2019

Racial Equity in Algorithmic Criminal Justice

Aziz Huq

Recommended Citation

Aziz Huq, "Racial Equity in Algorithmic Criminal Justice," 68 Duke Law Journal 1043 (2019).

Duke Law Journal

VOLUME 68 MARCH 2019 NUMBER 6

RACIAL EQUITY IN ALGORITHMIC CRIMINAL JUSTICE

AZIZ Z. HUQ†

ABSTRACT

Algorithmic tools for predicting violence and criminality are increasingly deployed in policing, bail, and sentencing. Scholarly attention to date has focused on these tools' procedural due process implications. This Article considers their interaction with the enduring racial dimensions of the criminal justice system. I consider two alternative lenses for evaluating the racial effects of algorithmic criminal justice: constitutional doctrine and emerging technical standards of "algorithmic fairness." I argue first that constitutional doctrine is poorly suited to the task. It often fails to capture the full spectrum of racial issues that can arise in the use of algorithmic tools in criminal justice. Emerging technical standards of algorithmic fairness are at least attentive to the specifics of the relevant technology. But the technical literature has failed to grapple with how, or whether, various technical conceptions of fairness track policy-significant consequences. Drawing on the technical literature, I propose a reformulated metric for considering racial equity concerns in algorithmic design: Rather than asking about abstract definitions of fairness, a criminal justice algorithm should be evaluated in terms of its long-term, dynamic effects on racial stratification. The metric of nondiscrimination for an

† Frank and Bernice J. Greenberg Professor of Law, University of Chicago Law School. Thanks in particular to Sharad Goel, Ravi Shroff, and Sam Corbett-Davies for their help over several conversations, in which they patiently explained the interaction of machine learning and statistical concepts to me. Emily Berman, Vincent Chiao, Jessica Clarke, Nancy King, Kiel Brennan Marquez, Debbie Hellman, Andrew Selbst, and Chris Slobogin all gave me helpful comments that improved my thinking and corrected my errors. I was also helped greatly by workshops at Vanderbilt Law School, Northwestern Law School, and the University of Toronto Law School. Faith Laken provided terrific research assistance. The editors of the *Duke Law Journal* did superlative work editing this article. I am very grateful to them. All remaining errors are mine alone.

algorithmically assigned form of state coercion should focus on the net burden thereby placed on a racial minority.

TABLE OF CONTENTS

INTRODUCTION

From the cotton gin to the camera phone, new technologies have scrambled, invigorated, and refashioned the terms on which the state coerces. Today, we are in the throes of another major reconfiguration. Police, courts, and parole boards across the country are turning to sophisticated algorithmic instruments to guide decisions about the where, whom, and when of law enforcement.[1] New predictive algorithms trawl immense quantities of data, exploit massive computational power, and leverage new machine-learning technologies to generate predictions no human could conjure. These tools are likely to have enduring effects on the criminal justice system. Yet law remains far behind in thinking through the difficult questions that arise when machine learning substitutes for human discretion.

My aim in this Article is to isolate one important design margin for evaluating algorithmic criminal justice: the effect of algorithmic criminal justice tools on *racial equity*. I use this capacious term to capture the complex ways in which the state's use of a technology can implicate normative and legal concerns related to racial dynamics. The Article considers a number of ways in which legal scholars and computer scientists have theorized how criminal justice interacts with racial patterning in practice. It analyzes the utility of each lens for evaluating new algorithmic technologies. A primary lesson concerns the parameter that best captures racial equity concerns in an algorithmic setting: I suggest that the leading metrics advanced by computer scientists are not sufficient, and propose an alternative. A secondary lesson relates to the fit between problems of race in the algorithmic context on the one hand, and legal or technical conceptions of equality on the other. Reflection on technological change, that is, casts light on the approaching desuetude of equal protection doctrine.

Racial equity merits a discrete, detailed inquiry given the fraught racial history of American criminal justice institutions.[2] Since the turn of the twentieth century, public arguments about criminality have been entangled, often invidiously, with generalizations about race and the putative criminality of racial minorities.[3] Today, pigmentation

1. Reed E. Hundt, *Making No Secrets About It*, 10 ISJLP 581, 588 (2014) ("[The G]overnment now routinely asks computers to suggest who has committed crimes.").

2. This is a familiar thought. Matthew Desmond & Mustafa Emirbayr, *To Imagine and Pursue Racial Justice*, 15 RACE ETHNICITY & EDUC. 259, 268 (2012) ("One of the most racially unjust institutions today in American society is the American criminal justice system.").

