47, at 203 ("It is now widely accepted that Aristotle's prescription to treat like cases alike is essentially tautological, or, as Peter Westen puts it, empty.").

judgment about what constitutes unjustified inequality in some decision-making process is also a determination about the legitimate criteria for that decision, and one cannot identify unjustified inequality without choosing, or assuming, some answer to that underlying question.[151]

To pursue equality in statistical risk assessment, it is necessary to specify the appropriate grounds for a risk score, and thus what renders two individuals relevantly alike, such that they should receive the same score. But this is not really up for debate. The very concept of risk assessment presumes an answer: statistical risk is the appropriate basis for statistical risk assessment. Risk assessment is nothing *other* than a statement of statistical risk. Two people are therefore alike for purposes of statistical risk assessment if they present the same statistical risk. This is the conception of equality that Part I.C termed "individual-risk equality."

Because it follows from the nature of the activity, individual-risk equality is a *sine qua non* of risk assessment. If a risk assessment algorithm, when faced with two people who pose precisely the same statistical risk, says "high-risk" in one case and "low-risk" in another, the algorithm is failing in the most basic way. Its statements of risk cannot be meaningful, for they do not reliably state the underlying risk. Whether a given degree of risk is high or low may require a normative judgment, but it cannot coherently be both. This is to say that a mandate of individual-risk equality is a corollary of the very concept of statistical prediction.

A demand for equality in false-positive or false-negative rates corresponds to a different judgment about what renders people relevantly alike. Equality in false-positive rates demands an equal error rate for two groups: black versus white defendants who will *not* actually go on to commit crime—the eventual law-abiders. Equality in false-negative rates demands equality between the black and white groups who *will* go on to commit crime—the eventual law-breakers. Implicit in this equality demand is the judgment that two people or groups are relevantly alike if they have the same eventual outcome. Eventual law-abiders should be treated the same regardless of race. So should eventual law-breakers.

At first blush, this makes sense. It seems fairer to condition treatment on actual events than on mere probabilities. And if the thing we aspire to predict and prevent is crime, surely the actual occurrence of crime must be the best possible measure of risk!

In fact, however, this view is deeply incoherent. To hold that ultimate outcomes are what render two people (or groups) alike for purposes of risk

[151] To appreciate this fact in the context of criminal justice risk assessment, notice that the schema of equality metrics in Part I.C is incomplete. It is possible to create new metrics of equality by subdividing the ones enumerated there. Rather than inquiring about the percentage of black versus white arrestees who are classified as high-risk, for instance (total population impact), one might inquire about the percentage of black versus white *male* arrestees so classified, or the percentage of black versus white male arrestees under 25 who receive that designation, or the percentage of black versus white male arrestees under 25 with a prior felony conviction who do. In fact, there are a nearly infinite number of possible equality metrics. That is because the key question for defining a metric—who are the relevant comparators?—admits of a nearly infinite number of answers. And who one deems to be the relevant comparators depends on what one believes to be a legitimate basis for assigning risk.

assessment is to hold that outcomes are a good basis for risk assessment. But outcomes cannot be the basis for risk assessment, because at the time of assessment they are unknown. This is why we resort to risk assessment in the first place. Even this formulation, moreover, affords outcomes more stability than they have, for not only are outcomes unknown; if chance plays any role in our lives, they are also *unknowable*.

The point is not a technical one. As a technical matter, risk assessment algorithms can be engineered to produce equal false-positive or false-negative rates across racial groups. The point, rather, is conceptual. The demand for equal algorithmic treatment for same-outcome groups equates to the judgment that outcomes are the appropriate basis for prediction. And that judgment is nonsensical.[152]

More concretely, structuring an algorithm to equalize false-positive and false-negative rates will almost certainly violate individual-risk equality. If the base rate of the predicted event differs across racial groups, equalizing false-positive and false-negative rates will likely require setting different risk thresholds by race for each risk classification. It might require, for instance, classifying white defendants as high-risk at a rearrest probability of 15% or above, while classifying black defendants as high-risk only at a probability of 25% or higher. In a scenario like that, a person with a 20% chance of rearrest will be classified as high-risk if he is white but not if he is black. To achieve equality across groups that have not yet come into existence, the algorithm must produce different risk assessments for people who pose the same degree of risk.

It is worth recalling, too, that the very notion of "error" in risk assessment is contested.[153] False positives are the group of people for whom we can say *in retrospect*, evaluating a test run of an algorithm, that they committed no harm. But at the point of assessment we do not know for whom this will be true. All we have is a probability. Even in retrospect, the fact that a risk does not materialize does not mean a high-risk classification was incorrect. Sometimes high risks do not materialize. That is what differentiates risks from certainties.

In sum: To demand equality for same-outcome groups at the cost of equality for same-risk individuals is to reject the project of statistical risk assessment. It precludes risk assessment on the basis of risk. It conditions risk assessment on future outcomes shaped, in part, by chance.

