2018

Bias In, Bias Out

Sandra G. Mayson
University of Pennsylvania Carey Law School

THE YALE LAW JOURNAL

SANDRA G. MAYSON

Bias In, Bias Out

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 impacts. 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 per se. 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 it is of the algorithmic tools now replacing it. Algorithmic risk assessment has revealed the inequality inherent in all prediction, forcing us to confront a problem much larger than the challenges of a new technology. Algorithms, in short, 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. Further, to the extent that we can reliably assess risk, criminal system actors should strive whenever possible to respond to risk with support rather than restraint. Counterintuitively, algorithmic risk assessment could be a valuable tool in a system that supports the risky.

AUTHOR. Assistant Professor of Law, University of Georgia School of Law. I am grateful for extremely helpful input from David Ball, Mehrsa Baradaran, Solon Barocas, Richard Berk, Stephanie Bornstein, Kiel Brennan-Marquez, Bennett Capers, Nathan Chapman, Andrea Dennis, Sue Ferrere, Melissa Hamilton, Deborah Hellman, Sean Hill, Mark Houldin, Aziz Huq, Gerry Leonard, Kay Levine, Truman Morrison, Anna Roberts, Bo Rutledge, Hannah Sassaman, Tim Schnacke, Andrew Selbst, Megan Stevenson, Lauren Sudeall, and Stephanie Wykstra; for thoughtful comments from fellow participants in the 2017 Southeastern Junior/Senior Faculty Workshop, CrimFest 2017 and 2018, and the 2017 and 2018 UGA-Emory Faculty Workshops; for invaluable research support from T.J. Striepe, Associate Director for Research Services at UGA Law; and for extraordinary editorial assistance by the members of the *Yale Law Journal*, especially Yasin Hegazy and Luis Calderón Gómez. Title credit to Maron Deering, way back in 2016.

ARTICLE CONTENTS

INTRODUCTION

"There's software used across the country to predict future criminals. And it's biased against blacks."[1] So proclaimed an exposé by the news outlet ProPublica in the summer of 2016. The story focused on a particular algorithmic tool, COMPAS,[2] 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 such programs will drive our cars, but for now 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 is human judgment. It eliminates the irrational biases that skew so much of our decision-making. But it has become abundantly clear that machines too can discriminate.[3] 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 spread 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.[4] The concern with future crime is not new; police, prosecutors, judges, probation officers, and parole officers have long been tasked with making subjective determinations of dangerousness. The recent shift is from subjective

1. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/4G83-MDAS].

2. An acronym for Correctional Offender Management Profiling for Alternative Sanctions. *Id.*

3. *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2017); SAFIYA UMOJA 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 CALIF. L. REV. 671 (2016) (discussing the role of bias in data and what can be done about it); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017) (proposing that police should make "algorithmic impact statements" to assess the potential discriminatory impact of predictive policing technologies).

4. Most risk-assessment tools, however, do not actually measure the likelihood of future crime commission but instead measure the likelihood of future *arrest*, which is a poor proxy. *See infra* Section II.B.

to actuarial assessment.[5] 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 linchpin of the bail-reform movement;[6] the cutting edge of policing;[7] and increasingly used in charging,[8] sentencing,[9] and allocating supervision resources.[10]

This development has sparked profound concern about the racial impact of risk assessment.[11] 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.[12]

---

5. *See, e.g.*, Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59, 61 n.1 (2017) (noting that "[p]redictive technologies are spreading through the criminal justice system like wildfire" and citing scholarship cataloging the spread). This is not to say that actuarial risk assessment is entirely new to criminal justice. Parole boards have used risk-assessment instruments since the 1930s, and some jurisdictions have relied on algorithms for predicting certain kinds of offenses, like sex offenses, for decades past. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 7-18 (2007).

6. *See, e.g.*, Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 490 (2018) ("Bail reform is gaining momentum nationwide."); Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303 (2018) (studying the use of pretrial risk assessment as a mandatory component of bail decisions in Kentucky); Shaila Dewan, *Judges Replacing Conjecture with Formula for Bail*, N.Y. TIMES (June 26, 2015), https://www.nytimes.com/2015/06/27/us/turning-the-granting-of-bail-into-a-science.html [https://perma.cc/Y86J-GHU4] (highlighting growing support for algorithmic risk assessments in bail decision-making).

7. *See, e.g.*, Selbst, *supra* note 3, at 113 ("[P]redictive policing [is] a popular and growing method for police departments to prevent or solve crimes."); Letter from Jonathan J. Wroblewski, Dir., Office of Policy & Legislation, to Hon. Patti B. Saris, Chair, U.S. Sentencing Comm'n 2 (July 29, 2014) [hereinafter DOJ Letter to U.S.S.C.] ("Predictive Policing—the use of algorithms that combine historical and up-to-the-minute crime information to do the work of hundreds of traditional crime analysts and produce real-time targeted patrol areas—is spreading.").

8. *See generally* Andrew Guthrie Ferguson, *Predictive Prosecution*, 51 WAKE FOREST L. REV. 705, 705-08 (2016) (explaining "predictive prosecution" and exploring its "promise and perils").

9. *See, e.g.*, Erin Collins, *Punishing Risk*, 107 GEO. L.J. 57 (2018) (critically assessing the rise of actuarial sentencing); Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583 (2018) (proposing principles for how risk-assessment tools should be used, particularly in the sentencing context).

10. *Issue Brief: Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders*, PEW CTR. ON STATES 2 (Sept. 2011), https://www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2011/pewriskassessmentbriefpdf.pdf [https://perma.cc/38CG-D395] (describing the use of risk assessment to allocate supervision resources).

11. *See infra* Section I.A.

12. Bernard E. Harcourt, *Risk as a Proxy for Race: The Dangers of Risk Assessment*, 27 FED. SENT'G REP. 237, 237 (2015).

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 in the predictive process to ensure racial equity. Yet this raises the difficult question of what racial equity looks like. The challenge is that there are many possible metrics of racial equity in statistical prediction, and some of them are mutually exclusive.[13] The law provides no useful guidance about which to prioritize.[14] In the void, data scientists are exploring different statistical measures of equality and different technical methods to achieve them.[15] Legal scholars have also begun to weigh in.[16] Outside the ivory tower, this debate is happening in courts,[17] city-council chambers,[18] and community meetings.[19] The stakes are real. Criminal justice institutions must decide whether to adopt risk-assessment tools and, if so, what measure of equality to demand that those tools fulfill. They are making these decisions even as this Article goes to print.[20]

---

13. *See infra* Section I.B.

14. Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1053, 1083-1102 (2019) (explaining why constitutional law "provides no creditable guidance" for pursuing racial equity in risk assessment).

15. *See infra* Section I.C.

16. *See, e.g.*, Huq, *supra* note 14, at 1123-33.

17. *E.g.*, State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (finding the use of risk assessment at sentencing constitutionally permissible).

18. *E.g.*, *Interim Report - Fall 2016: A Shift from Re-Entry to Pre-Entry*, PHILA. CITY COUNCIL SPECIAL COMMITTEE ON CRIM. JUST. REFORM 12 (Fall 2016), http://phlcouncil.com/wp-content/uploads/2016/11/SCFall2016InterimReport.pdf [https://perma.cc/W5HX-D9PY] ("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.").

19. *E.g.*, Chris Palmer & Claudia Irizarry-Aponte, *Dozens of Speakers at Hearing Assail Pa. Plan to Use Algorithm in Sentencing*, PHILA. INQUIRER (June 6, 2018), https://www.philly.com/philly/news/crime/philadelphia-pennsylvania-algorithm-sentencing-public-hearing-20180606.html [https://perma.cc/P7R4-C8R2].

20. The Pennsylvania Commission on Sentencing, for instance, held five public hearings on its proposed risk assessment tool in December 2018 at which it encountered considerable opposition. To the author's knowledge, the Commission has not yet determined how to proceed as this Article goes to press. *See Proposed Sentence Risk Assessment Instrument*, PA. COMMISSION ON SENT'G, http://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument [https://perma.cc/CB9Y-TZ6Y] (providing link to testimony received at public hearings); *Risk Assessment*, PA. COMMISSION ON SENT'G, https://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment [https://perma.cc/65PV-T74T] (collecting information relating to the Commission's project to develop a risk-assessment tool with public input).

Among racial-justice advocates engaged in the debate, a few common themes have emerged.[21] 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"). Lastly, critics argue that, if algorithmic risk assessment cannot be made meaningfully race neutral, the criminal justice system must reject algorithmic methods altogether.[22]

This Article contends that these three strategies—colorblindness, efforts to equalize predictive outputs by race, and the rejection of algorithmic methods—are at best inadequate, and at worst counterproductive, 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. Predictive analysis, in effect, holds a mirror to the past. It distills patterns in past data and interprets them as projections of 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."[23] To be more specific, if the thing that we undertake to predict—say arrest—happened more frequently to black people than to white people in the past data, then a predictive analysis will project it to happen more frequently to black people than to white people 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 frequently for people of color.

---

21. *See infra* Part III.

22. Aziz Huq offers a more nuanced set of prescriptions, but his analysis is addressed to equity in the allocation of coercion rather than equity in the assessment of risk per se. Huq, *supra* note 14, at 1111-12. His prescriptions and mine might be read as complementary. *See infra* note 274 and accompanying text.

23. 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.

Limiting input data cannot eliminate the disparity without compromising 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, arrest rates are racially distorted relative to offending rates for any given category of 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 that it reflects neither the data nor any demonstrable reality. Along similar lines, calls to equalize adverse predictions across racial lines require an algorithm that forsakes 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.

This 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, bending it to our liking 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.

Actuarial risk assessment, in other words, has revealed the racial inequality inherent in *all* crime 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.[24] To start, we should be more thoughtful about what we want to learn from the past, and more honest about what we *can* learn from it. 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, then there is no basis for predicting serious crime 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 fairly represent the incidence of serious crime, however, the place to redress racial disparity is not in the measurement of risk, but in the *response* to

24. *See infra* Part IV.

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 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 for restraint.

This 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.[25] The Article offers an accessible taxonomy of potential measures of equality in prediction, synthesizing recent work in computer science with legal-equality constructs. The second contribution is a 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.

Although this Article is about criminal justice risk assessment, it also offers 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 these larger connections.

A few 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.[26] 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 the purpose 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 by necessity, to explain competing metrics of equality with as much clarity as pos-

25. A handful of seminal articles, however, have helped to bridge the gap. *See generally* Barocas & Selbst, *supra* note 3; Huq, *supra* note 14; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Selbst, *supra* note 3.

26. Stevenson, *supra* note 6.

sible, and in part to recognize that the criminal justice system itself tends to deploy this reductive schema. The reader may judge whether this approach is warranted.

It is important to note, though, that much of the Article's analysis generalizes to other minority groups. Although the criminal-legal apparatus has inflicted unique harm on African Americans over the past two hundred years, the data that generate predictions may also include disparities with respect to other groups, and this data will in turn produce predictive inequality. The manifold equality metrics presented in Section I.C apply to *any* intergroup comparison, as do the trade-offs among them. And there is every reason to be concerned about predictive disparities for other marginalized populations. Melissa Hamilton has recently shown that the very same prediction data set that ProPublica analyzed for black/white disparities manifests even greater disparities between Hispanic and white defendants.[27] As the debate on equality in algorithmic prediction evolves, the analysis here is meant to serve as a template with broader applications.

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 taxonomy of potential metrics of predictive equality. Part II lays out the 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 past-crime data relative to crime rates, and a difference in crime rates by race. Accounting for both, Part III explains why the prescriptions for racial equity that currently dominate the 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*,[28] of interest to major mainstream media

27. Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. (forthcoming 2019).

28. The Daily Show with Trevor Noah, *Disrupting the Legal System with Robots*, YOUTUBE (Mar. 7, 2018), https://youtu.be/VkizYljxcD8.

outlets,[29] and the subject of a vibrant and growing body of scholarship.[30] 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-learning algorithms that represent the future.[31] It does not include clinical assessment or instruments used for "structured professional judgment" (SPJ).[32]

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";[33] legal

29. *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), https://fivethirtyeight.com/features/prison-reform-risk-assessment [https://perma.cc/9UP9-U86D] (employing simulations to demonstrate risk-assessment outcomes and disparate racial impact); Dewan, *supra* note 6.

30. *See, e.g.*, Collins, *supra* note 9; 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); Harcourt, *supra* note 12; Huq, *supra* note 14; John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725 (2018); Mayson, *supra* note 6; Anne Milgram et al., *Pretrial Risk Assessment: Improving Public Safety and Fairness in Pretrial Decision Making*, 27 FED. SENT'G REP. 216 (2015); John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CLINICAL PSYCHOL. 489 (2016); Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671 (2015); Slobogin, *supra* note 9; Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014); Stevenson, *supra* note 6; Deborah Hellman, Measures of Algorithmic Fairness (Jan. 25, 2019) (unpublished manuscript) (on file with author).

31. For a brief explanation of the difference, see Mayson, *supra* note 6, at 509-11, 511 n.97. *See also* Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222 (2015); Marion Oswald et al., *Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART model and 'Experimental' Proportionality*, 27 INFO. & COMM. TECH. L. 223 (2018) (describing a machine-learning algorithm); *Assessing Offender Risk*, CTR. FOR SCI. & L., http://scilaw.org/risk-assessment [https://perma.cc/6K9U-DSA9] (promoting sophisticated risk-assessment software that operates on hand-held tablets).

32. *See, e.g.*, Chris Baird, *Structured Professional Judgment Models,* NAT'L COUNCIL ON CRIME & DELINQ. (2017), https://www.nccdglobal.org/sites/default/files/structured_professional_judgment_models.pdf [https://perma.cc/25SZ-QF9K] (explaining structured professional judgment (SPJ) and critiquing SPJ instruments used in a criminal justice context).

33. Matt Stroud, *The Minority Report: Chicago's New Police Computer Predicts Crimes, but Is It Racist?*, VERGE (Feb. 19, 2014, 9:31 AM), https://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist [https://perma.cc/E6HJ-A7QP].

scholar Sonja Starr argued that the U.S. Constitution prohibits the use of race-, gender-, or income-correlated variables in risk-assessment tools used at sentencing;[34] and the DOJ flagged both "the promise and danger of data analytics in sentencing and corrections policy."[35] 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."[36] The following year, Bernard Harcourt expanded on the problem.[37] The nation's long history of social and economic oppression of African Americans — including criminal laws and law enforcement targeting black men — has produced higher rates of arrest, prosecution, conviction, and incarceration among black Americans than white Americans. The result is that criminal history now correlates with race.[38] Any form of risk assessment that relies on criminal history will have a disparate impact on black communities, and on black men in

34. Starr, *supra* note 30. 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 "are highly correlated with race." *Id.* at 838; *see also* Sonja B. Starr, *Sentencing, by the Numbers*, N.Y. TIMES (Aug. 10, 2014), https://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html [https://perma.cc/FK9B-YLQZ].

35. DOJ Letter to U.S.S.C., *supra* note 7, at 1, 7 (formatting and capitalization altered) (cautioning that the use of risk assessment at sentencing "ultimately raises constitutional questions because of the use of group-based characteristics and suspect classifications in the analytics").

36. Eric H. Holder, Jr., U.S. Att'y Gen., Remarks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference (Aug. 1, 2014), https://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th [https://perma.cc/D87X-8GU8].

37. Harcourt, *supra* note 12, 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").

38. *See, e.g.*, Frank McIntyre & Shima Baradaran, *Race, Prediction, and Pretrial Detention*, 10 J. EMPIRICAL LEGAL STUD. 741, 759 (2013) (identifying black defendants' disproportionate likelihood of being arrested on drug charges as a potential cause of the race gap); Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 683-84, 704-06 (2016) (concluding that criminal history correlates with race in their data set).

particular.[39] Media, advocacy organizations, and other scholars echoed the concern.[40] In 2016, the ProPublica exposé supercharged the debate.[41]

Many people now focus on the possible 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 pre-

39. Harcourt, *supra* note 12, 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.").

40. *See, e.g.*, 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"); Hamilton, *supra* note 30, at 242 (discussing challenges, including constitutional considerations, relating to racial classifications); Anna Maria Barry-Jester et al., *The New Science of Sentencing: Should Prison Sentencing Be Based on Crimes That Haven't Been Committed Yet?*, MARSHALL PROJECT (Aug. 4, 2015, 7:15 AM), https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing [https://perma.cc/J4UW-BDKP]; Barry-Jester et al., *supra* note 29 (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, 9:00 AM), https://billypenn.com/2017/09/25/can-phillys-new-technology-predict-recidivism-without-being-racist [https://perma.cc/93S7-G5BH]; *Race & Justice News: Risk Assessment or Race Assessment?*, SENT'G PROJECT (July 23, 2015), http://www.sentencingproject.org/news/race-justice-news-risk-assessment-or-race-assessment [https://perma.cc/K3LC-73S6].

41. Angwin et al., *supra* note 1.

trial context.[42] Legal scholars[43] and policy organizations[44] are also increasingly attentive to the problem, as are computer scientists and econometricians who write about criminal justice.[45] Aziz Huq has laid out both the history of racial oppression in criminal justice that makes the concern so acute and the inadequacy of current constitutional doctrine to address it.[46]

Notwithstanding this growing interest, the debate remains hampered by ambiguous terms.[47] For some people, to say that a decision procedure is "biased" is to say that it is statistically unsound.[48] A risk-assessment algorithm is racially

---

42. 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 Civil Rights Concerns*, LEADERSHIP CONF. EDUC. FUND, http://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf [https://perma.cc/TQ83-TKGA] [hereinafter *Use of Pretrial "Risk Assessment" Instruments*]. *See also, e.g.*, *Predictive Policing*, MEDIA MOBILIZING PROJECT, https://mediamobilizing.org/predictive-policing [https://perma.cc/Y3FK-W7JS].

43. *See, e.g.*, Megan Stevenson & Sandra G. Mayson, *Pretrial Detention and Bail*, *in* 3 REFORMING CRIMINAL JUSTICE: A REPORT OF THE ACADEMY FOR JUSTICE, BRIDGING THE GAP BETWEEN SCHOLARSHIP AND REFORM 21, 34-39 (Erik Luna ed., 2017) (evaluating the risk of pretrial risk assessments and noting accuracy, racial-equality, and procedural concerns); Anupam Chander, *The Racist Algorithm*, 115 MICH. L. REV. 1023, 1025 (2017) (arguing that the real-world facts on which algorithms used in criminal justice risk assessment are based are "deeply suffused with invidious discrimination"); Eaglin, *supra* note 30, at 94-99 (discussing how risk assessment might "compromise[e] equality"); Mayson, *supra* note 6, at 494-96; Selbst, *supra* note 3; *see also* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 863-64 (2017) (exploring the racial effects of algorithmic prediction in the employment context).

44. *See, e.g.*, Criminal Justice Policy Program, *Moving Beyond Money: A Primer on Bail Reform*, HARV. L. SCH. (Oct. 2016), http://cjpp.law.harvard.edu/assets/FINAL-Primer-on-Bail-Reform.pdf [https://perma.cc/24PK-Z9XP]; Andrea Woods & Portia Allen-Kyle, *A New Vision for Pretrial Justice in the United States*, ACLU (Mar. 2019), https://www.aclu.org/sites/default/files/field_document/aclu_pretrial_reform_toplines_positions_report.pdf [https://perma.cc/PU6E-ZC5D].

45. *See, e.g.*, Stevenson, *supra* note 6; Jon Kleinberg et al., *Human Decisions and Machine Predictions* (Nat'l Bureau of Econ. Research, Working Paper No. 23180, 2017), https://www.nber.org/papers/w23180.pdf [https://perma.cc/3WHJ-TWLJ].