3. *See generally* KHALIL GIBRAN MUHAMMAD, THE CONDEMNATION OF BLACKNESS:

regrettably remains for many people a de facto proxy for criminality. That proxy distorts everything from residential patterns to labor market opportunities.[4] Police respond to black and white suspects in different ways.[5] So do judges and prosecutors.[6] Partly as a result of these dynamics, roughly one in three black men (and one in five Latino men) will be incarcerated during their lifetime.[7] At the same time, the criminal justice system imposes substantial socioeconomic costs on minority citizens not directly touched by policing or prosecutions. In particular, minority children of the incarcerated bear an unconscionable burden as a result of separation from their parents.[8] More generally, there is substantial evidence that spillover costs of producing public safety fall disproportionately on minority groups.[9] As

RACE, CRIME, AND THE MAKING OF MODERN URBAN AMERICA (2010) (exploring how at the beginning of the twentieth century, policymakers in northern cities began linking crime to African Americans on the basis of genetic and predispositional arguments).

4. *See, e.g.*, Lincoln Quillian & Devah Pager, *Black Neighbors, Higher Crime? The Role of Racial Stereotypes in Evaluations of Neighborhood Crime*, 107 AM. J. SOC. 717, 718 (2001) (finding "that the percentage of a neighborhood's black population, particularly . . . young black men, is significantly associated with perceptions of the severity of the neighborhood's crime problem"); Robert J. Sampson & Stephen W. Raudenbush, *Seeing Disorder: Neighborhood Stigma and the Social Construction of "Broken Windows*,*"* 67 SOC. PSYCHOL. Q. 319, 319–23 (2004) (finding that perceptions of disorder in a neighborhood were better predicted by the racial composition of a neighborhood than by actual disorder).

5. For evidence, see CHARLES R. EPP, STEVEN MAYNARD-MOODY & DONALD HAIDER-MARKEL, PULLED OVER: HOW POLICE STOPS DEFINE RACE AND CITIZENSHIP 32–33 (2014) (describing the use of racial profiling in drug arrests); Aziz Z. Huq, *The Consequences of Disparate Policing: Evaluating Stop and Frisk as a Modality of Urban Policing*, 101 MINN. L. REV. 2397, 2407–13 (2017) [hereinafter Huq, *Disparate Policing*] (discussing evidence of such disparities in police stop-and-frisks).

6. For two different perspectives, emphasizing intentional bias and disparate racial impacts, see Richard S. Frase, *What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations?*, 38 CRIME & JUST. 201, 265 (2009) (finding that "seemingly legitimate sentencing factors such as criminal history scoring can have strongly disparate impacts on nonwhite defendants"); Sonja B. Starr & M. Marit Rehavi, *Mandatory Sentencing and Racial Disparity: Assessing the Role of Prosecutors and the Effects of* Booker, 123 YALE L.J. 2, 25–30 (2013) (documenting racial disparities in federal prosecutorial charging decisions related to the application of mandatory minimum sentences in drug cases).

7. *See* Cassia Spohn, *Race, Crime, and Punishment in the Twentieth and Twenty-First Centuries*, 44 CRIME & JUST. 49, 55 (2015) (noting that in 2001 "the chances of ever going to prison were highest among black males (32.2 percent) and Hispanic males (17.2 percent)"); *see also* BRUCE WESTERN, PUNISHMENT AND INEQUALITY IN AMERICA 31–39 (2006) (describing the growth of the incarcerated population over time and describing racial inequalities).

8. *See* SARA WAKEFIELD & CHRISTOPHER WILDEMAN, CHILDREN OF THE PRISON BOOM: MASS INCARCERATION AND THE FUTURE OF AMERICAN INEQUALITY 41 (2014) (discussing the racially disparate spillover effects of paternal and maternal incarceration on children).

9. *See id.* (noting the disproportionate number of minority children in foster care, for instance).

a result, criminal justice elicits racial stratification.[10] Such downstream consequences of existing criminal justice institutions raise weighty moral and legal questions.[11] Even if one demurs to the analogy commonly drawn between our criminal justice system and early twentieth-century debt peonage,[12] it is clear that the criminal justice system is an institution in which racial identity has meaningful effects and that these in turn have influences on the role that race plays in larger American society.[13] In crude terms, it can be both racist and race making.