A similar argument applies to statistical parity. Statistical parity requires that the same proportion of each racial group (of people subject to assessment) be classified as high-risk. It presumes that the most relevantly "alike" units are the entire racial groups subject to assessment, such that these groups should be treated alike regardless of statistical differences between

[152] Keep in mind, too, that short of perfect prediction it is not possible for an algorithm to treat every two individuals who will ultimately have the same outcome identically. What equality in conditional procedure accuracy demands is equality across *groups*: black eventual-law-abiders versus white, and white eventual-law-breakers versus black.

[153] *See supra* note 74 and accompanying text.

them. It thus rejects the premise of risk assessment—statistically informed action.

Having read this far, some readers might conclude that this line of argument offers a case in favor of algorithmic affirmative action rather than against it. Yes, equalizing error rates or requiring statistical parity does fundamentally compromise statistical crime prediction. And that, some may feel, is a good thing.

Perhaps these critics are right, and the criminal justice system should get out of the business of crime prediction altogether. There are many grounds on which one might reach that conclusion.[154] The merits of those arguments are beyond the scope of this Article.

But this is the debate we should be having. If we want to reject criminal justice risk assessment, the rejection should be considered and direct, not accomplished obliquely, and perhaps inadvertently, through an equality mandate. Risk assessment constrained to produce equal false-positive and false-negative rates is not really risk assessment. It is race-specific risk-sorting. To undertake that activity under the guise of risk assessment has the potential to do more harm than good. It may actually increase the burden on communities of color, as detailed above. And it might foster deep resentment. Better to engage in a frank debate about whether the disparate racial impact of crime prediction outweighs its benefit.

C. Rejecting Algorithmic Methods

If the question is whether the disparate racial impact of prediction outweighs its benefits, many critics may be inclined to answer yes—and on that basis to advocate the rejection of algorithmic methods. But rejecting algorithmic methods is also a superficial solution, because it is the nature rather than the mode of prediction that makes the disparities inevitable. Any consideration of criminal history as a risk factor, for instance, will entail similar inequality, whether the consideration is actuarial or subjective.[155]

---

[154] Bernard Harcourt, for instance, argues that (1) predictive crime control efforts might do more harm than good; (2) they might produce a "ratchet effect" in which the disparate impact of prediction on black communities compounds over time; and (3) the technical allure of prediction can distort and displace moral conceptions of justice. *See* HARCOURT, *supra* note 28, at 240; *see also* Kristian Lum & William Isaac, *supra* note 93; Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger & Suresh Venkatasubramanian, *Runaway Feedback Loops in Predictive Policing* (2017), arxiv.org/abs/1706.09847; Starr, *supra* note 28, at 804-06 (describing a small experiment suggesting that risk assessment during sentencing may distort judicial perceptions of justice). In addition to these arguments, one might contend that, because any present racial disparity in crime-risk is the product of historical oppression, it is an inappropriate basis for coercive state action: it is unjust for the state to condition coercion on crime-risk that our society has unjustly produced. Alternately, one might believe that the data is simply wrong, and the risk at issue is really uniform across racial lines. Finally, and most profoundly, one might believe that crime-risk is an incoherent concept, because all people who are self-determining agents have an equal capacity to avoid wrongdoing.

[155] *See* MODEL PENAL CODE § 6B.07(1)(c) (Tentative Draft #4 2016), robinainstitute.umn.edu/publications/model-penal-code-sentencing-tentative-draft-no-4 (noting "the danger that the use of criminal-history provisions to increase the severity of sentences may have disparate impacts on racial or ethnic minorities, or other disadvantaged groups"); *id.* § 6B.07(4) (instructing sentencing commissions to "monitor the effects of including criminal history as a sentencing factor," giving

It is true that there are concerns unique to algorithmic methods. Algorithmic assessment carries a scientific aura, which can produce unwarranted deference or a mistaken impression of objectivity.[156] Some algorithms are opaque. Algorithmic systems may be vulnerable to entrenchment, because they require specialized skill and resources to alter. Finally, if algorithmic assessment operates on a much larger scale than subjective assessment, it can inflict damage on a much larger scale.[157] And of course, if algorithmic assessment is imposed *on top of* subjective risk assessment, it is likely to produce additional disparity.

There are also concerns unique to subjective methods, however. Subjective prediction is vulnerable to explicit and implicit bias. Individual judges may generalize to a greater extent, and with less grounding, than sophisticated statistical models.[158] They may harbor animosity toward one racial group that infects their decision-making. In general, subjective risk assessment is far more opaque, and far less accountable, than assessment by algorithm.[159] The human being who judges a person to be a good risk or a bad one may not understand herself why she has done so.[160]

This Article does not take a position on the relative merit of algorithmic versus subjective crime prediction in terms of racial justice. There is no empirical research as of yet comparing their actual racial effects.[161] Everything depends, moreover, on the specifics of the method and the details of its implementation. A thoughtful judge with broad experience may be more effective at assessing risk than a rudimentary algorithm, but a sophisticated algorithm with the benefit of broad data may be more effective than a bad judge, and a good judge operating with the benefit of a good algorithm may be most effective of all.