46. Huq, *supra* note 14.

47. *Cf.* Selbst, *supra* note 3, at 123 (noting that "[t]he words 'discrimination,' 'fairness,' and 'bias' evoke a family of related concepts").

48. In econometrics, "bias" describes any systematic deviation of a statistical calculation from the true value of the thing calculated. *See* Bruce E. Hansen, Econometrics 105 (Dec. 2018) (unpublished manuscript), https://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf [https://perma.cc/QS4D-BEF4] ("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, https://www.merriam-webster.com

biased in this sense if it systematically over- or understates the average risk of one racial group relative to another.[49] 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 for the other.[50] "Discrimination" also carries ambiguity; it can mean any "act of making or perceiving a difference,"[51] or only an *unjustified* act of making or perceiving a difference.[52] Along similar lines, although Jennifer Skeem and Christopher Lowenkamp have contested Harcourt's claim that criminal history serves as a "proxy" for race in risk assessment, in fact

/dictionary/bias [https://perma.cc/TF8T-KLFQ] (giving as possible definitions of the term "bias" "deviation of the expected value of a statistical estimate from the quantity it estimates" and "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others").

49. William Dieterich et al., *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, NORTHPOINTE INC. 1, 2-3, 8-13 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/K4GM-RBQY] (suggesting 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, 40 (2016) (arguing that "well-established and accepted standards exist to test for bias in risk assessment"); Skeem & Lowenkamp, *supra* note 38, at 685 (asserting that if "a given score [has] the same meaning regardless of group membership," the instrument is "unbiased").

50. *E.g.*, Kim, *supra* note 43, 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 (2003); Jeffrey S. Morrow, *Insuring Fairness: The Popular Creation of Genetic Antidiscrimination*, 98 GEO. L.J. 215, 230-32 (2009). 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).

51. *Discrimination*, MERRIAM-WEBSTER ONLINE, https://www.merriam-webster.com/dictionary/discrimination [https://perma.cc/7WZH-RRVK].

52. *Id.* (defining "discrimination" as "prejudiced or prejudicial outlook, action, or treatment"); *see also* 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) (discussing "fair" and "unfair" discrimination). Note that the term "discrimination" can also be used in a technical legal sense, to mean only such differential treatment or impact as would incur liability pursuant to antidiscrimination law. *See infra* note 63.

they just define "proxy" differently than he does.[53] These ambiguous terms can obscure the questions at stake, which are already complex enough.

B. The Problem of Equality Trade-offs

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

The ProPublica saga illustrates the problem. ProPublica concluded that the algorithmic tool COMPAS was "biased against blacks."[54] Using data from a county where COMPAS was used to assess the likelihood that a pretrial defendant would be rearrested if she or he remained at liberty, the ProPublica researchers compared COMPAS's risk classifications with defendants' actual outcomes—whether each defendant was rearrested or not—over the subsequent two years. Northpointe, the company that owns COMPAS, responded with indignation: ProPublica's own data showed that COMPAS was demonstrably race neutral![55]

The fascinating thing was that both ProPublica and Northpointe were right; they were just emphasizing different metrics of equality.[56] The fact that led Northpointe to claim race neutrality was that black and white defendants classified as high risk by COMPAS were in fact rearrested at equal rates. A high-risk classification meant the same chance of rearrest for a black defendant as for a

---

53. Skeem & Lowenkamp, *supra* note 38, 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. *Id.* at 700. 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. *Id.* (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. Harcourt, *supra* note 12, at 238.

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

55. Dieterich et al., *supra* note 49, at 1; *see also* Flores et al., *supra* note 49, at 41 (reporting the results of an independent study of the same data and concluding that COMPAS was equally predictive for white and black defendants).

56. For a detailed analysis of this discourse, see Melissa Hamilton, Debating Algorithmic Fairness (unpublished manuscript) (on file with author).

white one (approximately 60% on the any-arrest-risk scale and 20% on the violent-arrest-risk scale, over a two-year period).[57] 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 *not* be rearrested within the study period was much more likely to be classified as high risk (44.9%) than a white defendant who would not be rearrested (23.5%).[58] In statistical terms, the false-positive rate was much higher for the black defendants than the white defendants.[59] Meanwhile, a white defendant who *would* be rearrested was more likely to be deemed low risk (47.7%) than a black defendant who would be rearrested (28.0%).[60] The false-negative rate was much greater for white defendants than for black defendants. ProPublica saw these racial differences in COMPAS's error rates as a serious injustice.

The racial disparity in error rates was not, however, the result of invidious distortion in the COMPAS algorithm itself.[61] It was a mathematical result of the divergent rates of arrest between the black and white defendants in the underlying data set. Because the rate of arrest was higher among the black defendants, they, on average, had higher arrest-risk profiles. When the average risk is higher for one group than for another, a greater proportion of the former group will be predicted to be rearrested, and a greater proportion of that group will also be *mistakenly* predicted to be rearrested. This is true no matter how carefully designed the algorithm is, so long as the algorithm is also striving to have equal predictive accuracy for each racial group.

To see this aspect of prediction more clearly, consider a stylized hypothetical. Figure 1 below depicts two groups of ten arrestees each—gray and black—who are subject to risk assessment. Say that the algorithm in question predicts rearrest within a year. For clarity, presume that it makes binary decisions: for each figure, it predicts either rearrest or no rearrest. A rearrest 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-rearrest prediction is a "negative." The figures depicted in outline

57. Dieterich et al., *supra* note 49, at 4. 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 to white defendants. *See* Flores et al., *supra* note 49, 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.").

58. Angwin et al., *supra* note 1.

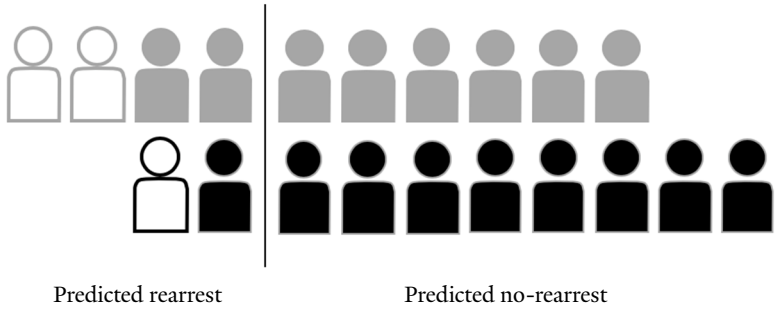
59. Whether or not the statistical concepts of "false positives" and "false negatives" are applicable in the context of risk assessment is debatable and is discussed below. *See infra* Section III.B.2.

60. Angwin et al., *supra* note 1.

61. 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* Section III.B.2.b.

only represent those who will ultimately be rearrested. The solid figures represent those who will not be rearrested. Note that the groups have different base rates of rearrest: a greater proportion of the gray group will actually be rearrested (2/10) than the black group (1/10). The dividing line between the figures, finally, represents the algorithm. The algorithm predicts rearrest for the figures to the left of the line. The figures to the right of the line are predicted not to be rearrested.

FIGURE 1.
GROUPS WITH DIFFERENT BASE RATES OF REARREST; 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 the gray groups, 50% of those forecast for rearrest (the figures to the left of the line) are indeed rearrested (the figures depicted in outline only). When the algorithm is deployed prospectively, a positive prediction for any individual will mean a 50% chance of 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.

In other ways, however, the algorithm produces unequal results. Consider the rate of false predictions among those who will *not* be rearrested—the false-positive rate. Of the eight gray figures who will not be rearrested (the solid gray figures), two are mistakenly forecast for rearrest. Of the nine black figures who will not be rearrested (the solid 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 one (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 figures as black ones 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.

TABLE 1.
GROUPS WITH DIFFERENT BASE RATES OF REARREST; PREDICTIVE PARITY

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

It is possible to modify the algorithm to equalize the false-positive rates for the two groups, but at a cost. Figure 2 below represents one possible modification: predicting arrest for a greater proportion of the black group. For both the black and the gray groups, now 25% of the *non-rearrestees* (the solid figures) are mistakenly forecast for rearrest. That is, the false-positive rate is 25% for each group. The total number of people forecast for rearrest is also much closer across groups. But notice the effect on the accuracy of the rearrest forecasts themselves (depicted by the dividing line between figures). For the gray group, a prediction of rearrest is still 50% likely to be true. But it is only about 30% likely to be true for the black group. When the algorithm is deployed prospectively, a rearrest forecast will mean something different depending on whether the figure is gray or black.

FIGURE 2.
GROUPS WITH DIFFERENT BASE RATES OF REARREST; PARITY IN FALSE-POSITIVE RATES

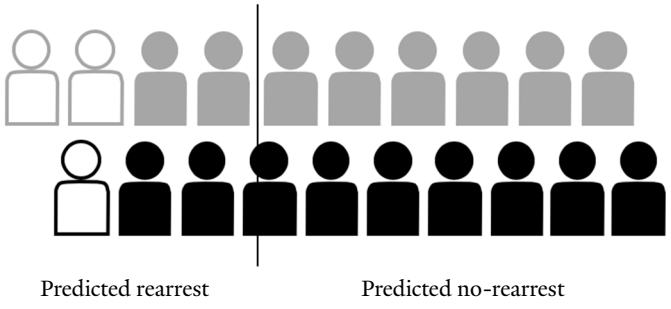


TABLE 2.
GROUPS WITH DIFFERENT BASE RATES OF REARREST; PARITY IN FALSE-POSITIVE RATES

	Gray	Black	
Percentage of Rearrest Forecasts That Are Correct	50	31	Disparate Predictive Accuracy
Percentage of No-Rearrests Falsely Forecast for Rearrest	25	25	Parity in False-Positive Rates
Percentage 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 rearrest is forecast, as depicted in Figure 3 below. But that will introduce a new disparity. Now, among those who *are* rearrested (the figures depicted only in outline), the algorithm correctly predicts rearrest 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 REARREST; PARITY IN FALSE-POSITIVE RATES AND PREDICTIVE PARITY

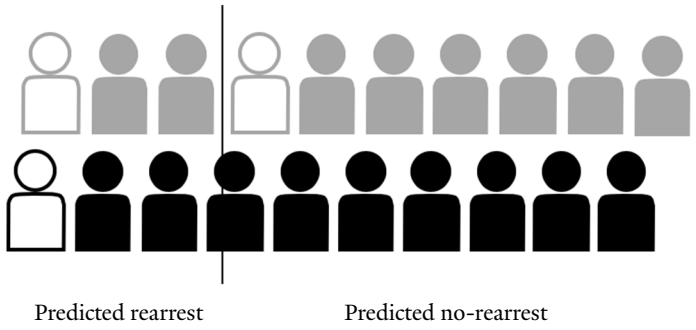


TABLE 3.
GROUPS WITH DIFFERENT BASE RATES OF REARREST; PARITY IN FALSE-POSITIVE RATES AND PREDICTIVE PARITY

	Gray	Black	
Percentage of Rearrest Forecasts That Are Correct	33	31	~ Predictive Parity
Percentage of No-Rearrests Falsely Forecast for Rearrest	25	25	Parity in False-Positive Rates
Percentage of Group Forecast for Rearrest	30	30.3	~ Statistical Parity
Percentage of Rearrests Missed	50	0	Disparate False-Negative Rates

As this example illustrates, 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.[62] When base rates differ, we must 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 intergroup equality. This Section briefly charts the most important such metrics, synthesizing the recent computer-sci-

62. *See, e.g.*, Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017); Jon Kleinberg et al., *Inherent Trade-offs in the Fair Determination of Risk Scores*, LEIBNIZ INT'L PROC. INFORMATICS, Jan. 2017, at 43:1, 43:4; *see also* Richard A. Berk et al., *Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions*, 13 J. EMPIRICAL LEGAL STUD. 94, 103-04 (2016) (illustrating the impossibility using real arraignment data and a machine-learning algorithm that forecasts whether a new arrest for a domestic-violence offense will occur within a period of twenty-four months); Huq, *supra* note 14, at 1055, 1123-24 (explaining this "impossibility result").

ence 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 measures of intergroup equality in a format accessible to both lawyers and statisticians. Those readers who are already immersed in the field or who prefer to avoid technical detail may wish to skip directly to Part II.

U.S. law divides racially unequal action into two major frameworks: disparate treatment and disparate impact.[63] Neither triggers legal liability if the differential treatment or impact is adequately justified, but for purposes of this taxonomy 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.[64] This division also aligns loosely with the distinction between "individual" and "group" equality metrics, although that distinction is not a clean one.[65]

---

63. There are two primary vehicles for asserting discrimination claims: the Equal Protection Clause of the Federal Constitution (and analogous state constitutional provisions) 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. Antidiscrimination 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, substantive 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, 1354-56 (2010).

64. To be clear, none of these output measures correspond to disparate impact *liability* under current law. As noted, only the first step in a legal disparate impact analysis is about outputs; the ultimate question is whether the challenged disparate impact is justified, a question that is arguably just as much about the decision-making process as a disparate treatment analysis. *See generally* Noah D. Zatz, *Disparate Impact and the Unity of Equality Law*, 97 B.U. L. REV. 1357, 1362 (2017) (arguing that disparate impact and disparate treatment liability are "separated superficially by the presence or absence of discriminatory intent but united fundamentally in addressing a common injury: status causation").

65. Much recent work in algorithmic fairness has categorized measures of 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., a single-threshold rule 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.

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 a protected characteristic, such as race.[66] 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 of process equality that can be understood as prohibitions on disparate treatment.

The first possible metric of process equality is *colorblindness*, which would prohibit the use of race as an input variable for prediction (or the intentional use of race proxies). The rationale for colorblindness is that if race can affect one's risk score, then there will be some set of people with otherwise identical risk prognoses who receive different risk scores on the basis of race.[67] A mandate of colorblindness would align with anticlassification conceptions of equality under law.[68]

The second possible process-equality metric is a requirement that two individuals who present the same statistical risk receive the same risk score. Statisticians refer to this requirement as a *single-threshold rule* for risk classification.[69] A single-threshold rule would prohibit the algorithm from assigning, on the basis of race, different scores 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. A single-threshold rule might seem synonymous with colorblindness, but it is not. Whereas colorblindness prohibits consideration of race in the calculation of risk, a single-threshold rule kicks in

---

66. *See, e.g.*, Ricci v. DeStefano, 557 U.S. 557, 577 (2009) (describing disparate treatment as another term for "intentional discrimination"); *see also* Washington v. Davis, 426 U.S. 229, 239-41 (1976) (holding that differential treatment of people of different races violates the Equal Protection Clause only if motivated by "discriminatory racial purpose").

67. 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.

68. *See* Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10-11 (2003) (explaining the distinction between the anticlassification and the antisubordination approaches to equality law). For this reason, Sam Corbett-Davies and Sharad Goel refer to colorblindness as "anti-classification." Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning 5 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.pdf [https://perma.cc/8PM6-L5MU].

69. Huq, *supra* note 14, at 1116-21; Corbett-Davies & Goel, *supra* note 68, at 6-8.

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 "six" 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). A single-threshold rule prohibits different "cut points" for risk classification (or classification "thresholds") by race.[70]

Both colorblindness and a single-threshold rule can be understood to reflect the Aristotelian notion that similarly situated individuals should be treated alike. They just reflect different judgments about which individuals are similarly situated for purposes of risk assessment. Colorblindness presumes that two individuals are similarly situated if they present the same statistical risk, calculated without reference to race. A single-threshold rule presumes that two individuals are similarly situated if they present the same statistical risk, calculated with as much precision as possible. If race moderates the predictive value of other factors, the two can be mutually incompatible.[71]

2. *Disparate Impact (Output-Equality) Metrics*

Disparate impact refers to the differential effects of some decision-making process on members of one racial group.[72] It concerns 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. Since these are inherently statistical concepts, it is necessary to have a sizable number of the algorithm's predictions for members of each racial group to evaluate an algorithm by any one of these measures and, in most cases, to know how many of the predictions were ultimately correct. Output-equality metrics align with antisubordination conceptions of equality.[73]

The following schema presents a core set of potential output-equality metrics. Like the figures above, Figure 4 depicts two groups, gray and black, with

---

70. 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 30, at 86.

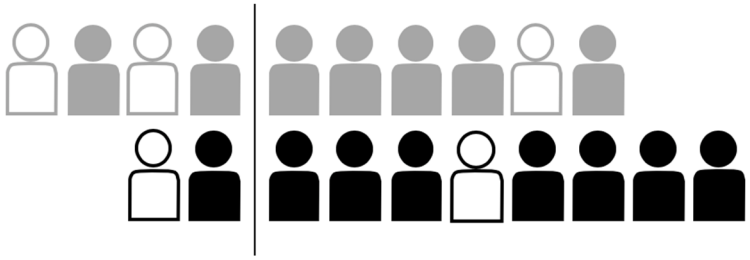
71. For a fuller explanation of this possibility, see *infra* Section III.A.

72. *See, e.g.*, Civil Rights Act of 1964 § 703, 42 U.S.C. § 2000e-2(a)(2) (2018); Griggs v. Duke Power Co., 401 U.S. 424, 430-31 (1971).

73. *See* Balkin & Siegel, *supra* note 68, at 9 ("Antisubordination theorists contend that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups.").

different base rates of the outcome in question—say rearrest for violent crime. Assume that the algorithm makes binary rearrest/no-rearrest forecasts. Once again, the figures depicted only in outline will ultimately be rearrested and the line represents the algorithm (those persons forecast for rearrest appear to the left of the line).

FIGURE 4.
GROUPS WITH DIFFERENT BASE RATES OF REARREST, AGAIN



a. 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 intergroup equality. It is also the one that dominates disparate impact law. EEOC guidance, for example, provides that too great a divergence from statistical parity is prima facie evidence of "adverse impact."[74] 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 is forecast for rearrest (the figures to the left of the line).[75] Statistical parity is some-

74. 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) (2018).

75. 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

times called "demographic parity." Related metrics in the computer-science literature include the Calders-Verwer (CV) score[76] and the "p%-rule."[77]

b. Predictive Parity

Predictive parity, the metric that Northpointe emphasized in its debate with ProPublica,[78] measures the algorithm's rate of accuracy among those who receive the same forecast. If the algorithm's rearrest forecasts are correct at an equal rate for each group, the algorithm achieves parity in *positive predictive value*. If the no-rearrest forecasts are correct at an equal rate for each group, the algorithm 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"[79] and "conditional use accuracy equality."[80] In our example in Figure 4, the algorithm achieves parity in positive predictive value only. For both the black and the gray groups, 50% of those forecast for rearrest are indeed rearrested (the figures depicted only in outline and to the left of the dividing line).

c. Equal False-Positive and True-Negative Rates (Equal Specificity)

ProPublica, however, argued that equality requires parity in false-positive rates. The false-positive rate and its inverse, the true-negative rate, measure the

instance, want to ensure that, among the subgroup of defendants with equivalent criminal histories and other "legitimate" predictors of arrest outside of race, the percentage forecast for future arrest is the same for each racial group. Scholars call this "conditional statistical parity." *E.g.*, Sam Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness 798 n.2 (2017) (unpublished manuscript) (on file with author).

76. *See* Toshihiro Kamishima et al., *Fairness-Aware Classifier with Prejudice Remover Regularizer*, *in* MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES 35, 37 (Peter A. Flach et al. eds., 2012) (defining the CV score as a difference, rather than ratio, of outcome rates between two groups).