To sharpen this point, it is useful to have at hand two examples of how new technologies can prompt debates about racial equity. I present the first at greater length because it has become a focal point in public debates. First, the Correctional Offender Management Profiling for Alternative Sanctions ("COMPAS") software application, created by the Northpointe Institution for Public Management, is used across the country to inform bail and parole decisions. COMPAS is organized around an algorithm that uses the answers to some 137 questions about a criminal suspect to rank them on a scale of 1 to 10.[14] This scale is supposed to capture the suspect's risk of reoffending and violent recidivism, with higher scores indicating a greater risk of recidivism. In 2016, journalists from the ProPublica organization did a quantitative analysis of COMPAS scores for roughly ten thousand people arrested and evaluated in Broward County,

10. For a synoptic view of this claim that is dated, but still insightful, see generally RANDALL KENNEDY, RACE, CRIME, AND THE LAW (1997).

11. I think it is important for legal scholars to be candid in distinguishing their normative judgments from their analytic, doctrinal, and empirical claims. The following paragraph states my normative position; it is a premise of what follows, not a conclusion I seek to defend here. *See infra* Part III.A (further defending this position).

12. *See, e.g.*, MICHELLE ALEXANDER, THE NEW JIM CROW: MASS INCARCERATION IN THE AGE OF COLORBLINDNESS (2010) (comparing the modern criminal justice system to the Jim Crow South). For nuanced criticism of Alexander's paradigm, see James Forman, Jr., *Racial Critiques of Mass Incarceration: Beyond the New Jim Crow*, 87 N.Y.U. L. REV. 21, 42–43 (2012).

13. *See* James Q. Whitman, *Equality in Criminal Law: The Two Divergent Western Roads*, 1 J. LEGAL ANALYSIS 119, 122 (2009) (describing the various stages of the criminal justice system through which minorities face discriminatory treatment).

14. Ed Young, *A Popular Algorithm Is No Better at Predicting Crimes Than Random People*, ATLANTIC (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646 [https://perma.cc/NY69-DF3L]. COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions. For descriptions of the COMPAS algorithm, *see* EQUIVANT, PRACTITIONERS GUIDE TO COMPAS CORE 1–2 (2017), http://www.equivant.com/assets/img/content/Practitioners_Guide_COMPASCore_121917.pdf [https://perma.cc/7ML8-NC9U]; *see also In re* Hawthorne v. Stanford, 22 N.Y.S. 3d 640, 641–42 (N.Y. App. Div. 2016) (describing the COMPAS assessment tool).

Florida. By comparing COMPAS scores to a person's behavior in the two years after bail was granted, ProPublica was able to evaluate the instrument's accuracy and, in particular, to investigate whether it had differential effects on different racial groups.

ProPublica estimated that the COMPAS instrument correctly predicted recidivism rates 61 percent of the time and violent recidivism rates 20 percent of the time.[15] ProPublica also concluded that the algorithm "was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants."[16] To reach this conclusion, ProPublica isolated the group of black suspects who had not reoffended in the two years following their evaluation. It found that 45 percent of that group was labeled high risk by the algorithm.[17] ProPublica then looked at the group of white suspects who had not reoffended and found that only 23 percent of that group had been labeled high risk. In other words, the ratio of false positives to true negatives within the pool of defendants who did not go on to recidivate was higher for blacks than for whites.[18] Correspondingly, ProPublica also found that the ratio of false negatives to true positives was lower for whites than for blacks.[19]

Not surprisingly, the company responded by sharply contesting ProPublica's analysis. Northpointe data scientists insisted that COMPAS was well calibrated in the sense that white and black defendants assigned the same risk score were equally likely to recidivate.[20] This constituted evidence, the company argued, that

15. Jeff Larsen, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm [https://perma.cc/YS62-AXYX]. Note that rates of violent crime tend to be so low that an "accurate" instrument would be one that simply classified everyone as low risk.

16. Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/6L7T-ELPG]. ProPublica treated "medium" and "high" risk rankings as higher risk.