"particular attention" to whether it "contributes to punishment disparities among racial and ethnic minorities, or other disadvantaged groups"); *id.*,§ 6.07 Commentary, at 90 ("An accumulating body of research indicates that criminal-history formulas in sentencing guidelines are responsible for much of the [ ] disparities in black and white incarceration rates . . . ."); *id.* at 101 (noting that African American defendants appear in criminal courtrooms, on average, with larger numbers of past convictions than white defendants, and citing relevant research).

[156] On the normative judgments that the construction of a risk assessment algorithm entails, *see generally* Eaglin, *supra* note 28.

[157] *See generally* O'NEIL, *supra* note 2 (chronicling and illustrating the dangers of ostensible scientific objectivity, opacity, entrenchment and scale).

[158] *Accord* Starr, *supra* note 28, at 824 ("There is, to be sure, considerable statistical research suggesting that judges (and prosecutors) do on average treat female defendants more leniently than male defendants."); Hamilton, *supra* note 28, at 284–85 ("[I]f constitutionally or ethically suspect variables are excised [from risk assessment tools], it is likely that fact-finders would consider [them] informally anyway, rendering their use less reliable, transparent, and consistent").

[159] *See* Kroll *et al*., *supra* note 131.

[160] Ralph Richard Banks & Richard Thompson Ford, *(How) Does Unconscious Bias Matter? Law, Politics, and Racial Inequality***,** 58 EMORY L.J. 1055 (2009); John A. Bargh, *Unconscious Thought Theory and Its Discontents: A Critique of the Critiques*, 29 SOC. COGNITION 629 (2011); Martie G. Haselton *et al*., *Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias*, 27 SOC. COGNITION 733 (2005).

[161] *Accord* Stevenson, *supra* note 5, at X. Even on the more basic question of comparative accuracy, the jury is still out. *See, e.g.*, Starr, *supra* note 28, at 855 (concluding that "the shibboleth that actuarial prediction outperforms clinical prediction is—like the actuarial risk predictions themselves—a generalization that is not true in every case"); Stevenson, *supra* note 5, at X (surveying existing evidence).

Whatever their relative costs and benefits on other fronts, algorithmic and subjective prediction share a common structure. They look to the past as template of the future. This is the source of the racial disparity that prediction entails. And because this is the source of the problem, rejecting actuarial methods does not solve it.

The new contribution that actuarial risk assessment does make is to illuminate—in formal, quantitative terms—the way in which prediction replicates and magnifies inequality in the world. Statistical prediction holds an especially precise mirror to the past. So long as it seeks to predict an event that, in the past, occurred more frequently in communities of color, any decent algorithm will predict that event more frequently for people of color in future. Tweaking the input data or the algorithm itself will not solve the problem. Nor will rejecting algorithmic methods, because the problem is inherent to prediction itself. The predictive inequality exposed by algorithmic methods should, instead, cause us to rethink a central strategy in contemporary U.S. criminal justice: identification and coercive control of the "dangerous."

IV. RETHINKING RISK

What algorithmic prediction makes painfully explicit are the racial fault lines in the risk-management model that has come to dominate criminal justice. In 1992, Malcolm Feeley and Jonathan Simon diagnosed the "New Penology," a shift in the orientation of the U.S. criminal justice system.[162] The "Old Penology" saw the primary goal and responsibility of the criminal justice system as the adjudication of guilt for specific criminal acts. The New Penology sees the system's primary goal and responsibility as the management of "dangerous groups."[163] Many others have since expanded on the diagnosis.[164] Scholars have long argued that a criminal justice system designed to incapacitate the risky will perpetuate racial injustice. Actuarial analytics illustrate precisely how.

One response is to refute the significance of risk itself, to lament the New Penology and argue for a return to the Old. Plenty of scholars do. Whether or not that would be the best outcome, it is a very unlikely one. This Article does not pursue it.

The other possible response is to accept the significance of risk to criminal justice decision-making but to nudge the system toward a more

[162] Malcolm M. Feeley & Jonathan Simon, *The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications*, 30 CRIMINOLOGY 449 (1992).

[163] *Id.* at 449.

[164] *See, e.g.*, Jennifer C. Daskal, *Pre-Crime Restraints: The Explosion of Targeted, Noncustodial Prevention*, 99 CORNELL L. REV. 327 (2014); Eisha Jain, *Arrests as Regulation*, 67 STAN. L. REV. 809 (2015); Issa Kohler-Hausmann, *Managerial Justice and Mass Misdemeanors*, 66 STAN. L. REV. 611 (2014); Sandra G. Mayson, *Collateral Consequences and the Preventive State*, 91 NOTRE DAME L. REV. 301, 348 (2015); Erin Murphy, *Paradigms of Restraint*, 57 DUKE L.J. 1321, 1405–06 (2008); Carol S. Steiker, *Foreword: The Limits of the Preventive State*, 88 J. CRIM. L. & CRIMINOLOGY 771, 774 (1998) (describing the constellation of government efforts to incapacitate the dangerous as "the preventive state").

thoughtful treatment of it. Given the inequality inherent in all crime prediction, what risk really matters? When there is such risk, how should the system respond? Does the answer change if we separate judgments of risk from judgments of blame? Engagement with predictive inequality requires engagement with these deep questions. This Article cannot answer them conclusively—if indeed they can be answered—but the remainder of the discussion argues for four key steps toward a more rational approach to risk.