77. *See* Muhammad Bilal Zafar et al., Fairness Constraints: Mechanisms for Fair Classification 2 (2017) (unpublished manuscript), http://proceedings.mlr.press/v54/zafar17a/zafar17a.pdf [https://perma.cc/5V2V-F77E] (generalizing the 80% threshold favored by the EEOC to thresholds of any arbitrary *p* value).

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

79. Kleinberg et al., *supra* note 62, at 4.

80. Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 2018 SOC. METHODS & RES. 1, 10.

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 rearrest—the law abiders mistakenly projected to commit future 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 (the solid gray figures), two are mistakenly forecast for rearrest (29%), whereas of the eight black people who will *not* be rearrested (solid black figures), only one is mistakenly forecast for rearrest (12%). The proportion of non-rearrestees who are *correctly* predicted is the true-negative rate (or the algorithm's "specificity").[81] Statisticians and computer scientists have referred to equal specificity both as "balance for the negative class" and as "predictive equality."[82]

There is disagreement about whether this statistical vocabulary for forecasting errors is appropriate to risk assessment. 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."[83] Nonetheless, the binary language of true versus false prediction is a helpful heuristic to explain where the costs of uncertainty fall.

d. Equal False-Negative and True-Positive Rates (Equal Sensitivity)

Whereas specificity measures the algorithm's accuracy among the true negatives (people who are *not* ultimately rearrested), sensitivity measures the algorithm's accuracy among the true positives—people who *are* ultimately rearrested. The proportion of this group correctly forecast for rearrest is the true-positive rate; the proportion mistakenly forecast for no-rearrest is the 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 figures depicted only in outline) 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 (the figures depicted in outline only) is correctly predicted and one is not, so both the true-positive and false-negative rates are 1/2 (50%).

---

81. In our model, this is the percentage of solid figures correctly left to the right of the dividing line (five of seven gray (71%) and seven of eight black (88%)).

82. Corbett-Davies et al., *supra* note 75, at 798 ("predictive equality"); Kleinberg et al., *supra* note 62, at 4 ("balance for the negative class").

83. For further discussion of this point, *see infra* note 201 and accompanying text.

Computer scientists have referred to parity in true-positive rates as "balance for the positive class"[84] and as "equal opportunity," because it means that a true positive will have an equal chance of being correctly predicted regardless of group membership.[85] 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."[86]

e. Equal Rate of Correct Classification

It is also possible to conceive of equality as parity in the rate of correct classification overall, or the percentage of each group correctly predicted. In our model, 70% of the gray figures are correctly classified (two actual rearrests—the figures depicted only in outline—to the left of the dividing line, and five no-rearrests to the right of the line). Of the black group, 80% are correctly classified (one actual rearrest to the left of the line and seven no-rearrests to the right).[87] Richard Berk and colleagues call parity in the rate of correct classification "overall accuracy equality."[88]

f. 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." This 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

84. Kleinberg et al., *supra* note 62, at 4.

85. Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, NIPS PROC. (2016), https://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf [https://perma.cc/LD25-6RR7].

86. *See* sources cited *infra* note 96.

87. Inversely, only 20% of the black group, but 30% of the gray group, is classified incorrectly.

88. Berk et al., *supra* note 80, at 13.

project in Philadelphia, for instance, stakeholders determined that missing a new arrest for domestic violence was ten times worse than incorrectly predicting a new arrest.[89] 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. Parity in cost ratios is also known as "treatment equality."[90]

g. Area-Under-the-Curve (AUC) Parity

There are also a number of measures that express an algorithm's overall performance at sorting people along a risk spectrum 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 posing a higher risk than the other). Parity in AUC scores is yet another measure of equality in predictive accuracy.

Table 4 charts these output metrics, their values in the black/gray example, and terms for each in the statistics and computer-science literature.[91]

---

89. Berk et al., *supra* note 62, at 104; *see also* Melissa Hamilton, *Adventures in Risk*, 47 ARIZ. ST. L.J. 1, 33 (noting that the contrary judgment is also reasonable); 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) ("[I]t 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).").

90. Berk et al., *supra* note 80, at 15. The question of the relative cost of a false positive and false negative in the prediction context evokes the famous Blackstone ratio, which asserts a position on the relative costs of false negatives and false positives in the context of criminal conviction and punishment. *See* 4 WILLIAM BLACKSTONE, COMMENTARIES *352 ("[T]he law holds, that it is better that ten guilty persons escape, than that one innocent suffer."); *cf.* Alexander Volokh, n *Guilty Men*, 146 U. PA. L. REV. 173 (1997) (chronicling variants on the Blackstone ratio through history); Megan Stevenson & Sandra Mayson, *n* Dangerous Men (unpublished manuscript) (on file with author) (exploring the analogue of the Blackstone ratio for preventive detention).

91. Like the text above, Table 4 simplifies the relevant concepts in at least three ways. It (1) treats risk assessment as binary prediction; (2) ignores the issue of whether the validation data will correspond to the population on which the tool is applied; and (3) ignores whether the classifier is an asymptotically unbiased estimator—called the tool's "estimation accuracy." *See* Berk et al., *supra* note 80, at 16.

TABLE 4.
DISPARATE IMPACT (OUTPUT-EQUALITY) METRICS

Stats. / Comp. Sci. Equality Terms	Parity in . . .	Gray	Black
Statistical Parity, Demographic Parity (related: Conditional Statistical Parity)[92]	Population Impact: Percentage of group predicted P	40%	20%
	Inverse: Percentage predicted N	60%	80%
Predictive Parity, Calibration Within Groups, Conditional Use Accuracy Equality[93]	Positive Predictive Accuracy: Percentage of P predictions that are correct	50%	50%
	Negative Predictive Accuracy: Percentage of N predictions that are correct	83%	88%
Balance for the Negative Class, Predictive Equality[94]	True-Negative Rate (Specificity): Percentage of Ns correctly predicted	71%	88%
	False-Positive Rate: Percentage incorrectly predicted	29%	12%

92. Benjamin Fish et al., *A Confidence-Based Approach for Balancing Fairness and Accuracy*, *in* PROCEEDINGS OF THE 2016 SIAM INTERNATIONAL CONFERENCE ON DATA MINING 144, 144 (Sanjay Chawla Venkatasubramanian & Wagner Meira eds., 2016) (using the term "statistical parity"); Toshihiro Kamishima et al., *Considerations on Fairness-aware Data Mining*, *in* PROCEEDINGS OF THE 2012 IEEE INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS 378, 382 (Mohammed J. Zaki et al. eds., 2012) (same); Berk et al., *supra* note 80, at 13 (same); Huq, *supra* note 14, at 1116 (same); Kleinberg et al., *supra* note 62, at 8 (same); Corbett-Davies & Goel, *supra* note 68, at 6 (describing "classification parity"); Corbett-Davies et al., *supra* note 75, at 2 (using the term "conditional statistical parity" to refer to parity in classification rates after controlling for "'legitimate' risk factors"); Hardt et al., *supra* note 85, at 1 (using the term "demographic parity"); Richard Zemel et al., *Learning Fair Representations*, 28 PROC. MACHINE LEARNING RES. 1 (2013) (using the term "statistical parity").

93. Berk et al., *supra* note 80, at 14 (describing "conditional use accuracy equality"); Kleinberg et al., *supra* note 62, at 4 (discussing "calibration within groups"); Corbett-Davies & Goel, *supra* note 68, at 2 (describing "calibration"); Dieterich et al., *supra* note 49 (describing "predictive parity"); Hardt et al., *supra* note 85, at 5 (describing "a score that is *calibrated within each group*").

94. Kleinberg et al., *supra* note 62, at 4 (describing "balance for the negative class"); Corbett-Davies et al., *supra* note 75, at 798 ("predictive equality").

Stats. / Comp. Sci. Equality Terms	Parity in . . .	Gray	Black
Balance for the Positive Class, Equal Opportunity[95]	True-Positive Rate (Sensitivity): Percentage of Ps correctly predicted	67%	50%
	False-Negative Rate: Percentage incorrectly predicted	33%	50%
Balance for Both Classes, Equalized Odds, Conditional Procedure Accuracy Equality, Equality of Opportunity[96]	Both True Positive and True Negative Rates		
Overall Accuracy Equality,[97] Overall Procedure Accuracy[98]	Overall Rate of Correct Classification: Percentage of group correctly predicted	70%	80%
	Inverse: Percentage incorrectly predicted	30%	20%
Treatment Equality[99]	Distribution of Errors b/t FP & FN ("Cost Ratio")	2:1	1:1
Total Fairness[100]	Everything Above		

D. Trade-offs, Reprise

An algorithm can be designed to achieve any one of the above metrics of output equality, but not *all* of them together. That is, an algorithm cannot be designed to achieve "total fairness." This Section offers a very brief overview of the

95. Kleinberg et al., *supra* note 62, at 4 (discussing "balance for the positive class"); Hardt et al., *supra* note 85, at 4 (defining "equal opportunity").

96. Berk et al., *supra* note 80, at 13-14 (defining "conditional procedure accuracy equality"); Kleinberg et al., *supra* note 62, at 2-3 (discussing "balance" for both classes); Hardt et al., *supra* note 85, at 3 (describing "equalized odds"); Matthew Joseph et al., *Fair Algorithms for Infinite and Contextual Bandits*, FAIRNESS, ACCOUNTABILITY & TRANSPARENCY MACHINE LEARNING 1 (2016), https://www.fatml.org/media/documents/better_fair_algorithms_for_infinite_contextual_bandits.pdf [https://perma.cc/F6BG-PDQL] (discussing "equality of opportunity").

97. Berk et al., *supra* note 80, at 13.

98. *Id.*

99. *Id.* at 14-15.

100. *Id.* at 15.

likely trade-offs between equality and overall accuracy and between different metrics of equality.

1. Equality/Accuracy Trade-offs

When base rates of the predicted outcome differ across groups, the most accurate algorithm possible will predict that outcome at different rates across groups. Imposing certain metrics of output equality will therefore have a cost in accuracy. The nature and magnitude of the trade-off will depend on how dramatically the underlying base rates diverge and on the nature of the fairness intervention.

Recent work in computer science has demonstrated this trade-off in practice. Sam Corbett-Davies and colleagues explored the possibility of designing a machine-learning risk-assessment algorithm to achieve three separate metrics of fairness, using the same Broward County, Florida data that ProPublica used.[101] They found that imposing any one of those metrics compromised the algorithm's accuracy. The optimal algorithm, from a public-safety perspective, was "unconstrained" with respect to group fairness. They concluded that "[a]dhering to past [group] fairness definitions can substantially decrease public safety; conversely, optimizing for public safety alone can produce stark racial disparities."[102] Other studies have offered similar demonstrations.[103]

2. Equality/Equality Trade-offs

There are also inescapable trade-offs between different metrics of equality. As discussed in Section I.B above, whenever base rates of the event we have undertaken to predict diverge across racial groups, it will be impossible to achieve

---

101. Corbett-Davies et al., *supra* note 75, at 798 (constraining the algorithm to produce, respectively, statistical parity, conditional statistical parity, and parity in false-positive rates).

102. *Id.* at 797.

103. *See, e.g.*, Richard Berk, *Accuracy and Fairness for Juvenile Justice Risks Assessments*, 16 J. EMPIRICAL LEGAL STUD. 175 (2019) (finding that modifying data to achieve statistical parity produced extremely high false-negative rates); Kleinberg et al., *supra* note 45, at 6; Zachary Lipton et al., *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, NIPS PROC. 1 (2018), https://arxiv.org/pdf/1711.07076.pdf [https://perma.cc/X8VF-EY53] (demonstrating that enforced blindness to protected traits can have a substantial cost in accuracy); Joan Petersilia & Susan Turner, *Guideline-Based Justice: Prediction and Racial Minorities*, 9 CRIME & JUST. 151, 174 (1987) (reporting that omitting factors correlated with race from a recidivism prediction algorithm significantly reduced the accuracy of the model).

equality by every metric at once.[104] To achieve any particular measure of equality, it will likely be necessary to sacrifice at least one of the others.

In a recent study, for instance, Richard Berk adjusted juvenile justice data to produce statistical parity in predictions of rearrest for violent crime.[105] This resulted in extremely high rates of false negatives; the algorithm missed 92% of violent-crime rearrests of white juveniles and 98% of violent-crime rearrests of black juveniles.[106] Such high false-negative rates are a hefty cost in accuracy in and of themselves. But the increase in error also had a disparate racial impact. Because the base rate of arrest for violence in this data was significantly higher among black juveniles than white juveniles,[107] and because most violent crime is intraracial,[108] the astronomical false-negative rates would mean a much greater *absolute* number of violent-crime arrests missed in the black community than in the white community. To the extent that violent-crime arrests correspond to violent crimes, the effort to achieve statistical parity for black juveniles subject to risk assessment comes at the cost of disparate harm to black victims.[109]

Given the trade-offs between certain equality metrics and overall accuracy, and between the equality metrics themselves, what should "equality" mean—what measure of equality should risk-assessment tools be required to meet? As Aziz Huq has thoroughly explained, the law does not answer this question.[110] Huq instead suggests a return to first principles and a commitment to ensuring that predictive programs do not impose a net burden on communities of color.[111]

---

104. Economics and statistics literature sometimes refers to this phenomenon as the problem of infra-marginality. *See, e.g.*, Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131, 135 (2002); Camelia Simoiu et al., *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS APPLIED STAT. 1193 (2017).

105. Berk, *supra* note 103, at 190.

106. *Id.* at 186.

107. *Id.* at 187.

108. Rachel E. Morgan, *Race and Hispanic Origin of Victims and Offenders*, BUREAU JUST. STAT. 2012-15 (2017), https://www.bjs.gov/content/pub/pdf/rhovo1215.pdf [https://perma.cc/F2X6-N9KK].

109. For further illustrations of trade-offs between equality and accuracy, and between different equality measures, see Section III.B.2 and the Appendix. Berk and his colleagues have also provided another illustration of these equality/accuracy trade-offs, using real arraignment data and a machine-learning algorithm that forecasts new arrests for a domestic-violence offense within a period of twenty-one months. The base rate among black defendants was 11%, and the base rate among white defendants was 6%. Berk and his colleagues found that, if the algorithm is designed to achieve predictive parity with respect to a prediction of no-rearrest, the "false negative and false positive rates vary dramatically by race." Specifically, the false negative rate is 49% for black defendants and 93% for white defendants; the false positive rate is 2% for white defendants and 24% for black defendants. Berk et al., *supra* note 80, at 32.

110. Huq, *supra* note 14, at 1055.

111. *Id.* at 1111.

That commitment is surely worth making. The question is how to design predictive tools to honor it. And the problem is that no one measure of predictive equality can control the impact of prediction in the world. To understand why this is so, one must first understand the source of the problem: prediction itself.

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, absent intervention, history will repeat itself. So what prediction does is identify patterns in past data and offer them as projections about future events. If there is racial disparity in the data, there will be racial disparity in prediction too. It is possible to replace one form of disparity with another, but impossible to eliminate it altogether.

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 event that 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,[112] any disparity in prediction reflects 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.

B. Racial Disparity in Past-Crime Data

From a racial equity perspective, the key question for any predictive tool is what it predicts: what data point is labeled as a "positive" instance of the target variable. Most contemporary criminal justice risk-assessment tools purport to

112. That is, the algorithm achieved predictive parity. *See supra* notes 78-80 and accompanying text; *supra* Table 1.

predict future crime.[113] But that is not actually what they predict. They generally predict future arrest.[114]

The reason that risk-assessment tools predict arrest rather than crime is that the data do not allow for direct crime prediction. To determine who is likely to commit crime in the future, one would have to look at who has committed crimes in the past. But we do not know precisely who has committed crimes in the past. Most crimes are never reported; some are reported falsely; and crime reports do not reliably identify crime perpetrators. Law enforcement institutions strive to identify perpetrators, and toward that end they make arrests, file charges, and seek convictions. These institutional events are documented, but even the best law enforcement agency does not make an accurate arrest for every crime. Most crimes never result in arrest.[115] Some arrests are erroneous. The same is true of filed charges and of convictions. So our record of past crimes is really a record of crime reports and law enforcement actions, and the relationship of that record to actual crimes committed is opaque.[116] Given this fundamental data limitation, most contemporary criminal justice risk-assessment tools predict arrest on the

113. *See, e.g.*, *Overview of the LSI-R*, MULTI-HEALTH SYS., https://www.mhs.com/MHS-Publicsafety?prodname=lsi-r [https://perma.cc/AQ8X-5FM7] (purporting to predict, inter alia, "recidivism"); *Public Safety Assessment: Risk Factors and Formula*, ARNOLD FOUND. 2-3 (2016), https://www.arnoldfoundation.org/wp-content/uploads/PSA-Risk-Factors-and-Formula.pdf [https://perma.cc/R62H-HD8Z] (purporting to predict "new criminal activity").

114. Some tools have other target variables, but the analysis in this section applies to many other target variables too. In the pretrial context, for instance, risk-assessment tools also predict "failure to appear," defined in terms of data points that vary by jurisdiction. *See* Mayson, *supra* note 6, at 509-13. There are also risk-assessment instruments that purport to predict violence but in fact predict any allegation of violence, whether it results in arrest or not (let alone conviction). *See, e.g.*, Min Yang et al., *The Efficacy of Violence Prediction: A Meta-Analytic Comparison of Nine Risk Assessment Tools*, 136 PSYCHOL. BULL. 740, 742 (2010) (noting that "[t]he range of possible criterion variables for violence is wide, and "includes self-reports to third-party reports . . . , informal social service or police contact, formal contact or police charges, formal adjudication and court convictions, and incarceration").

115. *Crime in the United States 2017*, FBI tbl.425, https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/tables/table-25 [perma.cc/G2QQ-F34R] (reporting that, in data from reporting law enforcement agencies nationwide, only 45.6% of violent offenses and 17.6% of property offenses were cleared by arrest).

116. *Cf.* Cathy O'Neill, *Commentary: Let's Not Forget How Wrong Our Crime Data Are*, CHI. TRIB. (May 25, 2018), https://www.chicagotribune.com/news/opinion/commentary/ct-perspec-danger-marijuana-legalizing-crime-data-black-youth-facial-bias-0528-story.html [https://perma.cc/DMU4-XW38] (arguing that crime statistics are a poor proxy for actual crime).

premise that it is the best available proxy for crime commission.[117] A few predict arrest for a specified type of crime, but most assess the likelihood of arrest for any offense at all within a designated timespan.

The choice to predict arrest has profound consequences for racial equity because in most places, for nearly all crime categories, arrest rates have been racially disparate for decades. The recent DOJ investigations into the Ferguson and Baltimore police departments offered two dramatic examples.[118] But Ferguson and Baltimore are not unique. In 2014, a USA Today analysis of FBI data concluded that "[a]t least 1,581 other police departments across the USA arrest black people at rates even more skewed than in Ferguson."[119] The report explained: "Blacks are more likely than others to be arrested in almost every city for almost every type of crime. Nationwide, black people are arrested at higher rates for crimes as serious as murder and assault, and as minor as loitering and marijuana possession."[120] The most recent data are no better. In 2017, the black arrest rate nationwide was at least twice as high as the white arrest rate for every crime category

---

117. Whether arrest is actually the best available proxy for commission of crime is a difficult and contested question. *See* Anna Roberts, *Arrest as Guilt*, 60 ALA. L. REV. (forthcoming 2019) (manuscript at 9) (on file with author).

118. Civil Rights Div., *Investigation of the Baltimore City Police Department*, U.S. DEP'T JUST. 3 (2016), https://www.justice.gov/crt/file/883296/download [https://perma.cc/2BHX-3QB4] [hereinafter *Baltimore Investigation*]; Civil Rights Div., *Investigation of the Ferguson Police Department*, U.S. DEP'T JUST. 2 (2015), https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/03/04/ferguson_police_department_report.pdf [https://perma.cc/GFX7-ZDWT] [hereinafter *Ferguson Investigation*].