17. Larsen et al., *supra* note 15. This disparity remained once ProPublica controlled for "prior crimes, future recidivism, age, and gender." *Id.*

18. *Id.*

19. *Id.*

20. *See* WILLIAM DIETERICH, CHRISTINA MENDOZA & TIM BRENNAN, NORTHPOINTE INC. RESEARCH DEPT., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 3 (2016), https://assets.documentcloud.org/documents/2998391/ProPublica-Commentary-Final-070616.pdf [https://perma.cc/GGR7-TJPV] (flagging "equal discriminative ability" of the algorithm for blacks and whites); *see also* Anthony W. Flores, Kristin Bechtel & Christopher T. Lowenkamp, *False Positives, False Negatives, and False Analyses: A Rejoinder to*

where it mattered to the imposition of state coercion (i.e., where there was a prediction of high risk), the COMPAS algorithm had equal error rates across groups. In addition, Northpointe made a number of (sharply contested) technical complaints about ProPublica's analysis related to the way it accounted for base recidivism rates and how it cut its sample between low- and high-risk defendants.[21] These complaints lacked the force of Northpointe's central claim—that its risk predictions were equally accurate where it counted, regardless of race. This dialogue was not the end of the matter. Other analysts raised a cautionary flag to warn against accepting the terms of the debate as framed by ProPublica and Northpointe: Something more complex, they worried, seemed at stake, although they did not explain fully how to resolve these problems.[22] As a result, the debate about COMPAS—and in particular the question of which measure of fairness should be used to evaluate a predictive algorithm—persists as a locus for normative concern.

A second example of the race-related questions potentially raised by algorithmic criminal justice arises in the policing context, where officers are increasingly using such tools in determining where to deploy and whom to apprehend.[23] In Chicago, police faced with a wave of deadly street violence[24] have deployed a "Strategic Subjects List," or

"Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.," 80 FED. PROBATION 38, 38 (2016) (describing the ProPublica analysis as "faulty"). For a different result using a reconstruction of the COMPAS algorithm, see Razieh Nabi & Ilya Shpitser, *Fair Inference on Outcomes*, *in* THE THIRTY-SECOND AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE 1931, 1938 (2018), https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/download/16683/15898 [https://perma.cc/FW6V-XZ5E].

21. *See* DIETERICH, MENDOZA & BRENNAN, *supra* note 20, at 32–33.

22. *See, e.g.*, Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.f8164ea2cd2c [https://perma.cc/63EU-HSMZ]; Matthias Spielkamp, *Inspecting Algorithms for Bias*, MIT TECH. REV. (June 12, 2017), https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias [https://perma.cc/W3QW-XNW3]. For my analysis of the COMPAS algorithm, see *infra* notes 335–38 and accompanying text.

23. Mara Hvistendahl, *Can 'Predictive Policing' Prevent Crime Before It Happens?*, SCIENCE (Sept. 28, 2016, 9:00 AM), http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens [https://perma.cc/L7BZ-46M3] (noting the adoption of policing tools "which incorporate everything from minor crime reports to criminals' Facebook profiles"); *see also* Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109, 1120–44 (2017) (providing a careful catalogue of predictive policing tools).

24. Monica Davey, *Chicago Tactics Put Major Dent in Killing Trend*, N.Y. TIMES (June 11, 2013), https://www.nytimes.com/2013/06/11/us/chicago-homicides-fall-by-34-percent-so-far-this-

SSL. This is an algorithm developed by data scientists at the Illinois Institute of Technology using U.S. Department of Justice funds.[25] The SSL ranks individuals known to police for the risk of involvement in a shooting using eight data points.[26] Its aim, according to the Chief of Organizational Development for the department, was "to figure out now . . . how does that data inform what happens in the future."[27] Yet despite the fact that the SSL algorithm explicitly accounted for neither race nor gender,[28] interventions based on SSL were quickly condemned for directing disproportionate attention to African American men.[29] Other algorithms that guide the allocation of policing resources on geographic rather than individual terms have elicited kindred concerns about racial targeting.[30]

Questions about algorithmic criminal justice are poised to become more complex. COMPAS and the SSL are both relatively straightforward instruments. Each applies a fixed regression equation with a limited array of parameters to a static data set. Advances in what is called *machine learning*, however, will soon render this sort of tool passé. Machine learning is a "general purpose technology"[31] that, in broad terms, encompasses "algorithms and systems that improve their

---

year.html [https://perma.cc/ZC7R-M2MH].

25. City of Chicago, *Strategic Subject List: Public Safety*, CHI DATA PORTAL, https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np [https://perma.cc/VYQ6-3FKS].

26. *Id.*

27. Jeremy Gorner, *Chicago Police Use 'Heat List' As Strategy to Prevent Violence*, CHI. TRIB. (Aug. 21, 2013), http://articles.chicagotribune.com/2013-08-21/news/ct-met-heat-list-20130821_1_ chicago-police-commander-andrew-papachristos-heat-list [https://perma.cc/R3DR-GKAP].