A. Risk of What?[165]

As an initial matter, we should simply stop treating the likelihood of any arrest as a meaningful measure of risk, whether by actuarial or subjective methods. The average arrest offense is too insignificant to have much probative value, and the racial skew in arrest rates vis-à-vis offending rates is too prejudicial. This point has been argued elsewhere,[166] so it warrants no further elaboration here.

Risk assessment should, instead, be limited to assessing the risk of serious harm. This may not always be possible to do. The Pennsylvania Sentencing Commission, for instance, recently and to its credit, concluded that it could not predict future violence with sufficient accuracy to justify handing risk scores to judges.[167] When that is the case, we should not resort to predicting a more statistically significant event, but should simply recognize that our objectives exceed our ability.

B. A Supportive Response to Risk

The hardest problem arises if a predictive algorithm does predict some category of serious crime with sufficient accuracy, and reflects a difference in the base rate of that crime across racial lines. There are no easy answers to this problem. Distorting the predictive mirror or tossing it out does nothing to fix the disparities it reflects. The problem is not in the algorithm; the problem is on the ground. Solutions must target ground-level conditions too.

What if support, rather than jail, were the default response to risk? Risk, after all, is neither intrinsic nor immutable. It is possible to change the odds.[168] A supportive, needs-oriented response to risk might help to change the odds for high-risk groups in the long term. In the short term, it would

---

[165] For a thoughtful discussion of this question in the pretrial context, *see* Schnacke, *supra* note 111, at 109-14.

[166] *E.g.* Roberts, *supra* note 41; Mayson, *supra* note 5, at 562; Stevenson & Mayson, *supra* note 40, at 28-30; Slobogin, *supra* note 8, at 591; Schnacke, *supra* note 111, at 109-14.

[167] Pennsylvania Comm'n on Sentencing, Development and Validation of the Risk Assessment Scale (May 2018), available at http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment.

[168] *Cf.* Patrick Sharkey, Gerard Torrats-Espinosa & Delaram Takyar, *Community and the Crime Decline: The Causal Effect of Local Nonprofits on Violent Crime*, 82. AM. SOC. REV. 1214, 1234 (2017) (estimating that "the addition of 10 community nonprofits per 100,000 residents leads to a 9 percent decline in the murder rat, a 6 percent decline in the violent crime rate, and a 4 percent decline in the property crime rate.").

mitigate the immediate racial impact of prediction. If a high-risk classification meant greater access to social services and employment, a higher false-positive rate among black defendants would be less of a concern.

This proposal is not original. As a logical matter, it is what the "least-restrictive-means" principle encoded in many risk-management systems requires; an offer of support is certainly less restrictive than monitoring or detention. Pretrial and sentencing laws generally include some version of the least-restrictive-means principle.[169] A supportive response to risk is also built into the "risk-needs" model prevalent in more mature risk-management systems.[170] As a conceptual matter, scholars who study algorithmic fairness arrive at the same recommendation in other contexts.[171] Although it is true that algorithmic methodology poses some unique dangers,[172] whether it exacerbates or mitigates social inequality is entirely a function of the use to which it is put. In the aggregate, after all, crime-risk of the kind that contemporary criminal justice risk assessment tools measure—"any arrest," or arrest for a "violent crime"—is a function of disadvantage. If algorithms targeted the disadvantaged for support rather than further disadvantage, their effects in the world would be very different.

Nor does a supportive response to risk amount to coddling criminals. It does not diminish the state's authority to punish. Risk assessment is designed not to determine just punishment, but rather to evaluate risk in order to manage it. There is no reason that risk management should exclude support. The idea that the state must extend nothing other than condemnation to criminal-justice involved people runs contrary to law, as well as to the ideals of our criminal justice system.

Lastly, a default supportive response to risk need not mean obliviousness to danger. We know very little about what risk management strategies are most effective in run-of-the-mill cases. Meaningful support has just as much promise as electronic monitoring. For those who pose an acute threat to an identifiable person or group, the default could yield. Support for the many does not preclude preventive restraint, even detention, for a few.

Certainly a shift toward a default supportive response to risk would present a practical and political challenge. The ascendant policing model known as "focused deterrence" offers a cautionary tale. The model directs police to focus on a small number of people most likely to be involved in violent crime (as either perpetrator or victim). In concept, the model requires

[169] *See, e.g.*, 18 U.S.C. § 3533 (2012) (requiring judges to impose a sentence that is "sufficient but not greater than necessary" to accomplish goals of punishment); Richard S. Frase, *Sentencing Principles in Theory and Practice*, 22 CRIME & JUST. 363, 375-78 (1997) (explaining "parsimony principle"); Am. Bar Assoc., Standards for Pretrial Release § 10-1.2 (providing that "[i]n deciding pretrial release, the judicial officer should assign the least restrictive condition(s) of release that will reasonably ensure a defendant's attendance at court proceedings and protect the community, victims, witnesses or any other person").