119. Brad Heath, *Racial Gap in U.S. Arrest Rates: "Staggering Disparity,"* USA TODAY (Nov. 18, 2014), https://www.usatoday.com/story/news/nation/2014/11/18/ferguson-black-arrest-rates/19043207 [https://perma.cc/V9MY-K2WN].

120. *Id.* In fact, the aggregate national arrest rate for black people was at least *twice* as high as the aggregate white arrest rate every year from 1980 through 2014. *Arrest Data Analysis Tool*, BUREAU JUST. STAT., https://www.bjs.gov/index.cfm?ty=datool&surl=/arrests/index.cfm (click "National Estimates," then "Trend Graphs by Race," and then select the race and "All offenses") (last visited Feb. 15, 2019). A similar trend holds among misdemeanors. *See* Megan Stevenson & Sandra Mayson, *The Scale of Misdemeanor Justice*, 98 B.U. L. REV. 731, 758-63 (2018) (finding that "the [national] black arrest rate [for an index of misdemeanor offenses] has hovered around 1.7 times the white arrest rate since 1980"). The starkest disparities may be in more serious offense categories. For every year from 1980 through 2012, the black arrest rate for what the Bureau of Justice Statistics designates the "violent crime index" was at least three times the white arrest rate, and from 1980 through 1989 it was more than six times the white arrest rate. *Arrest Data Analysis Tool*, BUREAU JUST. STAT., https://www.bjs.gov/index.cfm?ty=datool&surl=/arrests/index.cfm (click "National Estimates," then "Trend Graphs by Race," and then select the race and "Violent Crime Index") (last visited Feb. 15, 2019).

except driving under the influence, violations of liquor laws, and "drunkenness."[121] For murder and robbery, the black arrest rate was approximately seven times the white arrest rate.[122] Given these pervasive and persistent trends, it is likely that many past-crime data sets will manifest racial disparity in arrest rates for many categories of crime.

C. Two Possible Sources of Disparity

There are two possible explanations for such disparities. The first is that they represent a racial distortion relative to the underlying rate of crime commission: white and black people commit the crime at equal rates, but racial skew in enforcement or reporting practices distorts this ground truth. The second possible explanation is that the disparity reflects a difference in offending rates across racial lines. This evokes one of the most pernicious themes in racist ideology—the association of blackness with criminality.[123] 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;[124] others assume that it reflects differences in

---

121. I calculated 2017 arrest rates by race and offense category using the arrest totals reported in the FBI's Uniform Crime Reports series and national population estimates reported by the U.S. Census Bureau. These sources have serious limitations, but to my knowledge are the best available basis for calculating national arrest rates by race. *See Crime in the United States 2017*, FBI tbl.43A, https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/tables/table-43 [https://perma.cc/EKG9-YD6G] (showing arrest totals by offense category and race in reporting jurisdictions); *Quick Facts: Population Estimates, July 1, 2017 (V2017)*, U.S. CENSUS BUREAU, https://www.census.gov/quickfacts/fact/table/US/PST045217#PST045217 [https://perma.cc/989D-GWNG] (reporting that white people constituted 76.6% of the national population of 325,719,178 (or 249,500,890) and that black people constituted 13.4% (or 43,646,370)).

122. *See Crime in the United States 2017*, *supra* note 121, tbl.43A.

123. *See, e.g.*, RANDALL KENNEDY, RACE, CRIME, AND THE LAW 137 (1997); KATHERYN RUSSELL-BROWN, THE COLOR OF CRIME 128 (1998); *cf. Crime in the United States 2017*, *supra* note 121, tbl.43A.

124. *See, e.g.*, *Hearing on the Proposed Pennsylvania Risk Assessment Tool for Sentencing* 8-9 (June 13, 2018) (testimony of Mark Houldin, Phila. Def. Ass'n), https://www.hominid.psu.edu/specialty_programs/pacs/guidelines/archived-sentence-risk-assessment/testimony/mark-f.-houldin-policy-director-defenders-association-of-pennsylvania.-harrisburg-june-13-2018/view [https://perma.cc/VE6Z-TPGN].

underlying crime rates.[125] So long as these conflicting assumptions go unstated, the debate cannot proceed.

Without confronting the two possible sources of disparity, moreover, it is impossible to remedy them because each one demands a different response. Distortions in the data or risk-assessment process can sometimes be corrected. And if correction is not possible—if the data cannot be made to reliably reflect the underlying incidence of crime—then they should not serve as the basis for risk assessment at all. But if the data *do* reliably reflect the underlying incidence of crime, and predictive disparity flows from a difference in underlying crime rates, then the disparity cannot be eliminated within the data or the predictive process. Nor is the answer to jettison algorithmic assessment in favor of subjective prediction. So long as the data reliably reflect the incidence of some event that is worth predicting, algorithmic risk assessment may have a valuable role to play.[126]

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

1. Disparate Law Enforcement Practice?

There is no question that, in many places, police have disproportionately arrested people of color relative to the rates at which black people and white people, respectively, commit crimes. Marijuana arrest rates are an oft-cited example: although black and white people use marijuana at approximately equal rates, black people have been arrested for marijuana much more frequently.[127] This also appears to be the case with drug arrests overall.[128] Recent DOJ investigations have

---

125. *See, e.g.*, *id.* at 8 (citing research commissioned by the Pennsylvania Sentencing Commission as interpreting racial differences in arrest rates to reflect racial differences in commission rates).

126. Part IV considers this possibility.

127. *The War Against Marijuana in Black and White*, ACLU 16-18 (2013), https://www.aclu.org/report/report-war-marijuana-black-and-white?redirect=criminal-law-reform/war-marijuana-black-and-white [https://perma.cc/S5RY-WUB7].

128. *See, e.g.*, MODEL PENAL CODE § 1.02(2), Reporters' Note 31 (AM. LAW INST., 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); 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, 304-08 (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

revealed racial disparities in arrest rates in New Orleans, Ferguson, and Baltimore that are not explicable on the basis of underlying crime rates alone.[129] Some scholars argue that distortions to police data are so pervasive that such data should never be taken to reflect crime patterns, but should instead be understood to document "the practices, policies, biases, and political and financial accounting needs of a given [police] department."[130]

To the extent that racial disparities in past-arrest rates derive from disparate law enforcement practice, that distortion makes "future arrest" a racially skewed proxy for "future crime." As between a black and a white defendant who are equally likely to commit crime, the black defendant may be more likely to be arrested.[131] Conversely, the fact that a black defendant is more likely to be arrested may not mean she or he is more likely to commit crime. There is thus reason to think that tools assessing the likelihood of "any arrest" may be racially biased in the sense that a given score – which corresponds to some likelihood of

SIGNIFICANCE 14, 19 (2016) (discussing bias in predictive policing); David Huizinga et al., *Disproportionate Minority Contact in the Juvenile Justice System: A Study of Differential Minority Arrest/Referral to Court in Three Cities*, NAT'L CRIM. JUST. REF. SYS. 3 (July 28, 2007), https://www.ncjrs.gov/pdffiles1/ojjdp/grants/219743.pdf [https://perma.cc/6KW3-SDE4] (evaluating longitudinal data from three cities and finding substantial racial differences in police contact after controlling for differences in self-reported offending).

129. *Baltimore Investigation*, *supra* note 118, at 72 ("In sum, [the Baltimore Police Department]'s stops, searches, and arrests disproportionately impact African Americans and predominantly African-American neighborhoods and cannot be explained by population patterns, crime rates, or other race-neutral factors."); *Ferguson Investigation*, *supra* note 118, at 62-79 (concluding that dramatic racial disparities in traffic stops, citations, and arrests were "not the necessary or unavoidable results of legitimate public safety efforts" and "stem[med] in part from intentional discrimination"); Civil Rights Div., *Investigation of the New Orleans Police Department*, U.S. DEP'T JUST. 34 (Mar. 16, 2011), https://www.justice.gov/sites/default/files/crt/legacy/2011/03/17/nopd_report.pdf [https://perma.cc/Z5VP-84B7] [hereinafter *New Orleans Investigation*] (finding "reasonable cause to believe that there is a pattern or practice of unconstitutional conduct and/or violations of federal law with respect to discriminatory policing"); *id.* at 39 (concluding that "the level of [racial] disparity [in arrests] for youth is so severe and so divergent from nationally reported data that it cannot plausibly be attributed entirely to the underlying rates at which these youth commit crimes"); *see also* Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 192 (2019).

130. Richardson et al., *supra* note 129, at 8.

131. *See, e.g.*, Jeffrey Fagan & Tracey L. Meares, *Punishment, Deterrence and Social Control: The Paradox of Punishment in Minority Communities* 6 OHIO ST. J. CRIM. L. 173, 178-80 (2008); Preeti Chauhan et al., *Trends in Arrests for Misdemeanor Charges in New York City, 1993-2016*, at 21, MISDEMEANOR JUST. PROJECT 21 (Feb. 1, 2018), https://misdemeanorjustice.org/wp-content/uploads/2018 /01/2018_01_24_MJP.Charges.FINAL.pdf [https://perma.cc/SP33-C6JA].

arrest—will mean a different risk of crime *commission* for black versus white defendants.[132] The tool will systematically overstate the riskiness of black men relative to other groups. Predictive disparities produced by this kind of distortion are sometimes called "irrational discrimination," because the disparity does not track the underlying reality of 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.

2. *Disparate Rates of Crime Commission?*

The second possible explanation for racial disparity in past-arrest rates is a difference in the underlying incidence of crime. 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.[133] White men have committed the vast majority of mass shootings in the United States during the last thirty years.[134] Nationwide firearm homicide rates have been higher in recent decades in black communities than in white ones, but the degree of disparity varies by

132. *See, e.g.*, Kristian Lum, *Limitations of Mitigating Judicial Bias with Machine Learning*, 1 NATURE HUM. BEHAV. 1 (2017).

133. *See, e.g.*, Sean Esteban McCabe et al., *Race/Ethnicity and Gender Differences in Drug Use and Abuse Among College Students*, 6 J. ETHNICITY SUBSTANCE ABUSE 75 (2007) (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).

134. *Number of Mass Shootings in the United States between 1982 and November 2018, by Shooter's Race and Ethnicity*, STATISTA, https://www.statista.com/statistics/476456/mass-shootings-in-the-us-by-shooter-s-race [https://perma.cc/238C-PVZR].

state.[135] High-stakes financial crimes are disproportionately committed by people working in the upper echelons of financial-services firms, and these individuals are disproportionately white men.[136]

In the Broward County data, as well as several other data sets used in recent risk-assessment studies, arrest rates for offenses designated as "violent" were higher among the black population than the white population.[137] Jennifer Skeem and Christopher Lowenkamp have opined that the disparity represents differential offending rates rather than differential enforcement.[138] This Article does not take any position on whether that is so; I have neither the data nor the expertise to judge.

The point is that *if* underlying offense 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 functions as a mirror. If the black population in the relevant data is statistically riskier with respect to the designated crime category, risk-assessment tools will reflect as much. If the mirror is modified to ignore this statistical fact, that very blindness will have disparate racial

---

135. *See, e.g.*, Alexia Cooper & Erica L. Smith, *Homicide Trends in the United States, 1980-2008*, U.S. DEP'T JUST. 11 (Nov. 2011), https://www.bjs.gov/content/pub/pdf/htus8008.pdf [https://perma.cc/88XB-M3ZV]; Michael Planty & Jennifer L. Truman, *Firearm Violence, 1993-2011*, U.S. DEP'T JUST. 5 (May 2013), https://www.bjs.gov/content/pub/pdf/fv9311.pdf [https://perma.cc/B2Y4-5XSW] (showing rates of firearm victimization by race); *see also* Corinne A. Riddell et al., *Comparison of Rates of Firearm and Nonfirearm Homicide and Suicide in Black and White Non-Hispanic Men, by U.S. State*, 168 ANNALS INTERNAL MED. 712 (2018).

136. *See* Brian Clifton et al., *Predicting Financial Crime: Augmenting the Predictive Policing Arsenal*, NEW INQUIRY (Apr. 25, 2017), https://whitecollar.thenewinquiry.com/static/whitepaper.pdf [https://perma.cc/QS9Y-JDG6] (synthesizing data on location of financial crimes); *cf.* Stacy Jones, *White Men Account for 72% of Corporate Leadership at 16 of the Fortune 500 Companies*, FORTUNE (June 9, 2017), https://fortune.com/2017/06/09/white-men-senior-executives-fortune-500-companies-diversity-data [https://perma.cc/67YB-ZYKR]; Susan E. Reed, *Corporate Boards Are Diversifying. The C-suite Isn't.*, WASH. POST (Jan. 6, 2019), https://www.washingtonpost.com/outlook/corporate-boards-are-diversifying-the-c-suite-isnt/2019/01/04/c45c3328-0f02-11e9-8938-5898adc28fa2 [https://perma.cc/X3YL-HXLG]. Clifton, Lavigne, and Tseng offer a new predictive technology "trained on incidents of financial malfeasance from 1964 to the present day, collected from the Financial Industry Regulatory Authority (FINRA)." Brian Clifton et al., *White Collar Crime Risk Zones*, NEW INQUIRY (Apr. 26, 2017), https://thenewinquiry.com/white-collar-crime-risk-zones [https://perma.cc/2K85-3VCP].

137. Dieterich et al., *supra* note 49; *see also* Berk, *supra* note 103; Flores et al., *supra* note 49; Skeem & Lowenkamp, *supra* note 38, at 689-90.

138. Skeem & Lowenkamp, *supra* note 38, at 690 (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, 17 (2015) (finding "that age, sex, and race . . . were significantly related to violent recidivism").

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—crimes that, because most crime is intraracial, will disproportionately befall communities of color.[139] No matter how we alter the data or algorithm, then, 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 commit crime.[140] They signify 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.[141] 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.[142] Crime rates are a manifestation of deeper forces; racial variance in crime rates, where it exists, manifests the enduring social and economic inequality produced by centuries of racial subordination.

3. The Broader Framework: Distortion Versus Disparity in the Event of Concern

The two possible sources of racial disparity in past-arrest rates—differential enforcement and differential offending—belong to a broader framework. There

---

139. This was the scenario in the example from the Berk study. *See supra* notes 105-109 and accompanying text.

140. 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.*, KENNEDY, *supra* note 123, at 12-17 (analyzing race relations in the administration of criminal justice); RUSSELL-BROWN, *supra* note 123 (discussing race, crime, and law, beginning with slavery in the United States).

141. *See, e.g.*, MODEL PENAL CODE § 1.02(2) cmt. k (AM. LAW INST., 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"); MEHRSA BARADARAN, THE COLOR OF MONEY: BLACK BANKS AND THE RACIAL WEALTH GAP (2017); KENNEDY, *supra* note 123. This is not to disclaim all individual responsibility for criminal acts. But individual responsibility for particular acts does not amount to group responsibility for group crime rates.

142. *See supra* note 136 and accompanying text.

are always two fundamentally distinct kinds of explanation for intergroup disparities in predictions: (1) distortion in the data or predictive process, and (2) an actual difference, across group lines, in the historical base rate of the event we want to predict.

Distortion can take many forms. In the criminal justice context, the choice of a proxy target variable with racial skew (i.e., "any arrest" as a proxy for "commission of serious crime") may be the most important.[143] But racial distortion can also result if the data are systematically less reliable for one racial group than for another. This problem can arise if the data are simply more limited for one racial group.[144] Another potential source of distortion in prediction is intentional manipulation of the data or algorithm to disadvantage one racial group—what Solon Barocas and Andrew Selbst call "masking."[145] There is no evidence that this is a serious concern in the context of contemporary criminal justice risk assessment.[146] 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 are made public, as they should be, independent researchers can replicate the tool-design and validation process and check for signs 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

---

143. Corbett-Davies and Goel call this problem "label bias" and diagnose it as "perhaps the most serious obstacle facing fair machine learning." Corbett-Davies & Goel, *supra* note 68, at 18.

144. An algorithm developed for maximum accuracy will conform to the majority data, and may be less accurate for members of the underrepresented group. *See, e.g.*, Sue Shellenbarger, *A Crucial Step for Avoiding AI Disasters*, WALL ST. J. (Feb. 13, 2019, 9:57 AM ET), https://www.wsj.com/articles/a-crucial-step-for-avoiding-ai-disasters-11550069865?ns=prod/accounts-wsj [https://perma.cc/C28U-LAAE] (explaining this phenomenon and how diverse development teams are more alert to unrepresentative data sets). 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 data set. *See* Sukarna Barua et al., *MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning*, 26 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENGINEERING 405, 405-06 (2014). For a possible example of this phenomenon, see Hamilton, *supra* note 27 (manuscript at 29, 10), which demonstrates that COMPAS was significantly less accurate for Hispanic than for white defendants by several measures and suggesting that smaller numbers of Hispanic defendants might be the cause.

145. Barocas & Selbst, *supra* note 3, at 692-93. They call it "masking" because machine-learning technologies offer opportunities to intentionally distort an algorithm in ways that are difficult to detect. *Id.*

146. *See* Huq, *supra* note 14, at 1090.

scores differently than did judges in urban counties with a greater black population, with the result that the new process disproportionately benefitted white defendants.[147] In terms of actual outcomes, this potential source of disparity may be the most important of all.

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 doing so in 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.

* * *

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, at least in theory, be eliminated within a risk-assessment system itself. If they cannot, then the very core of the risk-assessment enterprise is compromised, and it should be abandoned. Disparities that flow from differential crime rates cannot be eliminated within the risk-assessment system. Unlike in the case of distortion, however, such disparity does not mean that the project of risk assessment is compromised and should be abandoned. If the data accurately represent crime rates, risk assessment can provide valuable information. That information will be inherently unequal, and so presents a difficult dilemma—but one that is nevertheless important to confront.

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.[148] 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 why it is important to distinguish between distortion and differential offense rates as possible sources of racial disparity in prediction.

---

147. *See* Stevenson, *supra* note 6.

148. *See, e.g.*, Timothy R. Schnacke, *"Model" Bail Laws: Re-Drawing the Line Between Pretrial Release and Detention*, CTR. FOR LEGAL & EVIDENCE-BASED PRACTICES 12-13 (Apr. 18, 2017), https://www.clebp.org/images/04-18-2017_Model_Bail_Laws_CLEPB_.pdf [https://perma.cc/WP33-359T] (emphasizing the importance of defining the relevant risks in the context of pretrial risk assessment).

Whatever the source, though, the three strategies most commonly advocated to redress predictive disparity are off the mark. Part III explains why.

III. NO EASY FIXES

As the risk-assessment-and-race debate accelerates, critics increasingly argue for three strategies to promote racial equity in prediction. The first is the exclusion of both race and factors heavily correlated with race as input variables.[149] 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.[150] In particular, advocates have urged intervention to ensure an equal rate of adverse predictions across racial groups (statistical parity),[151] or equal error rates among those in each racial group who have the same outcome (parity in false-positive and false-negative rates).[152] The discussion here will use the term "algorithmic affirmative action" to refer to these proposals collectively, acknowledging 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.[153]

---

149. *E.g.*, Chander, *supra* note 43, at 1039 (urging advocates to focus on "inputs and outputs" rather than algorithms themselves); Huq, *supra* note 14, at 1080 (discussing "the [p]roblem of [d]istorting [f]eature [s]election"); Danielle Kehl et al., *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y 34 (2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf [https://perma.cc/7V6D-JLLM] ("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.").