[170] *See, e.g.*, CORRECTIONS IN ONTARIO: DIRECTIONS FOR REFORM 110 (2017); Robin J. Wilson, Franca Cortoni, Andrew J. McWhinnie, 21 SEXUAL ABUSE J. RES. & TREATMENT 412 (2009).

[171] *See, e.g.*, O'Neil, *supra* note 2.

[172] Because of its capacity for scale, tendency toward opacity (especially in the private sector), and veneer of scientific objectivity. O'Neil, *supra* note 2; Noble, *supra* note 2; Barocas & Selbst, *supra* note 2.

police to both offer a carrot—increased social support—and threaten a stick—increased punishment for even small criminal infractions—to those targeted. In practice, the carrot tends to get lost.[173] Criminal justice system actors, for the most part, are not trained as social workers. A good-guy-bad-guy mentality pervades the system. Changing the default response to risk would require overcoming these institutional and cultural barriers.

But a shift in the way the system responds to risk is achievable over the long run. There are signs, in fact, that such a shift might be underway. For decades, legislatures bought political capital by codifying employment barriers and other civil disabilities for people with past convictions, which they justified as public safety measures. In the first five months of 2018, by contrast, twenty-one states enacted laws to improve opportunities for people with criminal records.[174] Even President Trump has signed on to the "second chance" agenda. In the criminal justice system itself, supportive reentry and "preentry" programs are gaining traction. And some risk assessment tool developers have begun to disclaim the idea that a risk score alone can justify increased restraint.[175] This is not the same as targeting at-risk people for support, but it is a step in the right direction.

C. Algorithmic Prediction as Diagnostic

Counterintuitively, actuarial risk assessment could be a valuable tool in any effort to redress racial disparities in prediction through a supportive response to risk.[176] Because they transparently reflect inequality in the data from which they are built, predictive algorithms can be deployed in reverse, as diagnostic tools. We can use risk statistics to diagnose and understand racial disparities in past arrest and crime rates. We can confront the image in the mirror and take responsibility for it. In other arenas, data scientists are working to deploy machine-learning to similar diagnostic ends. James Zou and colleagues, for instance, are using machine-learning to identify adjectives

[173] *See, e.g.*, Selbst, *supra* note 2, at 142-43 (noting that early evidence on the program's implementation in Chicago suggests that the support didn't happen); Susan Turner, Terry Fain & Amber Sehgal, *Validation of the Risk and Resiliency Assessment Tool for Juveniles in the Los Angeles County, Probation System*, RAND (2005), https://www.rand.org/content/dam/rand/pubs/technical_reports/2005/RAND_TR291.pdf.

[174] CCRC Staff, *More States Enact "Second Chance" Reforms*, COLLATERAL CONSEQUENCES RESOURCES CENTER (June 11, 2018), http://ccresourcecenter.org/2018/06/11/three-more-states-enact-major-second-chance-reforms/.

[175] The developers of the proposed Pennsylvania Risk Assessment Tool for sentencing, for instance, intend for the tool to be used only to identify people for whom an in-depth presentencing report is indicated. Pennsylvania Comm'n on Sentencing, Development and Validation of the Risk Assessment Scale (May 2018), available at http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment.

[176] This is not to endorse actuarial assessment overall. I take no position on the relative merit of actuarial versus subjective methodology. Each has dangers; each has advantages; the context and specifics of the tool matter enormously. *See supra* Part III.C. But given a choice between two methods of assessing crime risk that will likely entail equivalent racial disparity, it is worth recognizing the advantages that actuarial tools do have.

most frequently associated with different ethnic groups over the nineteenth and twentieth centuries to illuminate the history of discrimination.[177]

It is also possible to hold algorithms accountable for their calculations and outputs in a way that it is not possible to hold humans accountable for their mental decision-making process.[178] We can quantify an algorithm's racial impact, and demand that its predictions fulfill whatever measure of output equality we choose. Scholars and stakeholders have begun to elaborate the procedural and legal regimes necessary for this kind of accountability.[179] There are hurdles, of course, but the accountability prospects are far better for algorithmic than for subjective prediction. More than thirty years ago, Noval Morris and Marc Miller, arguing for a frank reckoning with the costs and benefits of preventive detention, wrote: "We propose to get the dragon out onto the plain."[180] Algorithmic prediction puts the dragon of racial inequality out on the plain. It is frightful, but at least we can see it.

To serve as a diagnostic tool of any kind, risk assessment must function with equal integrity across racial lines. This requires that it meet three metrics of predictive equality. First, individuals who pose the same statistical risk should receive the same risk score regardless of race ("individual-risk equality"). Second, a given risk score should communicate the same average risk regardless of the race of the person to whom it applies ("predictive parity").[181] Third, a predictive algorithm should order individuals along a spectrum of risk with equal accuracy for each racial group (have equivalent AUC-ROC scores by race).

The first two metrics may sound similar, but they are not co-extensive. Assigning the same risk score to all those who present the same risk will not necessarily produce predictive parity,[182] and an algorithm might achieve predictive parity without assigning the same risk score to all who present the same risk.[183] But individual-risk equality and predictive parity are conceptually related in that both require that the relationship between a risk score and risk itself be constant across racial groups. Individual-risk equality requires that the algorithm consistently translate a given degree of individual risk into the same risk score regardless of race, and predictive parity requires that a given risk score consistently express the same average risk regardless

[177] Nikhil Garg, Londa Schiebinger, Dan Jurafsky & James Zou, *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*, 115 PROC. NAT'L ACAD. SCI. U.S.A. E3635 (2018), http://www.pnas.org/content/115/16/E3635.full.

[178] *See, e.g.*, Kroll *et al.*, *supra* note 131.

[179] Kroll *et al.*, *supra* note 131; Selbst, *supra* note 2, at 169-180 (proposing "algorithmic impact statements" that "would require police departments to evaluate the efficacy and potential discriminatory effects of all available choices for predictive policing technologies"); Dillon Reisman *et al.*, *Algorithmic Impact Assessments*, AINOW (2018) ainowinstitute.org/aiareport2018.pdf; The Dataset Nutrition Label Project, http://datanutrition.media.mit.edu (proposing that datasets be required to include the equivalent of "nutrition labels" that disclose possible demographic skews or systemic inaccuracies in the data).

[180] Norval Morris & Marc Miller, *Predictions of Dangerousness*, 6 CRIME & JUST. 1 (1985).

[181] In other words, the statistical meaning of the score itself must not vary by race.

[182] If the risk class is broad—encompasses anyone who poses between a 20% and 99% chance of rearrest, for instance—and the distribution of risk within the class is different across racial groups.

[183] Corbett-Davies *et al*., *supra* note 67.

of race. Both are achievable, furthermore, even if base rates of the predicted outcome differ across racial lines.

None of these equity metrics, nor all of them in combination, render an algorithm race-neutral. On the contrary, achieving them may require race-conscious choices in the construction of the algorithm. Moreover, if the base rate of the predicted outcome differs across racial groups, an algorithm that achieves these equality metrics will also produce unequal false-positive and/or false-negative rates. But that disparate impact, as should now be clear, is an inevitable product of prediction.

CONCLUSION

On June 6, 2018, the Pennsylvania Sentencing Commission held a public hearing on the proposed new Pennsylvania Risk Assessment Tool for sentencing.[184] The room was packed. One by one, community members walked to the lectern and delivered passionate pleas against adoption of the tool. They argued that reliance on criminal history factors would have disparate impact, and that the likelihood of arrest is an artifact of racially skewed law enforcement rather than a meaningful measure of risk. Several speakers wondered why the system is so fixated on risk—on the prospect of failure—in the first place. Instead, they argued, it should direct its efforts to improving people's prospects for success.

The speakers at that meeting offered a profound critique . . . of *all* state coercion on the basis of risk. Some of their concerns were indeed specific to algorithmic methods and to the proposed Pennsylvania tool. But the deepest concerns of the community, the sources of its deepest outrage, applied equally to the subjective risk assessment that already pervades the criminal justice system.

What algorithmic methods have done is reveal the racial inequality that inheres in all forms of risk assessment, actuarial and subjective alike. Neither colorblindness, nor algorithmic affirmative action, nor outright rejection of actuarial methods will solve the underlying problem. So long as crime and arrest rates are unequal across racial lines, any method of assessing crime- or arrest-risk will produce racial disparity. The only way to redress the racial inequality inherent in prediction in a racially unequal world is to rethink the way in which contemporary criminal justice systems conceive of and respond to risk.

The analysis of racial inequality in criminal justice risk assessment also serves as a case study for broader questions of algorithmic fairness. The important distinction between the two possible sources of inter-group disparity in prediction—distortions in the data versus differential base rates of the event of concern—applies in any predictive context, as does the taxonomy of equality metrics. But the types of distortions that affect the data

[184] *See Proposed Risk Assessment Instrument* (Pennsylvania Commission on Sentencing), http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument (last visited Aug. 29, 2018).

or algorithmic process will differ by context.[185] So will the analysis of what equality metric(s) it makes sense to prioritize. This is because the right equality metric depends on the relevant basis for the action at issue. When an algorithm's very purpose is to accurately communicate statistical risk under status quo conditions, statistical risk is the only relevant basis for its action, such that two people who pose the same statistical risk must be treated alike. But in other contexts algorithms might have other purposes. Algorithms used to allocate loans, housing or educational opportunity might have distributional goals.[186] Algorithms that drive internet search engines might be programmed to maximize the credibility of top results or minimize representational harms.[187] Algorithms used to calculate lost-earnings damages in wrongful death suits should perhaps have objectives other than reflecting status quo earning patterns.[188] Not all algorithms should faithfully mirror the past.