150. *See, e.g.*, Chander, *supra* note 43, at 1039-41 (calling for "algorithmic affirmative action").

151. *E.g.*, Corbett-Davies et al., *supra* note 75 (identifying statistical parity as a "popular" definition of fairness in the risk-assessment and algorithmic-fairness literature); Michael Feldman et al., Certifying and Removing Disparate Impact (July 16, 2015) (unpublished manuscript), https://arxiv.org/pdf/1412.3756.pdf [https://perma.cc/NNQ4-NHUH] (an early work in the algorithmic-fairness literature that adopts a statistical-parity metric).

152. *E.g.*, Angwin et al., *supra* note 1 (criticizing disparity in false-positive rates as unjustified bias); *see also* Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in* State v. Loomis, 18 N.C. J.L. & TECH. 75, 86 (2016) (same).

153. *E.g.*, John Ralphing, *Human Rights Watch Advises Against Using Profile-Based Risk Assessment in Bail Reform*, HUM. RTS. WATCH (July 17, 2017, 12:00 AM EDT), https://www.hrw.org/news/2017/07/17/human-rights-watch-advises-against-using-profile-based-risk-assessment-bail-reform [https://perma.cc/95PJ-DBY4]; *Use of Pretrial "Risk Assessment" Instruments*, *supra* note 42.

This Part argues that all three of these strategies are misguided. Though well intentioned, they have the potential to compromise the goal of racial equity rather than to further it.[154]

A. Regulating Input Variables[155]

Input variables are often cited as the primary concern in the quest for racial equity in risk assessment. It is an almost-universal orthodoxy, in fact, that race must be excluded as an input to prediction.[156] Many people extend this principle to variables that correlate with race in a given locale, like zip code.[157] 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 approach to achieving racial equity.[158] 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 Chief 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."[159]

---

154. A comprehensive review of the "fair machine learning" literature by two scholars well versed in the field was developed contemporaneously with this Article, and arrived at much the same conclusions. *See generally* Corbett-Davies & Goel, *supra* note 68 (surveying popular fairness metrics, explaining their limitations, and advocating for "single-threshold" classification rules instead). Aziz Huq has also recently offered a set of nuanced prescriptions for racial equity in algorithmic criminal justice, grounded by the principle that predictive programs should strive to avoid imposing any net burden on communities of color. Huq, *supra* note 14, at 1129; *see also infra* text accompanying note 274-275 (discussing the difference between Huq's proposal and the proposal offered by this Article).

155. I explore this subject matter more comprehensively in a follow-on article: Sandra G. Mayson, Algorithmic Fairness and the Myth of Colorblindness (Jan. 10, 2019) (unpublished manuscript) (on file with author).

156. *See, e.g.*, Starr, *supra* note 30, at 812 ("There appears to be a general consensus that using race would be unconstitutional.").

157. *E.g.*, Corbett-Davies & Goel, *supra* note 68, at 8 ("[S]everal papers have suggested algorithms that enforce a broad notion of anti-classification, which prohibits not only the explicit use of protected traits but also the use of potentially suspect 'proxy' variables.").

158. *Accord id.* at 9-17.

159. Chief Justice Roberts, writing for the plurality in *Parents Involved in Community Schools v. Seattle School District 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) (plurality opinion). Justice Sotomayor rejoined, seven years later, that "[t]he way to stop discrimination on the

A simple example illustrates. When I worked in New Orleans as a public defender, the significance of arrest there 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, however, had three past arrests, 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.[160] A race-blind algorithm would not observe 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 arrests were a much less significant indicator of risk for a black man than for a white man.[161] 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 might therefore 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 people.[162] A few risk-assessment-tool developers have

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, 134 S. Ct. 1623, 1676 (2014) (Sotomayor, J., dissenting).

160. *Cf. New Orleans Investigation*, *supra* note 129, at ix-x (finding "racial disparities in arrests of whites and African Americans in virtually all categories, with particularly dramatic disparity for African-American youth"); *id.* at x ("The level of disparity for youth in New Orleans is so severe and so divergent from nationally reported data that it cannot plausibly be attributed entirely to the underlying rates at which these youth commit crimes . . . .").

161. 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").

162. 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. *See, e.g.*, Harold Berlak, *Race and the Achievement Gap*, *in* CRITICAL SOCIAL ISSUES IN AMERICAN EDUCATION: DEMOCRACY AND MEANING IN A GLOBALIZING WORLD 223, 227 (H. Svi Shapiro & David E. Purpel eds., 3d ed. 2005) (discussing the racial achievement gap

encountered the problem in practice, discovering that variables like past arrests or misdemeanor convictions are less predictive for black people.[163] The usual response is simply to eliminate the problematic input variables from the model. But that solution has a cost in accuracy,[164] which might fall disproportionately on communities of color, as discussed at greater length below.[165]

The alternative is to allow an algorithm to assess the significance of risk factors *contingent on* race. If race moderates the factors' predictive value, this would lower average risk scores for black defendants. It would achieve what a group of computer scientists have dubbed "fairness through awareness."[166] And it would improve, rather than compromise, the tool's accuracy. Under these circumstances, including race as an input variable would promote accuracy and racial equity at the same time.[167] This approach is not feasible for simple checklist tools, but it could be for the machine-learning programs that represent the future.[168]

---

in other standardized tests). 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.

163. Richard Berk and Marie Van Nostrand, along with others, have each reported finding, in different data sets, that past misdemeanor convictions were less predictive of future serious arrest for people of color than for white people. Berk, *supra* note 103, at 183; Christopher T. Lowenkamp et al., *Investigating the Impact of Pretrial Detention on Sentencing Outcomes*, LAURA & JOHN ARNOLD FOUND. (Nov. 2013), https://www.arnoldfoundation.org/wp-content/uploads/2014/02/LJAF_Report_state-sentencing_FNL.pdf [https://perma.cc/L9JF-KEHL]. The Pennsylvania Sentencing Commission recently rejected past arrests entirely as input variables because they had such different predictive significance across racial lines. *Risk Assessment Update: Arrest Scales*, PA. COMMISSION ON SENT'G 4-7 (Feb. 28, 2018), http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/research-and-evaluation-reports/risk-assessment [https://perma.cc/32WY-9F74].

164. The Pennsylvania Commission on Sentencing, for instance, has elected to rely on past conviction rather than past-arrest data despite the fact that it renders the model less accurate overall. *See Risk Assessment Update: Arrest Scales*, *supra* note 163, at 1.

165. *See infra* Section III.B.2.

166. Cynthia Dwork et al., Fairness Through Awareness (Nov. 30, 2011) (unpublished manuscript), https://arxiv.org/pdf/1104.3913.pdf [https://perma.cc/N8QE-27NY].

167. *See* Kim, *supra* note 43, 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."); Lipton et al., *supra* note 103 (arguing on the basis of statistical examples that a prohibition on race or sex data is counterproductive); Corbett-Davies & Goel, *supra* note 68, at 9 (explaining the "[l]imitations of anti-classification" as a fairness metric).

168. *See* Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS & PROC. 22, 23 (2018) (demonstrating, with national data, that including race as an input variable to a machine-learning college-admissions algorithm both "improves predicted GPAs of admitted students" and can increase "the fraction of admitted students who are black").

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.[169] The same is likely true for equalizing cost ratios across racial groups.[170] To achieve predictive parity, it may be necessary to manipulate the data to cancel out the effect of race on other observable variables,[171] or to assess the predictive import of every input variable contingent on race. As Solon Barocas and Andrew Selbst have noted, algorithmic prediction thus offers a particularly clear window on the conflict between anticlassification and antisubordination conceptions of equality.[172]

Yet neither excluding race and race-correlated factors nor including them can equalize outcomes entirely if the event 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 are made available, the facts that correlate with the target variable — and therefore

169. *See, e.g.*, Corbett-Davies et al., *supra* note 75; Hardt et al., *supra* note 85.

170. Berk, *supra* note 103, at 185-86 (explaining that, to equalize cost ratios across racial groups in a juvenile risk-assessment context, the author "separate[d] forecasting exercises" for white and black juveniles, respectively, and that the machine-learning forecasting algorithms the data produced were different for each racial group).

171. There are different ways to attempt this, and many risk-assessment-tool developers do. Marie VanNostrand, who has developed several of the checklist pretrial risk-assessment tools in current use, searches for risk factors that are equally predictive across racial lines and discards those that are not. Telephone Interview with Marie VanNostrand (Oct. 20, 2016) (notes on file with author). This approach is straightforward, but could have a steep cost in overall accuracy. *See* Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RES. 12-18 (July 2, 2018), https://journals.sagepub.com/doi/pdf/10.1177/0049124118782533 [https://perma.cc/LB82-47SB].

172. Barocas & Selbst, *supra* note 3, at 723 (explaining that "[d]ata mining discrimination will force a confrontation between the two divergent principles underlying antidiscrimination law: anticlassification and antisubordination"). For an introduction to anticlassification and antisubordination principles, see, for example, Balkin & Siegel, *supra* note 68, at 10; Helen Norton, *The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality*, 52 WM. & MARY L. REV. 197, 206-15 (2010); and Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 509-15 (2003) (discussing the normative grounds underlying racial-classification decisions). This theme will be explored at greater length in Mayson, *supra* note 155.

become the algorithm's predictors—will also correlate with race because the target variable does.[173] The only way to break the race correlation is by compromising the algorithm's ability to predict the target variable. Excluding criminal-history data, for instance, might dramatically reduce the disparate racial impact of predicting future arrest, but it will also dramatically compromise the algorithm's ability to predict future arrest. To eliminate racial disparity in the prediction of a racially disparate event is to undermine the predictive tool.

Some readers may feel that weakening predictive tools is a good thing. If a tool predicts a race-skewed target variable like "any arrest," for example, the tool has dubious value to begin with. In that situation, though, the better answer is to stop predicting the meaningless event entirely.[174] And if the target variable does *not* embed racial distortion, then undermining the predictive tool can be counterproductive because the loss in accuracy may inflict proportionally more "errors" on black communities than on white ones.[175]

The larger point is that colorblindness is not a meaningful measure of equality. It can exacerbate rather than mitigate racial disparity in prediction.[176] And even if it does mitigate disparity in prediction, that improvement may come at a cost to accuracy that itself has a racially disparate impact. As 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.[177]

B. Equalizing (Some) Outputs

Algorithmic affirmative action has similar shortcomings. As noted, for purposes of this discussion "algorithmic affirmative action" refers to an intervention to produce statistical parity, equal false-positive rates, or equal false-negative

---

173. *See* Corbett-Davies & Goel, *supra* note 68, at 9 (noting that "nearly every covariate commonly used in predictive models is at least partially correlated with protected group status; and in many situations, even strongly correlated").

174. *See infra* Section III.B.1.

175. *See infra* Section III.B.2 and Appendix.

176. *See* Huq, *supra* note 14, at 1100; Kim, *supra* note 43, at 867 ("[I]f 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."); Kroll et al., *supra* note 25, at 693-95.

177. *Cf.* Mayson, *supra* note 155; 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 . . . .").

rates. 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

First, consider 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, then risk scores will overstate the risk posed by black men relative to the risk of actual crime commission. The goal of algorithmic affirmative action is to adjust the data to cancel out this racial distortion in arrest rates.[178]

This strategy presumes that the scale of the distortion is known. If 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.[179] 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 this reality, the more direct solution to the problem is simply to avoid target variables that are likely to be racially skewed with respect to the thing we care about.[180] 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. The average arrest offense is too insignificant to have much probative value, and the racial skew in arrest rates relative to offending rates is too prejudicial.[181]

---

178. *See* Berk, *supra* note 103, at 189 (considering data modifications along these lines); *cf.* Sorelle A. Friedler et al., On the (Im)Possibility of Fairness (Sept. 23, 2016) (unpublished manuscript), https://arxiv.org/pdf/1609.07236 [https://perma.cc/BP4U-N7KM] (raising a similar scenario with respect to SAT scores and college-admissions algorithms designed to assess students' academic potential).

179. *See, e.g.*, Laurel Eckhouse et al., *Layers of Bias: A Unified Approach for Understanding Problems with Risk Assessment*, 46 CRIM. J. & BEHAV. 185, 197 (2019) ("The magnitude and pattern of the bias in the data cannot be measured directly with the techniques used by ProPublica, Northpointe, or any of the others studying these models, including us.").

180. *Principles for the Validation and Use of Personnel Selection Procedures*, SOC'Y FOR INDUS. & ORG. PSYCHOL., INC. 33 (2003), https://www.siop.org/_principles/principles.pdf [https://perma.cc/4FJR-47TC] ("Confidence in the criterion measure is a prerequisite for an analysis of predictive bias.").

181. *See* Mayson, *supra* note 6, at 562; Roberts, *supra* note 117, at 4-13; Schnacke, *supra* note 148, at 110-14; Slobogin, *supra* note 9, at 591; Stevenson & Mayson, *supra* note 43, at 29-31.

Risk-assessment tools should instead predict something closer to the harm we actually want to avoid.[182] The challenge is to identify that harm, both conceptually and in data.[183] I have argued elsewhere that risk-assessment tools should assess the risk of violent crime,[184] but that category is amorphous—does it include burglary? a bar fight? driving under the influence?—and its relative importance is contestable; perhaps we should be equally concerned with the risk of financial crime.[185] The point is that the decision about what to predict is a momentous one and should be made based on law and considered policy judgment, rather than on what data are most readily available.[186]

We should also acknowledge that resorting to more specific target variables may not solve the problem. To begin with, it may not be possible to produce useful predictions of low-frequency events like violent crime. The Pennsylvania Commission on Sentencing, to its credit, recently concluded that it could not predict future violence with sufficient accuracy to justify handing risk scores to judges.[187] When sufficiently accurate prediction is not possible, we should not resort to predicting a more statistically significant event but should simply recognize that our objectives exceed our ability.

Narrow target variables, moreover, may still embed racial distortion with regard to the actual harm of concern. Violent-crime arrest remains an inexact proxy for violent crime itself. Police sometimes arrest the wrong person. Many violent crimes never lead to arrest at all. There may therefore still be racial skew between arrest rates and underlying offense rates. This might be so even if arrest

---

182. For a thoughtful discussion of the "risk of what?" question in the pretrial context, see Schnacke, *supra* note 148, at 110-14.

183. *See* Selbst, *supra* note 3, at 131-33 (characterizing this task as "defin[ing] the problem" for prediction); Schnacke, *supra* note 148, at 110-14.

184. Mayson, *supra* note 6, at 562.

185. *See* Clifton et al., *supra* note 136.

186. As Andrew Selbst notes in his discussion of predictive policing, "Using data mining also tends to bias organizations toward questions that are easier for computers to understand." Selbst, *supra* note 3, at 132.

187. *Development and Validation of the Proposed Risk Assessment Scales*, PA. COMMISSION ON SENT'G 2-3 (May 2018), https://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment/phase-iii-reports/development-and-validation-of-the-risk-assessment-scale [https://perma.cc/7V7V-Z4UW] ("The sentence risk assessment score or category is not intended to be used by the court as an aggravating or mitigating factor."). The Commission postponed adoption of the risk assessment instrument and revised the risk scales based on a June 2018 public hearing. *Revisions to the Proposed Risk Assessment Instrument*, PA. COMMISSION ON SENT'G (Nov. 2018), https://pcs.la.psu.edu/guidelines/proposed-risk-assessment-instrument/additional-information-about-the-proposed-sentence-risk-assessment-instrument/revisions-to-the-proposed-risk-assessment-instrument-november-2018 [https://perma.cc/PMR9-EFY2].

rates track the incidence of reported crimes by race.[188] For example, if white communities report domestic violence with less frequency when it happens, violent-crime reports would embed racial skew relative to the actual rates of offending, and arrest rates that track report rates would carry the distortion forward.

Stated in more general terms, one might object that we can *never* be confident that our target variable is free from racial distortion.[189] We must rely on the past to predict the future, but we see the past only hazily through the splintered lens of data.[190] 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, and 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 our past-crime data are inadequate to serve as the basis for any prediction.[191] 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 past data do not reliably represent the events we want to avoid, we should stop consulting them as a guide for 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. There is a practical argument against such interventions and a deeper conceptual one.

---

188. The correspondence between arrest rates 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, e.g.*, Skeem & Lowenkamp, *supra* note 38, at 690.

189. As Selbst puts it, "[I]t may be impossible to tell *when* the disparate impact truly reflects reality." Selbst, *supra* note 3, at 167; *see also* Barocas & Selbst, *supra* note 3, at 682 ("So long as prior decisions affected by some form of prejudice serve as examples of *correctly* rendered determinations, data mining will necessarily infer rules that exhibit the same prejudice.").

190. *Cf.* 1 *Corinthians* 13:12 ("For now we see through a glass, darkly . . . .").

191. *See* Barocas & Selbst, *supra* note 3, at 682-84; 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, 121 (2013) ("What is not axiomatic is the straightforward application of assumptions about priors . . . to violence risk assessment—that remains a set of important empirical matters."); Selbst, *supra* note 3, at 140-43.

a. Practical Problems

The practical argument against intervention to equalize false-positive and false-negative rates is that it is unlikely 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. Moreover, even if it is possible to equalize both error rates simultaneously, 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.

This might occur because equalizing false-positive and false-negative rates does not mean equalizing the total number of errors for each racial group. Equalizing false-negative rates, rather, means equalizing the *proportion* of rearrests the algorithm misses for each racial group. If the algorithm misses 50% 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 of white defendants. The difference in the absolute number of false negatives could overwhelm any benefit to black communities that flows from equalized false-positive rates.[192] The Appendix below illustrates this possibility with an example drawn from real data.

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

This is to say not that equalizing error rates *will* necessarily increase the net cost of prediction borne by black communities, but that it *might*. It depends on the underlying base rates and what the false-positive and false-negative rates are. There is no basis to think that this metric is systematically more likely than any other to equalize the net burden of prediction. Moreover, if prioritizing equality in error rates has too great a cost in accuracy, it will eliminate the utility of prediction.[193]

---

192. Equalizing false-positive rates will result in fewer 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 high-base-rate group in the first place.

193. 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 7% increase in violent crime.

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. 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).[194] Depending on the error rates 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, constraining an algorithm to equalize cost ratios increased the disparity in the rate of adverse predictions for each racial group, as well as the disparity in false-positive rates.[195]

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 that difference will just produce disparate rates of error, which might increase the net burden on the very communities the intervention was intended to protect.

b. Conceptual Problems

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

Corbett-Davies et al., *supra* note 75, at 802. Furthermore, 17% of those detained would be low-risk people for whom detention was unwarranted. *Id.*

194. 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%). Berk, *supra* note 103, at 189. In addition, Berk found that for whites, the algorithm produced 5.25 false positives for every false negative; whereas for blacks, the algorithm produced 1.85 false negatives for every false positive. *Id.*

195. When Berk trained the algorithm only to optimize for overall accuracy, it forecasted rearrest for 17% of the white subgroup and 33% of the black subgroup (a 16 percentage-point difference); the false-positive rates were 16% for the white subgroup and 28% for the black subgroup (a 12 percentage-point difference). *Id.* at 180. When he altered it to equalize the cost ratios, it forecasted rearrest for 10% of the white subgroup and 29% of the black subgroup (a 19 percentage-point difference); the false-positive rates were 8% for the white subgroup and 22% for the black subgroup (a 14 percentage-point difference). *Id.* at 185.