The next few years will set the course of criminal justice risk assessment. To demand race-neutrality of tools that can only function by reflecting a racially unequal past is to demand the impossible. To reject algorithms in favor of subjective prediction is to discard the clear mirror for a cloudy one. The only sustainable path to predictive equity is a long-term effort to eliminate the societal inequality that the predictive mirror reflects. That path should include radical revision in how the criminal justice system understands and responds to crime-risk. And there is an opportunity now, with risk assessment and race in the public eye, to take it.

[185] It may be even more challenging in other arenas to find a target variable that doesn't encode racial skewing vis-à-vis the actual outcome of concern. In the employment context, for instance, employers want to predict success on the job. But the data on past success may be skewed by the company's past discrimination in hiring or promotion practices. There is nothing in the past data that reliably represents "job success" in a non-discriminatory environment.

[186] Corbett-Davies et al., *supra* note 67, at 9 (citing SCOTT E PAGE, THE DIFFERENCE: HOW THE POWER OF DIVERSITY CREATES BETTER GROUPS, FIRMS, SCHOOLS, AND SOCIETIES (2008)).

[187] See Noble, *supra* note 2.

[188] *See* Kimberly A. Yuracko & Ronen Avraham, *Valuing Black Lives: A Constitutional Challenge to the Use of Race-Based Tables in Calculating Tort Damages*, 106 CAL. L. REV. 325, 330 (2018).

APPENDIX A: THE PRACTICAL CASE AGAINST AAA – AN ILLUSTRATION

This Appendix offers further explanation of how equalizing false-positive and false-negative rates might increase the net burden of prediction on communities of color. Consider the following example.

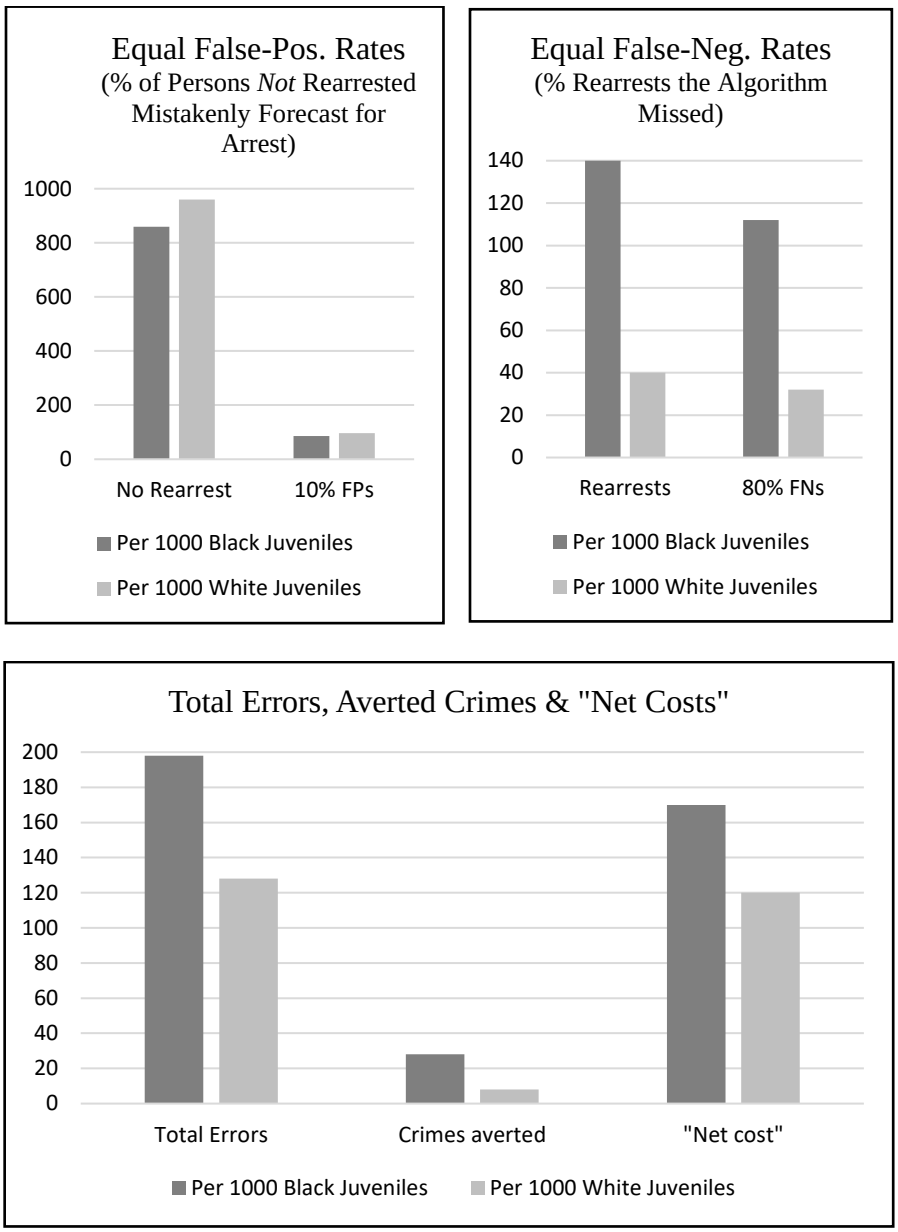
In the juvenile justice data recently examined by Richard Berk, there was a higher base rate of rearrest for violent crime among the black juveniles in the dataset than among the white. For every thousand white juveniles, 40 were rearrested and 960 were not. For every thousand black juveniles, 140 were rearrested and 860 were not. Say the false-positive rate (proportion of eventual non-rearrestees mistakenly forecast for rearrest) is ten percent for each group. For every thousand white juveniles, 96 (of the 960) non-rearrestees will be mistakenly forecast for arrest. For every thousand black juveniles, 86 (of the 860) non-rearrestees will be mistakenly forecast for arrest. Equal false-positive rates means fewer false positives per capita for black juveniles, because there are fewer non-rearrestees to start with.

But what if the false-negative rate (proportion of eventual rearrests the algorithm misses) is eighty percent for each group? Then the algorithm will miss 112 (of the 140) rearrests per thousand black juveniles, but only 32 (of the 40) rearrests per thousand white juveniles. Equal false-negative rates means many more false negatives per capita for the black juveniles, because there are many more rearrests to begin with. The difference in the total number of false negatives swamps the difference in the total number of false positives across racial groups. Altogether, there will be 128 errors for every thousand white kids and 198 for every thousand black kids. The overall error rate for black juveniles will be significantly higher.

Now, the algorithm also produces greater per capita benefit for black communities, because it successfully predicts a greater number of the black juvenile rearrests.[189] Nonetheless, the greater total error rate overwhelms the greater benefit. The result is a higher net cost to black communities. The following charts illustrate.

[189] This is on the assumption that violent-crime arrest corresponds to violent crime, and that violent crime is intra-racial.

Figure 4: High False-Negative Rates Can Produce Unequal "Net Costs"



The second reason that the increased net cost of a less accurate algorithm could fall disproportionately on black communities is that there might be more black people in the system than white. The example above assumed that there were equal numbers of black and white kids in the dataset. But suppose there are twice as many black kids arrested as white. In that case, the disparity in total errors and net costs will be doubled. In fact, even if the false-negative rates are low and the false-positive rates are high, such that the algorithm produces fewer per capita errors and lower per capita net cost for black people, it might *still* produce dramatically more errors in absolute terms, and have a greater net cost overall, for black communities. The

following chart shows the results if false-negative rates are equalized at 10%, false-positive rates are equalized at 40%, and there are twice as many black kids in the system as white.

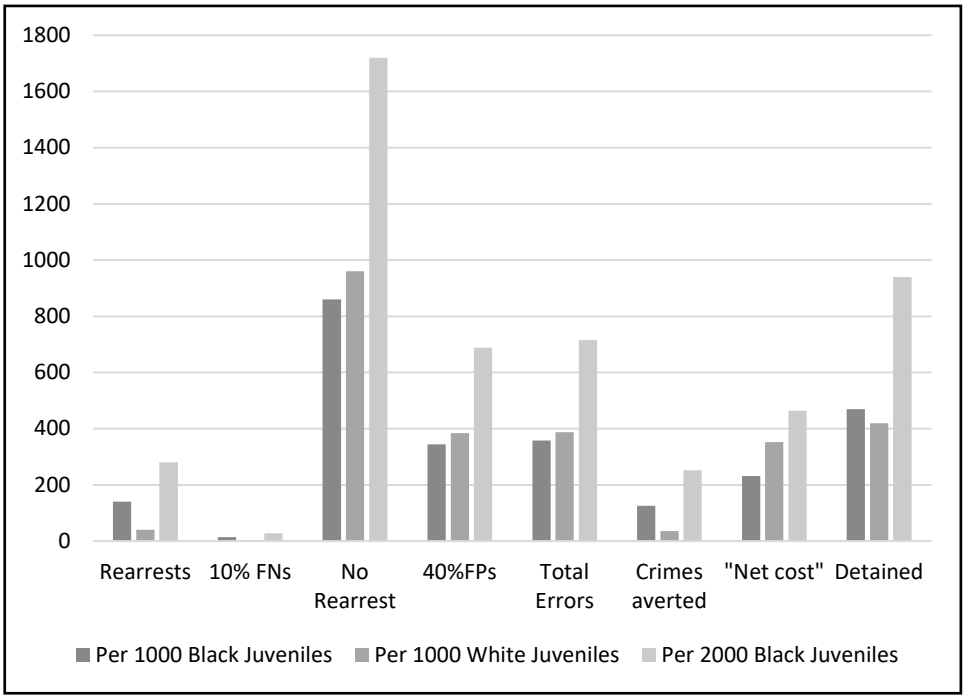


Figure 5: Even with Lower Per Capita "Net Costs" for Black Communities, Disparate Population Sizes Can Produce Unequal "Net Costs"

Lastly: If prioritizing equality in error rates has too great a cost in accuracy, it will eliminate the utility of prediction. Note that, in the second example above, the 40% false-positive rate means that almost half of those who will not be rearrested are misclassified, and the detention rate (if those forecast for arrest are detained) is nearly half of the entire assessed population.