This argument begins with the nature of equality. Equality is a formal concept. Although not all legal theorists agree that it is an "empty" one,[196] 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 "alike" for purposes of the action at hand.[197] Antidiscrimination laws, for instance, frequently require a claimant to show that she was treated differently from a "similarly situated" person outside the protected class in order to make out a prima facie case of discrimination.[198] To analyze such claims, judges must decide which traits are relevant for comparing one person to another. For purposes of a hiring decision, for example, skill and work experience 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. But 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 a particular action is really a question about the permissible grounds for that action. To judge that two people with equivalent skill and experience are relevantly "alike" for purposes of a hiring decision is to judge that skill and experience are good grounds on which to make such a decision. Likewise, 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 judgment about what constitutes unjustified inequality in some

---

196. 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.").

197. *See, e.g.*, H.L.A. HART, THE CONCEPT OF LAW 159 (Joseph Raz & Penelope A. Bulloch eds., 3d 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 resemblance and differences are relevant, 'Treat like cases alike' must remain an empty form."); SCHAUER, *supra* note 50, 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.").

198. *E.g.*, Tex. Dep't of Cmty. Affairs v. Burdine, 450 U.S. 248, 258 (1981) ("*McDonnell Douglas* teaches that it is the plaintiff's task to demonstrate that similarly situated employees were not treated equally."); Wilson v. B/E Aerospace, Inc., 376 F.3d 1079, 1087 (11th Cir. 2004) ("A plaintiff establishes a prima facie case of disparate treatment by showing that she was a qualified member of a protected class and was subjected to an adverse employment action in contrast with similarly situated employees outside the protected class.").

decision-making process is also a determination about what constitutes legitimate grounds for that decision, and one cannot identify unjustified inequality without choosing, or assuming, some answer to that underlying question.[199]

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 Section I.C described as a "single-threshold rule."

Because it follows from the nature of the activity, a single-threshold rule 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, then the algorithm is failing in the most basic way. Its determinations 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 single-threshold rule 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 lawbreakers. 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 lawbreakers.

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

199. To appreciate this fact in the context of criminal justice risk assessment, notice that the schema of equality metrics laid out in Section 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 (total population impact), for instance, one might inquire about the percentage of black versus white *male* arrestees so classified, or the percentage of black versus white male arrestees under twenty-five who receive that designation, or the percentage of black versus white male arrestees under twenty-five with a prior felony conviction who do. In fact, there is a nearly infinite number of possible equality metrics. That is because the key question for defining a metric—what are the relevant comparators?—admits of a nearly infinite number of answers. And those one deems to be the relevant comparators depend on what one believes to be a legitimate basis for assigning risk.

and prevent is crime, surely the actual occurrence of crime must be the best possible measure of risk!

In fact, this view is deeply incoherent. To hold that ultimate outcomes are what render two people (or groups) alike for purposes of risk assessment is to hold that outcomes are a good basis for risk assessment. But they 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, but if chance plays any role in our lives, they are also *unknowable*. The point is not a technical one—risk-assessment algorithms can be engineered to produce equal false-positive or false-negative rates across racial groups. The point is conceptual. The demand for equal algorithmic treatment for same-outcome groups amounts to a judgment that outcomes are the appropriate basis for prediction. And that judgment is nonsensical.[200]

More concretely, structuring an algorithm to equalize false-positive and false-negative rates will almost certainly violate the principle that people who present the same risk should receive the same risk score (a single-threshold rule). 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 (same-outcome groups), 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.[201] False positives are the group of people of whom we can say in retrospect 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. And even in retrospect, the fact that a risk does not materialize does not mean that a high-risk classification was incorrect. Sometimes high risks do not materialize. That is what differentiates risks from certainties.

---

200. *See* Huq, *supra* note 14, at 1119-22. 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 conditional procedure accuracy equality demands is equality across *groups*: black versus white eventual law abiders and white versus black eventual lawbreakers. *See* sources cited *supra* note 96.

201. *See supra* text accompanying note 83.

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 and conditions it 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 them. It thus rejects the premise of risk assessment—statistically informed action.[202]

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.[203] 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

202. It is worth noting here that predictive parity (calibration) too can be inconsistent with a single-threshold rule. At the very least, it does not guarantee or necessarily indicate a single threshold for risk classification. *See infra* notes 271-272 and accompanying text.

203. 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 5; *see also* Danielle Ensign et al., *Runaway Feedback Loops in Predictive Policing*, 81 PROC. MACHINE LEARNING RES. 1 (2018); Lum & Isaac, *supra* note 128, at 19; Starr, *supra* note 30, at 804-06 (suggesting that risk assessment during sentencing may distort 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—that, in other words, 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 are simply wrong and that 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.

than good. It may actually increase the burden on communities of color, as detailed above. And it might foster deep resentment. It would be better to engage in a frank debate about whether the disparate racial impact of crime prediction outweighs its benefits.

C. Rejecting Algorithmic Methods

The third and increasingly most prevalent strategy for promoting racial equity in prediction is to resist the use of algorithmic methods altogether. In August 2018, more than one hundred civil rights organizations released a joint statement of concerns with pretrial risk assessment. It began: "We believe that jurisdictions should not use risk assessment instruments in pretrial decisionmaking."[204] In Pennsylvania, grassroots advocacy groups have effectively halted the development of a risk-assessment tool for sentencing, notwithstanding a state law requiring the Pennsylvania Commission on Sentencing to create and implement one.[205] Recent advocacy materials urged constituents to "[s]ay NO to [the] racist risk assessment tool," on the ground that the tool was "rooted in the racial disparities already plaguing Pennsylvania's criminal justice system," and "[i]n no circumstance should people's fate within the criminal legal system be determined by an algorithm."[206]

The trouble with this strategy is that the default alternative—subjective risk assessment—is very likely to be worse. Judges engaging in subjective prediction assess the risk of the same events as do algorithmic tools, usually future arrest. They tend to rely on the same factors as actuarial prediction, with the same effect. Any consideration of criminal history, for instance, will entail racial inequality, whether the consideration is actuarial or subjective.[207] On top of this, subjective risk assessment is plagued by a set of pathologies that motivated the turn

---

204. *Use of Pretrial "Risk Assessment" Instruments*, *supra* note 42.

205. Samantha Melamed, *Pa. Officials Spent 8 Years Developing an Algorithm for Sentencing. Now, Lawmakers Want to Scrap It.*, PHILA. INQUIRER (Dec. 12, 2018), https://www.philly.com/news/risk-assessment-sentencing-pennsylvania--20181212.html [https://perma.cc/4XL7-ED77].

206. *Pennsylvania Commission on Sentencing: Say NO to Racist Risk Assessment Tool*, COLOR CHANGE, https://act.colorofchange.org/letter/pa_no_risk_assessment_email_action [https://perma.cc/9GSD-QG6Y].

207. *See* MODEL PENAL CODE § 6B.07(1)(c) (AM. LAW INST., Tentative Draft No. 4, 2016) (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 . . . incorporating offenders' criminal history as a factor relevant to sentencing," giving "particular attention" to whether it "contributes to punishment disparities among racial and ethnic minorities, or other disadvantaged groups"); *id.* § 6B.07 cmt. ("An accumulating body of research indicates

to actuarial tools in the first place. Subjective prediction is vulnerable to irrational bias. A 2016 metareview of risk-assessment instruments used in parole and probation contexts in the United States concluded that "[t]here is overwhelming evidence that risk assessments completed using structured approaches produce estimates that are more reliable and more accurate than unstructured risk assessments."[208] Other recent studies have reached similar conclusions.[209] This is because individual judges may generalize to a greater extent, and with less grounding, than statistical models do.[210] Human beings are prone to cognitive biases that distort rational judgment.[211] In the context of risk assessment, judges may overweight factors that have particular salience to them (including the current charged offense), fall victim to framing effects, and give undue significance to their own past experience.[212]

Irrational cognitive bias can fuel racial inequality in subjective prediction. Individual criminal justice actors tasked with subjective risk assessment may harbor animosity toward one racial group that infects their decision-making. Or the bias may be implicit. A significant and growing body of experimental literature

that criminal-history formulas in sentencing guidelines are responsible for much of the . . . disparities in black and white incarceration rates . . . ."); *id.* (noting that African American defendants appear in criminal courtrooms, on average, with larger numbers of past convictions than white defendants and citing relevant research).

208. Sarah L. Desmarais et al., *Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings*, 13 PSYCHOL. SERVICES 206, 206 (2016).

209. *See, e.g.*, Ben Green & Yiling Chen, Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments (2019) (unpublished manuscript), https://scholar.harvard.edu/files/19-fat.pdf [https://perma.cc/Q8QA-AHHL] (presenting the results of an experimental study in which human subjects "underperformed the risk assessment even when presented with its predictions"); Stevenson & Mayson, *supra* note 43, at 34-35 (describing recent studies suggesting that actuarial risk assessment can improve accuracy of pretrial risk judgments).

210. *See, e.g.*, Hamilton, *supra* note 30, 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."); Starr, *supra* note 30, 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.").

211. *See generally* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (Daniel Kahneman et al. eds., 1982) (reviewing multiple studies on human biases across various judgmental heuristics).

212. *See* Cass R. Sunstein, *Algorithms, Correcting Biases*, SOC. RES. (forthcoming), https://ssrn.com/abstract=3300171 (noting that empirical research on the accuracy of machine versus human predictions suggests the existence of a "current offense bias" that distorts judicial assessments).

has documented the effects of implicit bias in a range of criminal justice settings.[213] There is no reason to think subjective risk assessment is immune. On the contrary, ample and mounting evidence has documented otherwise inexplicable racial disparities in policing, charging, pretrial detention, and sentencing.[214] Notably, two recent studies of risk assessment in action have argued that it was the exercise of human discretion in *responding* to risk-assessment scores that injected racial disparity in outcomes, rather than the risk-assessment scores themselves.[215]

Lastly, subjective risk assessment is far more opaque, and far less accountable, than algorithmic assessment.[216] The human being who judges a person to be a good risk or a bad one may not herself understand why she has done so.[217] Most risk-assessment algorithms, by contrast, can be examined and interrogated; the trend is away from proprietary algorithms and toward transparency.[218] It is therefore 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 deliberations.[219] We can also quantify an algorithm's racial impact

---

213. *See, e.g.*, R. Richard Banks et al., *Discrimination and Implicit Bias in a Racially Unequal Society*, 94 CALIF. L. REV. 1169 (2006); Jennifer L. Eberhardt et al., *Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes*, 17 PSYCHOL. SCI. 383 (2006); David L. Faigman et al., *Implicit Bias in the Courtroom*, 59 UCLA L. REV. 1124 (2012); Justin D. Levinson, *Forgotten Racial Equality: Implicit Bias, Decisionmaking, and Misremembering*, 57 DUKE L.J. 345 (2007); Justin D. Levinson et al., *Guilty by Implicit Bias: The Guilty/Not Guilty Implicit Association Test*, 8 OHIO ST. J. CRIM. L. 187 (2010); L. Song Richardson & Phillip Atiba Goff, *Implicit Racial Bias in Public Defender Triage*, 122 YALE L.J. 2626 (2013).

214. *E.g.*, Eckhouse et al., *supra* note 179, at 202 ("[T]here is substantial evidence that defendants of color are disadvantaged in pretrial and sentencing decisions made without reference to risk-assessment models."); M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Sentences*, 122 J. POL. ECON. 1320, 1323 (2014) (finding that black defendants in the federal system were 1.75 times more likely to face a mandatory minimum charge than similarly situated white defendants); *supra* notes 127-130 (sources reporting and discussing racial disparity in policing actions not explicable by crime rates alone).

215. Green & Chen, *supra* note 209; Stevenson, *supra* note 6, at 53; *cf.* Samuel R. Wiseman, *Fixing Bail*, 84 GEO. WASH. L. REV. 417, 454 (2016) (arguing that "actuarial instruments should limit" rather than inform judicial discretion).

216. *See* Kroll et al., *supra* note 25.

217. For reviews and discussions of research on unconscious biases, see Ralph Richard Banks & Richard Thompson Ford, *(How) Does Unconscious Bias Matter?: Law, Politics, and Racial Inequality*, 58 EMORY L.J. 1053 (2009); John A. Bargh, *Unconscious Thought Theory and Its Discontents: A Critique of the Critiques*, 29 SOC. COGNITION 629 (2011); and Martie G. Haselton et al., *Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias*, 27 SOC. COGNITION 733 (2009).

218. *See* Kroll et al., *supra* note 25.

219. *See*, *e.g.*, *id.*

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.[220] There are hurdles, but the accountability prospects are far better for algorithmic prediction than for subjective prediction.

At least on paper, then, algorithms have distinct advantages over subjective assessments of risk. They eliminate the variability, indeterminacy, and apparent randomness—indeed, the subjectivity—of human prediction that has long pervaded criminal justice. They bring uniformity, transparency, and accountability to the task.

This is not to overstate the case for algorithms. The evidence for the superior accuracy of actuarial over subjective prediction is not watertight; a great deal depends on the algorithm at issue and the details of its use.[221] There is an urgent need for further research to document the comparative effects of the two methods on the ground.[222] It is also 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.[223] 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 does, it

220. *See id.* at 680-82; Selbst, *supra* note 3, at 110, 169-80 (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"); Sarah Holland et al., The Dataset Nutrition Label: A Framework to Drive Higher Quality Standards (May 2018) (unpublished manuscript), https://arxiv.org/pdf/1805.03677.pdf [https://perma.cc/FC79-BQX5] (proposing that data sets be required to include the equivalent of "nutrition labels" that disclose possible demographic skews or systemic inaccuracies in the data); Dillon Reisman et al., *Algorithmic Impact Assessments*, AI NOW INST. (Apr. 2018), https://ainowinstitute.org/aiareport2018.pdf [https://perma.cc/Q6PV-GRJQ].

221. A thoughtful judge with broad experience may be more effective at assessing risk than a rudimentary algorithm, but a sophisticated algorithm 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. *See* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES 1 (2018) (finding that untrained human participants performed nearly as well as COMPAS); Green & Chen, *supra* note 209; Starr, *supra* note 30, 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 6, at 14-19 (surveying existing evidence).

222. *See, e.g.*, Stevenson, *supra* note 6, at 57-58.

223. On the normative judgments that the construction of a risk-assessment algorithm entails, see generally Eaglin, *supra* note 30.

can also inflict damage on a much larger scale.[224] And of course, if algorithmic assessment is imposed *on top of* subjective risk assessment, it is likely to compound the racially disparate effects of both forms of assessment.

Still, given the state of practice and the state of our knowledge, there is every reason to expect that subjective risk assessment produces greater racial disparity than algorithmic risk assessment—and that it does so with less transparency and less potential for accountability or intervention. To the extent that this is true, rejecting algorithmic methods in favor of subjective risk assessment not only will fail to eliminate predictive inequality, but also might exacerbate it. At best, then, rejection of actuarial risk assessment is a superficial measure. At worst, campaigning against algorithms per se might distract from the real problem: the nature of prediction itself. Not only will subjective prediction continue to generate racial disparity, but in the absence of algorithmic methods, the disparity will be harder to see and to redress.

Actuarial risk assessment, in other words, has not created the problem of racially disparate prediction, but rather exposed it. Its contribution is to illuminate—in formal, quantitative terms—the way in which prediction replicates and magnifies inequality in the world. 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."[225] Algorithmic prediction puts the dragon of predictive inequality out on the plain. It is frightful, but at least we can see it. Rejecting the precise mirror of algorithmic prediction in favor of subjective risk assessment does not solve the problem. It merely turns a blind eye.

IV. RETHINKING RISK

The predictive inequality exposed by algorithmic methods should cause us to rethink a central strategy in contemporary U.S. criminal justice: identification and coercive control of the "dangerous." 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 American criminal justice system.[226] Under the "Old Penology," the system's primary goal and

224. *See generally* O'NEIL, *supra* note 3 (chronicling and illustrating the dangers of ostensible scientific objectivity, opacity, entrenchment, and scale).

225. Norval Morris & Marc Miller, *Predictions of Dangerousness*, 6 CRIME & JUST. 1, 2 (1985).

226. Malcolm M. Feeley & Jonathan Simon, *The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications*, 30 CRIMINOLOGY 449, 452, 455 (1992).

responsibility was the adjudication of guilt for specific criminal acts. The New Penology, by contrast, sees the system's primary goal and responsibility as the management of "dangerous groups."[227] Many others have since expanded on this diagnosis.[228] Scholars have long argued that a criminal justice system designed to incapacitate the risky will perpetuate racial injustice. Actuarial analytics illustrate precisely how.

There are two possible responses to the problem. The first is to refute the significance of risk itself, lament the New Penology, and argue for a return to the Old.[229] Regardless of whether such a return is preferable, it is very unlikely to occur. The other option is to accept the significance of risk and prediction in criminal justice decision-making, but to radically rethink its role. This Part argues for the latter approach. It further contends that actuarial risk assessment can—and should—play a central role in changing how the criminal justice system understands and responds to risk.

A. Risk as the Product of Structural Forces

As a mirror of the past, actuarial risk assessment provides a detailed image of the societal distribution of crimes and arrests. To the extent that that image is a picture of race and class disparity, the reason is not mysterious. Aggregate crime and arrest risk of the kind that contemporary criminal justice risk-assessment tools measure—"any arrest" or arrest for a "violent crime"—are functions of disadvantage.

This fact can be lost in individual cases. When a judge confronts a statistically risky person, the risk can seem like a feature of the person himself, something for which he can or should be held responsible. Perhaps this is especially

227. *Id.* at 456.

228. *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 "[t]he preventive state").

229. *E.g.*, Erin Collins, *Punishing Risk*, 107 GEO. L.J. 57, 107 (2018) (arguing that actuarial risk-informed sentencing "distorts traditional sentencing principles, by authorizing and encouraging the consideration of non-culpable and personal characteristics to predict future behavior"); *cf.* HARCOURT, *supra* note 5 (arguing "against prediction" and for randomization as a guiding principle of law enforcement action).

likely when judges are forced to assess risk and blame at the same time, as they are at sentencing.[230]

But judgments of risk are fundamentally different from judgments of blame.[231] Blame is a moral quality; a judgment of blame is a judgment about the moral responsibility a person bears for past choices he has freely made. Risk is an empirical quality, not a moral one. A judgment of risk has no inherent moral import; it is a factual judgment about the likelihood of a given future event. The moral import of the factors that render a person risky is irrelevant to the assessment of risk itself. A victim of circumstances may be as risky as a ruthless manipulator.

Judgments of risk therefore have a very different relationship to punishment than do judgments of blame.[232] The distinguishing features of punishment are the expression of moral condemnation and the purposeful infliction of suffering.[233] A judgment of condemnation is thus a necessary condition for imposing punishment.[234] Blame can authorize punishment, and punishment can be an ap-

---

230. *See* Paul H. Robinson, *Punishing Dangerousness: Cloaking Preventive Detention as Criminal Justice*, 14 HARV. L. REV. 1429, 1444 (2001) (arguing for a full institutional separation of punishment and prevention); *see also* Mayson, *supra* note 228 (arguing for a conceptual, if not institutional, separation).

231. *See* Mayson, *supra* note 228, at 324-27 (elaborating differences between judgments of risk and blame).

232. *Id.* at 317-33 (contrasting concepts of punishment and preventive restraint and exploring the implications of the contrast in legal doctrine and practice).

233. Notwithstanding the many points of dispute in punishment theory, there is broad consensus on this fact. *See, e.g.*, DOUGLAS HUSAK, OVERCRIMINALIZATION 95 (2008) (defining punishment as "the intentional infliction of a stigmatizing deprivation"); Henry M. Hart, Jr., *The Aims of the Criminal Law*, 23 LAW & CONTEMP. PROBS. 401, 404 (1958) ("What distinguishes a criminal from a civil sanction . . . is the judgment of community condemnation which accompanies and justifies its imposition.").

234. *E.g.*, *In re* Winship, 397 U.S. 358, 363-64 (1970) ("[A] society that values the good name and freedom of every individual should not condemn a man for commission of a crime when there is reasonable doubt about his guilt."); Richard S. Frase, *Limiting Retributivism: The Consensus Model of Criminal Punishment*, *in* THE FUTURE OF IMPRISONMENT 143, 144 (Michael Tonry ed., 2004) (noting "substantial agreement" among theorists and practitioners that "concepts of just deserts must place limits on the pursuit of crime control and other consequentialist goals" through punishment); Mayson, *supra* note 228, at 317-21, 327-29 (enumerating criminal-law doctrines that tie punishment to culpability); *cf.* John Rawls, *Two Concepts of Rules*, *in* PUNISHMENT 58, 62 (Joel Feinberg & Hyman Gross eds., 1975) (coining the term "telishment" for hypothetical sanctions not conditioned on culpability, and arguing that "punishment" is limited, by definition, to sanctions for a blameworthy act). Whether a judgment of blame is a *sufficient* condition for punishment is another matter. *See, e.g.*, Douglas Husak, *Lifting the Cloak: Preventive Detention as Punishment*, 48 SAN DIEGO L. REV. 1173, 1201 (2011) ("Sensitivity to the drawbacks of punishment undermines the thesis that desert suffices to justify penal sanctions.").

propriate response to a judgment of blame. Risk, on the other hand, cannot authorize punishment on its own: an empirical assessment cannot authorize a moral response. Nor can punishment be an appropriate response to an empirical fact.

Actuarial risk assessment may help to clarify the difference between judgments of risk and judgments of blame because, in the aggregate, big data and statistical analysis continually demonstrate that both arrests and some categories of crimes are concentrated in marginalized communities. This is to say that arrest and crime risk are products of structural forces. They are functions of disadvantage. Recognition of this fact has two corollaries: (1) predictive algorithms can be deployed "in reverse" to help diagnose areas of risk and need; and (2) one means of reducing risk is to target the structural conditions that produce it. The rest of this Part argues for these two approaches.

B. Algorithmic Prediction as Diagnostic

Because predictive algorithms transparently reflect inequality in the data from which they are built, they can also be deployed in reverse: as diagnostic tools to identify sites and causes of racial disparity in criminal justice. To deploy algorithms in this way, stakeholders and researchers must first audit an algorithm's predictions for racial disparity. If there is such disparity, by any equality metric, the next step is to identify why. It will either be the case that (1) the algorithm is less predictive for one group than for another (as was true for Hispanic defendants in Melissa Hamilton's study of COMPAS); or (2) the algorithm is equally predictive across racial lines and the disparity flows from a difference in the base rate of the predicted event (as was true for black versus white defendants in the ProPublica study).[235]

If the algorithm is less predictive for one racial group, stakeholders and researchers should attempt to figure out why. It may be that the data are more limited or of lower quality for one racial group. That conclusion would support policy efforts to increase data collection, improve data quality, or adjust the data directly to ensure adequate group representation for the development of a revised algorithm. Alternately, the source of the divergence in predictive accuracy might be that, for one group, certain input factors correlate differently with the relevant outcome. As noted above, some risk-assessment-tool developers have found that past arrests and misdemeanor convictions "mean less" about future risk for black people than for other demographic groups.[236] In places where this

235. *See supra* Section I.B; *see also* Hamilton, *supra* note 27. Hamilton evaluated COMPAS's performance on the basis of the same data set as ProPublica.

236. *Supra* notes 163, 171.

is true, it suggests that black communities have been disproportionately subject to past arrest and misdemeanor prosecution relative to rates of offense. This kind of evidence would support policy initiatives to counter overpolicing and overenforcement. More generally, identifying group differences in how input factors correlate with outcomes can help illuminate group differences in the causal pathways that generate crime and arrests.[237] That information is essential to informing crime-reduction policy.

The second possibility is that the algorithm is equally predictive across racial lines, and that the racial disparity in outputs flows mathematically from a difference in the underlying base rate of the predicted event—for instance, in the base rate of rearrest. When this is the case, stakeholders should question why the base-rate difference exists. Are the data simply less accurate for one racial group (that is, are rearrests systematically underreported for one group)? If so, the data should be improved. Or are the data accurate, and the problem is that the event we have chosen to predict is visited with unjustified frequency on minority communities, like arrest for low-level crimes? If that is the case, we should rethink the value of predicting that outcome at all.[238] Base-rate differences that reflect racially skewed enforcement rather than racial variance in crime commission also provide a powerful case for changes to policing policy. Lastly, a divergent base rate might reflect a difference across racial groups that we *do* want predictions to capture, like a difference in offense rates. If the rates for some category of crime vary by race in a given time and place, awareness of that fact is critical to meaningful policy intervention.

To summarize: Instead of discarding the statistical mirror, we could confront the image it reflects and take responsibility for it. In other arenas, data scientists are working to apply machine learning to similar diagnostic ends.[239] Deploying algorithmic prediction as a diagnostic tool would promote racial justice by identifying sites and causes of racial disparity in criminal justice, paving the way for targeted interventions. It would also promote the larger interests of the system

---

237. Hamilton, *supra* note 27, at 11, 29 (noting the likelihood that risk-assessment tools will underperform for Hispanic Americans unless they integrate cultural and situational factors that moderate risk, and concluding that it is "quite likely" that poorer performance of COMPAS for Hispanic defendants was due to failure to integrate "cultural differences" that moderate risk); *see also* Stephane M. Shepherd & Roberto Lewis-Fernandez, *Forensic Risk Assessment and Cultural Diversity: Contemporary Challenges and Future Directions*, 22 PSYCHOL. PUB. POL'Y & L. 427, 498 (2016).

238. *See supra* Section III.B.2.

239. One group of scientists, for instance, is using machine learning to identify adjectives most frequently associated with different ethnic groups over the nineteenth and twentieth centuries to illuminate the history of discrimination. Nikhil Garg et al., *Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes*, 115 PROC. NAT'L ACAD. SCI. U.S. E3635 (2018).

by troubleshooting data problems and by helping to illuminate the causal pathways of crime and arrest risk.

C. A Supportive Response to Risk

Recognizing that arrest and crime risk are the product of structural forces and directly confronting the statistical image of those patterns should lead us to fundamentally reconsider how the criminal justice system responds to risk. In the aggregate, these forms of risk are a function of disadvantage. To target the risky for restraint is thus to target the disadvantaged for more disadvantage. It is to compound the very social conditions that produce crime and arrest risk in the first place.[240]

Two foundational points are important to clarify here. First, as discussed above, a judgment of risk warrants a different kind of response than a judgment of blame. A judgment of risk, by itself, cannot authorize and does not warrant punishment. It is a familiar principle in criminal law and theory that punishment must be limited to blameworthy past acts.[241] Punishment is not an appropriate response to the possibility of a speculative future act. It may be that, when a person is subject to punishment for a culpable past act, risk is a permissible factor to consider in deciding what (deserved) punishment to impose in order to produce the most benefit.[242] But risk itself justifies a utilitarian, rather than a punitive, response. A risk assessment is an empirical judgment of the probability of some future harm occurring, given status quo conditions. The appropriate response is an intervention that minimizes the possibility of a net harm, taking into account any harm the intervention itself inflicts, and maximizes the possibility of a net benefit. A risk judgment alone provides no justification for a responsive intervention to express condemnation or inflict unnecessary suffering.

The second important point is that there is scant evidence on what interventions "work" to manage crime risk at the individual level. Custodial detention incapacitates a person from committing crime outside of the institution during

---

240. *Cf.* Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 105 (Hugh Collins & Tarunabh Khaitan eds., 2018) (arguing that there is a general duty to avoid compounding injustice and that widespread belief in such a duty animates disparate impact discrimination law).

241. *See, e.g.*, SANFORD H. KADISH ET AL., CRIMINAL LAW AND ITS PROCESSES 221-351 (10th ed. 2017) (identifying "culpability" as one of "[t]hree foundational principles [that] limit the imposition of punishment," and exploring actus reus and mens rea doctrines that strive to limit punishment to culpable past acts); *see also* sources cited *supra* notes 233-234.

242. *See, e.g.*, Frase, *supra* note 234 (describing forms of limiting retributivism that would accommodate such practice).

the period of confinement, but it is very costly (in both fiscal and human terms), does not prevent crime within the institution, and may have a criminogenic effect over the long term.[243] Noncustodial forms of restraint like GPS monitoring may deter some crime, but the evidence is weak, and they too impose hefty costs.[244] Robust empirical studies have mostly found that contemporary sex-offender registration and monitoring regimes, intended to prevent sex crimes, may do more harm than good.[245] There is simply very little evidence demonstrating the relative efficacy of different possible individual-scale interventions at reducing serious crime.

These two points suggest a way out of the cycle of compounding disadvantage: the default response to risk need not be coercion. What if it were support instead? Risk, after all, is neither intrinsic nor immutable. It is possible to change the odds.[246] In the short term, a supportive, needs-oriented response to risk would mitigate the immediate racial impact of prediction. If a high-risk classification meant greater access to support and opportunities, a higher false-positive rate among black defendants would be less of a concern. In the long term, a supportive response to risk might help to counter the social conditions that drive crime, for the benefit of all.

1. Objections

A skeptical reader may object at the outset that this argument goes too far. Crime, in the aggregate, is also the product of structural forces. Does that mean no one should be punished for committing one? It cannot be an argument

---

243. For a recent evaluation and synthesis of credible studies on the effect of incarceration on subsequent offending, see David Roodman, *The Impact of Incarceration on Crime*, OPEN PHILANTHROPY PROJECT (2017), https://www.openphilanthropy.org/files/Focus_Areas/Criminal_Justice_Reform/The_impacts_of_incarceration_on_crime_10.pdf [https://perma.cc/NUQ6-46R5]. *See also, e.g.*, Shima Baradaran Baughman, *Costs of Pretrial Detention*, 97 B.U. L. REV. 1 (2017) (estimating costs of detention); Paul Heaton et al., *The Downstream Consequences of Misdemeanor Pretrial Detention*, 69 STAN. L. REV. 711, 759-69 (2017) (finding, inter alia, that pretrial detention increased the likelihood that a defendant would accrue new criminal charges within eighteen months of a bail hearing).

244. *See, e.g.*, Stevenson & Mayson, *supra* note 43, at 45-47 (surveying evidence on efficacy of electronic monitoring and identifying key costs of electronic monitoring).

245. *E.g.*, J.J. Prescott, *Do Sex Offender Registries Make Us Less Safe?*, 35 REGULATION 48 (2012).

246. *Cf.* Patrick Sharkey et al., *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 rate, a 6 percent decline in the violent crime rate, and a 4 percent decline in the property crime rate").

against state coercion that the system compounds disadvantage when it punishes, for that would preclude law enforcement entirely.

But this extension of the point does not undercut it. Indeed, the realization that criminal prosecution and sentencing systematically burden the already disadvantaged, far beyond what is necessary to serve law enforcement goals, has galvanized widespread criminal justice reform over the past half decade. Criminal law and policy *should* account for the fact that coercive punishment will to some extent compound existing disadvantage. This is not an argument against all punishment, because punishment has retributive and expressive goals that may justify some coercion not otherwise justified on consequentialist grounds.[247] Risk assessment and management, however, are purely consequentialist endeavors. The questions they pose are empirical: what is the likelihood of future harm *X* under status quo conditions, and what intervention can minimize that risk at least cost to the individuals concerned and to the public? In that setting, addressing the underlying sources of risk should be the paramount objective.

The skeptical reader may also find the idea of supporting high-risk individuals to be dangerously naïve. Some people pose acute threats of serious harm that nothing but incapacitation can manage. But a default supportive response to risk need not mean obliviousness to danger. As noted above, 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, or even detention, of a few.

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.

2. Theoretical Framework

Proposing a supportive response to risk is not original. As a logical matter, it is what the "least-restrictive-means" principle already 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

247. *But see* Anders Kaye, *Radicalized Risk Assessment*, 36 BEHAV. SCI. & L. 610, 611 (2018) (arguing that "colonization" of criminal justice by risk assessment will illuminate the structural causes of crime to such an extent that it will galvanize a rejection of retributivism).

some version of the least-restrictive-means principle.[248] A supportive response to risk is also built into the "risk-needs-responsivity" (RNR) model prevalent in more mature systems of crime-risk management.[249] A related, newer model of crime-risk management, the Good Lives Model (GLM), is even more explicit in taking a supportive approach. The GLM is "a strengths-based rehabilitation theory that aims to equip clients with internal and external resources to live a good or better life—a life that is socially acceptable and personally meaningful."[250] Whereas the RNR model directs social workers or other practitioners to identify and address the risky individual's "criminogenic needs," the GLM directs practitioners to identify "internal or external barriers toward living a good life," which can then be "addressed within the broader strengths-based framework."[251]

All of these models justify, and arguably require, a default supportive response to risk. When a person is identified as presenting a substantial risk (but not so acute that immediate incapacitation is necessary), the core inquiry should be: "What can we do to help you succeed at *X*, such that harm *Y* does not transpire? What do you need?" It is of no great import whether this approach is characterized as pursuing the least-restrictive method of risk management, addressing criminogenic needs, or dismantling internal and external barriers to a better life. The support offered could include assistance in obtaining housing, education, training, employment, or counseling; accessing social services; obtaining a driver's license or restoring a suspended one; or pursuing medical, substance-abuse, or mental-health treatment.[252]

---

248. *See, e.g.*, 18 U.S.C. § 3553 (2018) (requiring judges to impose a sentence that is "sufficient, but not greater than necessary" to accomplish the goals of punishment); STANDARDS FOR CRIMINAL JUSTICE: PRETRIAL RELEASE § 10-1.2 (AM. BAR ASS'N 2007) (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"); Richard S. Frase, *Sentencing Principles in Theory and Practice*, 22 CRIME & JUST. 363, 364 (1997) (explaining the "parsimony" principle).

249. *See, e.g.*, Devon L.L. Polaschek, *An Appraisal of the Risk-Need-Responsivity (RNR) Model of Offender Rehabilitation and Its Application in Correctional Treatment*, 17 LEGAL & CRIMINOLOGICAL PSYCHOL. 1 (2012); *Corrections in Ontario: Directions for Reform*, INDEP. REV. ONT. CORRECTIONS 110 (Sept. 2017), https://www.mcscs.jus.gov.on.ca/sites/default/files/content/mcscs/docs/Corrections%20in%20Ontario%2C%20Directions%20for%20Reform.pdf [https://perma.cc/T2ED-SPG5].

250. Tony Ward et al., *The Good Lives Model and the Risk Needs Responsivity Model: A Critical Response to Andrews, Bonta, and Wormith (2011)*, 39 CRIM. JUST. & BEHAV. 94, 95 (2012).

251. *Id.*

252. *Cf.* Glenn A. Grant, *2017 Report to the Governor and the Legislature*, N.J. CTS. 26 (Feb. 2018), https://www.njcourts.gov/courts/assets/criminal/2017cjrannual.pdf [https://perma.cc/YT4W-8EWQ] ("Even when they are not court-ordered to submit to treatment or other

More broadly, recognizing that risk is the product of social conditions should lead us to seek responses that directly address those conditions—an approach Kelly Hannah-Moffat has called a "socio-structural analysis of risk."[253] As Tim Goddard and Randolph R. Myers have suggested, risk-assessment tools that identify geographic areas of high risk, or high-risk demographic groups, could justify community-support programs and targeted private or public investments in schools, jobs, housing, and economic development.[254] To the extent that a risk profile reflects law enforcement practice more than the likelihood of serious crime, risk-management measures could include modifications to law enforcement practice.[255] One simple structural intervention that would dramatically reduce risks of nonappearance and rearrest in the pretrial phase would be to dedicate the resources necessary to adjudicate cases more promptly.

3. Examples

A shift toward a default supportive response to risk would certainly present a practical challenge. The ascendant policing model known as "focused deterrence" offers a cautionary tale. It 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 police both to offer a carrot—increased social support—and to threaten a stick—increased punishment for even small criminal infractions—to those targeted. In practice, the carrot tends to get lost.[256] 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.

services, many defendants on pretrial release request assistance in areas such as mental health and drug treatment.").

253. Kelly Hannah-Moffat, *A Conceptual Kaleidoscope: Contemplating "Dynamic Structural Risk" and an Uncoupling of Risk from Need*, 22 PSYCHOL., CRIME & L. 33, 35 (2016).

254. Tim Goddard & Randolph R. Myers, *Against Evidence-Based Oppression: Marginalized Youth and the Politics of Risk-Based Assessment and Intervention*, 21 THEORETICAL CRIMINOLOGY 151, 162 (2017); *see also* Sharkey et al., *supra* note 246, at 1233-36.

255. Goddard & Myers, *supra* note 254, at 162; Hannah-Moffat, *supra* note 253, at 35 (arguing that modifications to criminal justice practices that "produce systemic conditions for recidivism . . . could make a measurable difference in recidivism and other correctional efficiencies").

256. *See, e.g.*, Selbst, *supra* note 3, at 142-43 (noting that early evidence on the program's implementation in Chicago suggests that the support did not happen).

But a shift in the way the system responds to risk is nonetheless achievable. There are programs across the nation that have already implemented strategies of support as a first-line response to risk, and that might offer useful models.

One such program is Supervision to Aid Reentry (STAR), an unusual reentry court operated by Magistrate Judge Rice of the U.S. District Court for the Eastern District of Pennsylvania and Judge Restrepo of the U.S. Court of Appeals for the Third Circuit.[257] Unlike most existing reentry courts, the program targets those designated as risky: it is open only to "returning citizens . . . who pose a *medium-to-high risk of recidivism for violent crime*," as assessed by the Federal Post-Conviction Risk Assessment Tool.[258] At each session, the participants update the presiding judge on their lives and challenges. The presiding judge listens. He asks how the court can help.

And help it does. Working collaboratively with federal probation officers, the U.S. Attorney's office, the federal defender, and a team of volunteers, Judges Rice and Restrepo strive to eliminate whatever obstacle is impeding smooth reentry. The reentry-court team assists participants in securing housing, employment, training, counseling, benefits, education, credit, and treatment. One partner organization helps participants clear their records of old arrests; another employs participants in the restoration of abandoned homes. A local university's psychology department has developed an intensive cognitive-behavioral therapy program for participants who opt in. Law school clinic students help participants navigate the labyrinthine traffic court to handle old fines, restore suspended licenses, and obtain new ones. Perhaps most importantly, the court has created a deep sense of community among the participants and the myriad individuals and organizations involved in its efforts. The program entails sanctions, too, for failure to abide by its relatively loose conditions, but the overwhelming emphasis is on support. In its eleven years, 9% of the program's graduates and 15% of all participants have been rearrested or had their parole revoked, as compared to 35.8% for similarly situated individuals not enrolled.[259]

Other models include the Harlem Parole Reentry Court, which "links parolees to a range of social services, including drug treatment, vocational services, and mental health treatment," and offers the same referrals to family members

257. By way of full disclosure, I volunteered with this program as a law clerk for Judge Restrepo.

258. Memorandum from L. Felipe Restrepo, Judge, U.S. Court of Appeals for the Third Circuit, & Timothy R. Rice, Magistrate Judge, U.S. District Court for the E. Dist. of Pa., to Lawrence F. Stengel, Chief Judge, U.S. Court of Appeals for the Third Circuit, Annual Report: Reentry Court Program 1 (July 17, 2018), http://www.fbacrimphila.org/files/2018-annual-report.doc [https://perma.cc/GR4S-QZVM] (emphasis added).

259. *Id.*

"[w]here appropriate . . . to help increase stability in the home."[260] This kind of holistic reentry support appears to be part of a trend.[261] In the pretrial context, there is considerable momentum toward more supportive risk-management approaches. In Tulsa, Oklahoma, the public defender's office is collaborating with the organization Uptrust to connect pretrial defendants with supportive services.[262] Community bail funds, the Good Call hotline, Silicon Valley De-Bug, and the recent mass bailout organized in New York City have also aspired to reduce risk through supportive interventions.[263] New York's "Supervised Release" program, piloted in 2013 and expanded citywide in 2016, secures the release of defendants who would otherwise remain detained pending trial.[264] Licensed caseworkers and social workers assess participants' needs and goals and then provide voluntary social-services referrals as well as court-date reminders by phone and text message. There is a light supervision component (check-ins that range from monthly to weekly, depending on the person), but the emphasis is on support, and early results have been promising.[265]

---

260. *Harlem Reentry Court*, CTR. FOR CT. INNOVATION, https://www.courtinnovation.org/programs/harlem-reentry-court [https://perma.cc/68CX-3NHM].

261. Jeff Mellow & Kevin Barnes-Ceeney, *Key Factors to Promote Successful Comprehensive Reentry Initiatives*, 81 FED. PROB. 22, 22 (2017) ("More and more, these collaborative efforts take the form of comprehensive or multi-faceted reentry initiatives that focus on strategic system-level change . . . .").

262. *See* Corey Jones, *Twist on Innovative Smartphone App Aims to Help Poor Tulsa County Defendants Make Court Dates*, TULSA WORLD (Oct. 28, 2018), https://www.tulsaworld.com/news/courts/twist-on-innovative-smartphone-app-aims-to-help-poor-tulsa/article_a03efe4c-46bd-5a5c-a641-d1d2bc5fe6d7.html [https://perma.cc/JUN4-8N2X].

263. *See* Jocelyn Simonson, *Bail Nullification*, 115 MICH. L. REV. 585, 603-04 (2017) (describing how different bail funds support individuals they bail out); Jeffrey C. Mays, *105 New York City Inmates Freed in Bail Reform Experiment*, N.Y. TIMES (Nov. 20, 2018), https://www.nytimes.com/2018/11/20/nyregion/bail-reform-rikers-rfk-nyc.html [https://perma.cc/Q8EL-X4J5] (noting that a mass bail-out effort included housing and transportation assistance for those released, as well as use of a cell phone and text-message court reminders); Ashley Southall, *Bronx Hotline Helps People Make the Right Call After an Arrest*, N.Y. TIMES (Dec. 22, 2017), https://www.nytimes.com/2017/07/23/nyregion/bronx-hotline-helps-people-make-the-right-call-after-an-arrest.html [https://perma.cc/CW8G-XMDX]; Raj Jayadev, *The Future of Pretrial Justice Is Not Money Bail or System Supervision — It's Freedom and Community*, SILICON VALLEY DE-BUG (Apr. 4, 2019), https://siliconvalleydebug.org/stories/the-future-of-pretrial-justice-is-not-money-bail-or-system-supervision-it-s-freedom-and-community [https://perma.cc/E8K9-WU5C] (explaining, among other things, Silicon Valley De-Bug's Community Release Project).

264. N.Y.C. Mayor's Office of Criminal Justice, A Guide to Pretrial Supervised Release (unpublished guidance) (on file with author).

265. Between July and September of 2018, court-appearance rates for program participants in the five boroughs ranged from 85% to 92%, and at least 90% of participants in each borough

As these examples show, supportive strategies to reduce crime and arrest risk are feasible. And the requisite political will to develop them might actually exist. The national mood toward returning citizens, for instance, has undergone a definite shift. For decades, legislatures bought political capital by codifying employment barriers and other civil disabilities for people with past convictions. They justified these laws as public-safety measures. By contrast, in the first five months of 2018, "21 states . . . enacted laws to improve opportunities for people with a criminal record."[266] Even President Trump has signed on to the "second chance" agenda.[267] 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.[268] This is not the same as targeting at-risk people for support, but it is a step in the right direction.

As scholars of algorithmic fairness have observed in other contexts, whether algorithms exacerbate or mitigate social inequality is entirely a function of the uses to which they are put.[269] If algorithms targeted the disadvantaged for support rather than for further disadvantage, their effects in the world would be very different. A supportive response to risk would not only serve to prevent new crimes and arrests; it would dramatically mitigate the harm of racial disparity in prediction and, over time, help to mitigate the structural inequalities that give rise to racially disparate risk patterns in the first place.

---

avoided felony rearrest. *Supervised Release Quarterly Scorecard*, N.Y.C. MAYOR'S OFF. CRIM. JUST. 2 (Oct. 2018), https://criminaljustice.cityofnewyork.us/wp-content/uploads/2018/11/SR_11.14.18.pdf [https://perma.cc/5GAD-ULV5].

266. CCRC Staff, *More States Enact "Second Chance" Reforms*, COLLATERAL CONSEQUENCES RESOURCES CTR. (June 11, 2018), https://ccresourcecenter.org/2018/06/11/three-more-states-enact-major-second-chance-reforms [https://perma.cc/7LG9-ZA55].

267. Donald J. Trump, *President Donald J. Trump Proclaims April 2018 as Second Chance Month*, WHITE HOUSE (Mar. 30, 2018), https://www.whitehouse.gov/presidential-actions/president-donald-j-trump-proclaims-april-2018-second-chance-month [https://perma.cc/2Q4F-VKRV].

268. 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 a court should order an in-depth presentencing report containing risk and needs information. *Risk Assessment Project Phase III: The Development and Validation of the Proposed Risk Assessment Scales*, PA. COMMISSION ON SENT'G (May 2018), http://www.hominid.psu.edu/specialty_programs/pacs/publications-and-research/risk-assessment/phase-iii-reports/development-and-validation-of-the-risk-assessment-scale/view [https://perma.cc/4AX6-PYTV].

269. *See, e.g.*, O'NEIL, *supra* note 3.

D. The Case for Predictive Honesty

To serve as a diagnostic tool for supportive interventions, risk assessment must measure existing risk patterns as faithfully as possible across racial lines. This requires that risk-assessment tools meet three metrics of predictive equality.

First, individuals who pose the same statistical risk should receive the same risk score regardless of their race (a "single-threshold rule" for risk classification). As discussed above, this is a sine qua non of risk assessment itself. Second, to the extent possible with a single-threshold rule in place, a given risk score should communicate the same average risk regardless of the race of the person to whom it applies ("predictive parity" or "calibration").[270]

These two metrics may sound similar, but they are not coextensive. Assigning the same risk score to all those who present the same risk will not necessarily produce predictive parity,[271] and an algorithm might achieve predictive parity without assigning the same risk score to all who present the same risk.[272] But a single-threshold rule 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. A single-threshold rule 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 of race. These two metrics are achievable in combination, furthermore, even if base rates of the predicted outcome differ across racial lines.

Third, a predictive algorithm should order individuals along a spectrum of risk with equal accuracy for each racial group (i.e., have equivalent AUC scores by race). An algorithm's AUC score relates to how well it differentiates between individuals who present differing degrees of risk. Although this measure says nothing about the meaning of an individual risk score, it is a valuable measure of an algorithm's utility overall, and algorithms should have equal utility across racial lines. Equality in AUC scores is at least potentially compatible with a single-threshold rule and predictive parity.

None of these equity metrics, nor any combination of them, renders an algorithm race neutral. On the contrary, achieving them may require race-conscious choices in the construction of the algorithm. And if the base rate of the predicted outcome differs across racial groups, the algorithm will still predict it

---

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

271. 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, then predictive parity will not necessarily result.

272. Corbett-Davies et al., *supra* note 75, at 798.

more frequently for the high-base-rate group. If base rates differ, an algorithm that achieves these equality metrics will also produce unequal false-positive and/or false-negative rates. The call for "predictive honesty" thus privileges racial equality in the accuracy of a tool's risk assessments themselves over statistical parity or parity in false-positive or -negative rates.

If the risk-assessment tool is deployed for diagnostic or supportive purposes, though, racial inequality in the false-positive rate or total rate of positive predictions should be less of a concern because a positive prediction does not mean an additional burden on the person deemed risky. On the contrary, if a positive prediction means support, a higher rate of total positives or false positives is an asset.

The proposal here should highlight the critical distinction between risk assessment—the estimation of the likelihood of a future harm under status quo conditions—and action taken in response. Risk-assessment tools only purport to measure risk, and only under status quo conditions. They do not decide what action to take in response. Policy makers do.

For algorithms tasked not with measurement but rather with *allocating* some benefit or burden directly, the analysis of what equality measure to prioritize might look different. Equality in the accuracy of an allocation algorithm's determinations might be relatively less important. (It can be difficult to specify what "accuracy" means in an allocation context with multiple values at stake anyway.) Risk-assessment instruments should strive to assess risk as precisely as possible. Decision-making about how to respond to risk should strive both to maximize the net benefit of policy interventions to society in general and to struggling communities in particular, and to minimize the net harm.[273]

The distinction between risk assessment itself and the action taken in response helps to explain the divergence between the present proposal and the one recently offered by Aziz Huq. Huq begins from the premise that criminal justice risk-assessment tools are "mechanisms to allocate coercion within the criminal justice system."[274] His analysis is addressed to equality in the allocation of coercion rather than equality in the assessment of risk per se. If one takes account of this difference, Huq's proposal and mine are not necessarily inconsistent. In another new contribution, Deborah Hellman urges greater attention to the difference between algorithms that tell us what to believe and algorithms that tell us

273. *Cf.* Corbett-Davies & Goel, *supra* note 68, at 8 (laying out a "utility-based framework" to reason about equality in the context of allocation algorithms and identifying conditions under which a single-threshold classification rule might not be optimal).

274. Huq, *supra* note 14, at 1169.

what to do.[275] Criminal justice risk-assessment tools purport only to tell us what to believe, and they can tell us what to believe about the risk of future events only under status quo conditions. It is up to us to decide what to do—and whether to do something that reinforces the status quo or rectifies it.

CONCLUSION

On June 6, 2018, the Pennsylvania Commission on Sentencing held a public hearing in Philadelphia on the newly proposed Pennsylvania Risk Assessment Tool for sentencing.[276] The room was packed. One by one, community members walked to the lectern and delivered impassioned pleas against adoption of the tool.[277] 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—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.

Algorithmic methods have revealed 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. As 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 intergroup disparity in prediction—

275. Deborah Hellman, *Measures of Algorithmic Fairness*, 106 VA. L. REV. (forthcoming 2020) (on file with author).

276. *See Proposed Risk Assessment Instrument*, PA. COMMISSION ON SENT'G, http://www.hominid.psu.edu/specialty_programs/pacs/guidelines/proposed-risk-assessment-instrument [https://perma.cc/U9S5-Q6EX].

277. *See* Palmer & Irizarry-Aponte, *supra* note 19.

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 or algorithmic process will differ by context.[278] So too 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.[279] Algorithms that drive internet search engines might be programmed to maximize the credibility of top results or minimize representational harms.[280] Algorithms used to calculate lost-earnings damages in wrongful-death suits should perhaps have objectives other than reflecting status quo earning patterns.[281] Not all algorithms, in other words, should faithfully mirror the past.

Given the frenzied uptake of criminal justice risk assessment and the furious resistance it has engendered, the present moment is crucial. The next few years will likely set the course of criminal justice risk assessment for decades to come. 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 social inequality that the predictive mirror reflects. That path should include a radical revision of how the criminal justice system understands and responds to crime risk. There is an opportunity now, with risk assessment and race in the public eye, to take it.

---

278. It may be even more challenging in other arenas to find a target variable that does not 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 nondiscriminatory environment.

279. *See* Corbett-Davies et al., *supra* note 75, at 805 (citing SCOTT E. PAGE, THE DIFFERENCE: HOW THE POWER OF DIVERSITY CREATES BETTER GROUPS, FIRMS, SCHOOLS, AND SOCIETIES (2008)).

280. *See* NOBLE, *supra* note 3, at 104.

281. *See* Kimberly A. Yuracko & Ronen Avraham, *Valuing Black Lives: A Constitutional Challenge to the Use of Race-Based Tables in Calculating Tort Damages*, 106 CALIF. L. REV. 325, 330-33 (2018).

APPENDIX: THE PRACTICAL CASE AGAINST ALGORITHMIC AFFIRMATIVE ACTION—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 data set than among the white juveniles.[282] For every 1,000 black juveniles, 140 were rearrested and 860 were not. For every 1,000 white juveniles, 40 were rearrested and 960 were not.[283] Say the false-positive rate (proportion of eventual non-rearrestees mistakenly forecast for rearrest) is 10% for each group. For every 1,000 white juveniles, 96 (of the 960) non-rearrestees will be mistakenly forecast for arrest. For every 1,000 black juveniles, 86 (of the 860) non-rearrestees will be mistakenly forecast for arrest. Equal false-positive rates mean fewer false positives per capita for black juveniles because there are fewer non-rearrestees to start with.

But what if the false-negative rate (the proportion of eventual rearrests the algorithm misses) is 80% for each group? Then the algorithm will miss 112 (of the 140) rearrests per 1,000 black juveniles but only 32 (of the 40) rearrests per 1,000 white juveniles. Equal false-negative rates mean 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 1,000 white kids and 198 for every 1,000 black kids. The overall error rate for black juveniles will be significantly higher.

Now, the algorithm also produces greater per capita benefits for black communities because it successfully predicts a greater number of the black juvenile rearrests.[284] Nonetheless, the greater total error rate overwhelms the greater per capita benefit. The result is a higher net cost to black communities. The following charts in Figure 5 illustrate this point.

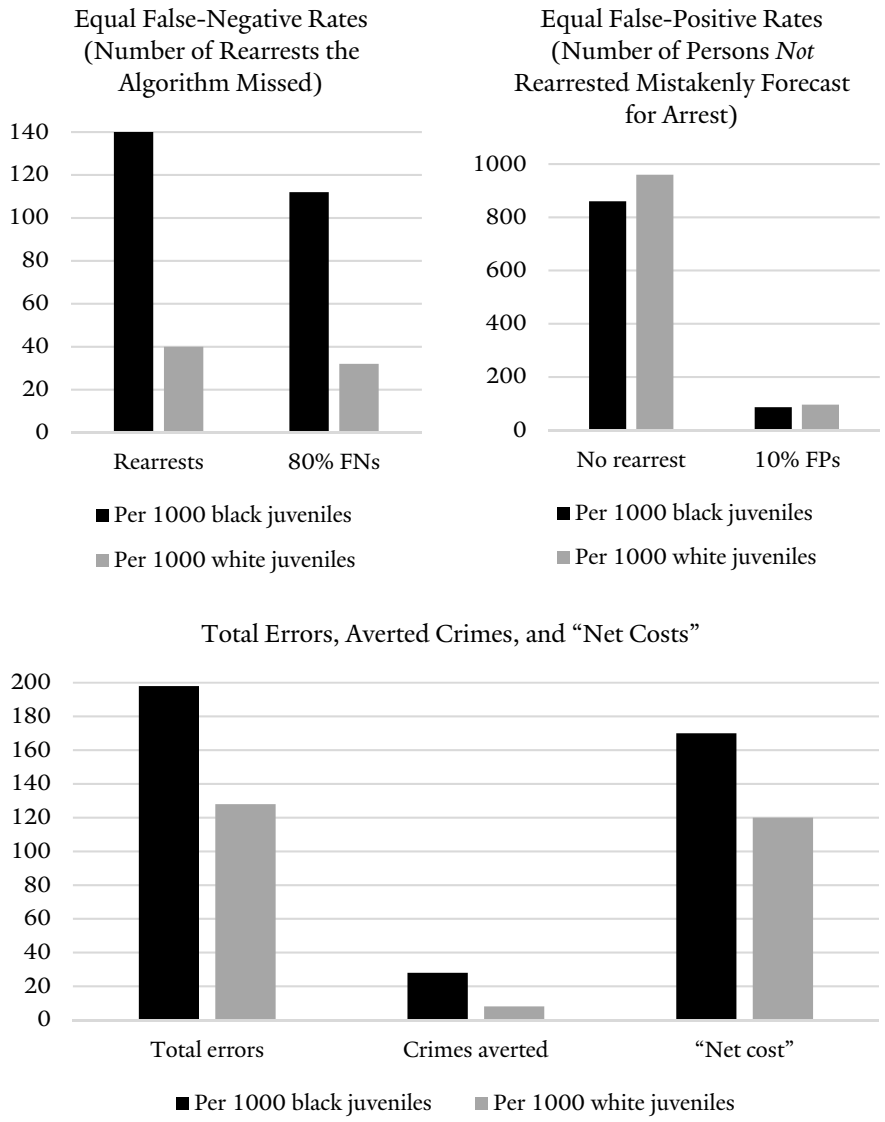
---

282. Berk, *supra* note 103, at 180.

283. *Id.*

284. This is on the assumption that violent-crime arrest corresponds to violent crime, and that violent crime is intraracial.

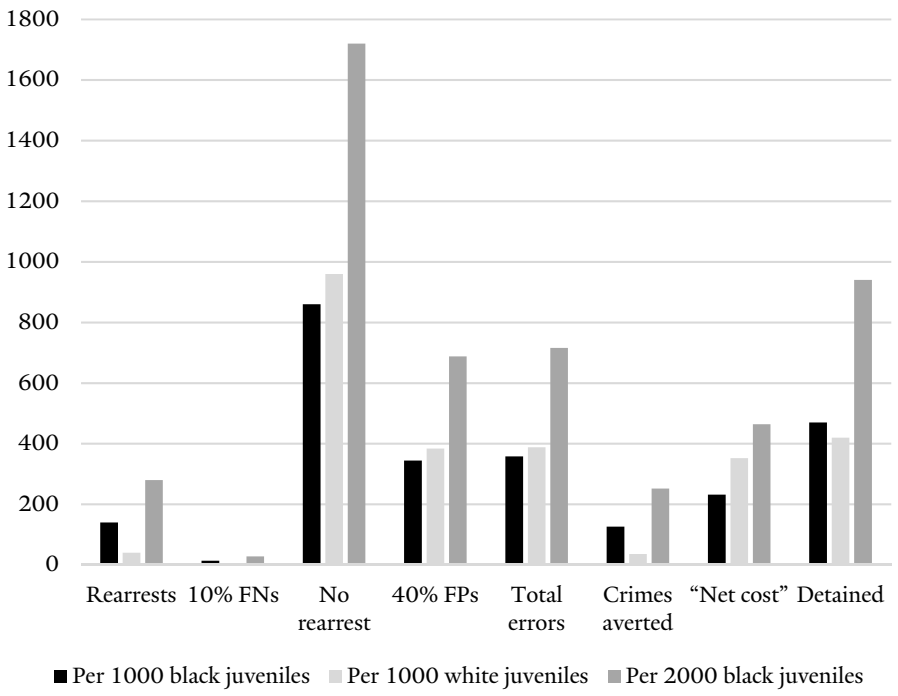
FIGURE 5.
HIGH FALSE-NEGATIVE RATES CAN PRODUCE UNEQUAL "NET COSTS"



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 than white people in the system. The example above assumed that

there were equal numbers of black and white juveniles in the data set. But suppose that twice as many black juveniles are arrested. 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 a 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 juveniles in the system as white juveniles.

FIGURE 6.
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 imposes 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.