28. City of Chicago, *supra* note 25. Other predictive policing instruments, however, do explicitly account for suspects' race. DAVID ROBINSON & LOGAN KOEPKE, UPTURN, STUCK IN A PATTERN: EARLY EVIDENCE ON "PREDICTIVE POLICING" AND CIVIL RIGHTS 3–5 (2016), https://www.upturn.org/static/reports/2016/stuck-in-a-pattern/files/Upturn_-_Stuck_In_a_ Pattern_v.1.01.pdf [https://perma.cc/8VXS-2A76].

29. *See, e.g.*, Matt Stroud, *The Minority Report: Chicago's New Police Computer Predicts Crimes, But Is It Racist?*, THE 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/JK3Y-5HMG].

30. Justin Jouvenal, *Police Are Using Software to Predict Crime. Is It a 'Holy Grail' or Biased Against Minorities?*, WASH. POST (Nov. 17, 2016), https://www.washingtonpost.com/local/public-safety/police-are-using-software-to-predict-crime-is-it-a-holy-grail-or-biased-against-minorities/2016/11/17/525a6649-0472-440a-aae1-b283aa8e5de8_story.html?utm_term=.72a9d2eb22ae [https://perma.cc/MZ5W-GCTS].

31. Erik Brynjolfsson & Tom Mitchell, *What Can Machine Learning Do? Workforce Implications*, 358 SCIENCE 1530, 1530 (2017).

knowledge or performance with experience."[32] So, a standard supervised machine-learning instrument—the species of machine learning likely most relevant in the criminal justice space[33]—begins with a so-called training set of examples that are "labeled" with some parameter values. The algorithm examines relations between various parameters associated with those examples to develop a wholly new criterion to classify new examples.[34] Unlike more familiar econometric tools such as regression analysis, a supervised machine-learning process classifies on the basis of *rules that the algorithm itself has developed*. Refining this process, the subset of machine-learning tools called "deep learning" deploy multilayered processes, account for billions of data points, and constantly adjust their classification rule.[35] Machine learning is now being deployed, for instance, in Cambridge, Massachusetts, to predict house burglaries,[36] and in Durham, England, to predict individual recidivism.[37] Deep learning is used in facial recognition and machine translation; it will likely find new uses as its capabilities are better understood. My use of the term "algorithmic criminal justice" is intended to capture both existing instruments, such as COMPAS and the SSL, and also machine-learning (including deep-learning) tools that are likely to be deployed for prediction purposes in the future. Such synoptic consideration is warranted because all of

---

32. PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 3 (2012); ETHEM ALPAYDIN, INTRODUCTION TO MACHINE LEARNING 2 (3d ed. 2014); *see also infra* notes 74–83 and accompanying text (providing a fuller account of machine learning); *infra* notes 85–90 and accompanying text (discussing deep learning).

33. *See* Susan Athey, *Beyond Prediction: Using Big Data for Policy Problems*, 355 SCIENCE 483 (2017) (noting the use of structured machine learning to solve prediction problems).

34. COMM. ON THE ANALYSIS OF MASSIVE DATA ET AL., NAT'L RESEARCH COUNCIL OF THE NAT'L ACAD. OF SCI., FRONTIERS IN MASSIVE DATA ANALYSIS 104 (2013) (noting that in supervised learning, the analyst must actively specify a variable of interest); Athey, *supra* note 33, at 483 (explaining that machine-learning "programs take as input training data sets and estimate or 'learn' parameters that can be used to make predictions on new data"); M. I. Jordan & T. M. Mitchell, *Machine Learning: Trends, Perspectives, and Prospects*, 349 SCIENCE 255, 257 (2015) (defining supervised learning as a process in which "the training data take the form of a collection of (x, y) pairs and the goal is to produce a prediction y^* in response to a query x^*").

35. Jordan & Mitchell, *supra* note 34, at 257. Deep learning uses a process called stochastic gradient descent to improve predictive quality continuously. Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 437 (2015).

36. Cynthia Rudin, *Predictive Policing: Using Machine Learning to Detect Patterns of Crime*, WIRED (Aug. 22, 2013), https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime [https://perma.cc/4U26-3BB9].

37. Chris Baraniuk, *Durham Police AI to Help with Custody Decisions*, BBC NEWS (May 10, 2017), http://www.bbc.com/news/technology-39857645 [https://perma.cc/26MV-F4ST]. See *infra* Part I.C for a catalog of more examples of how machine learning is used in the criminal justice context.

these tools leverage historical data to generate predictions for new, out-of-sample data.[38]

Algorithmic tools in criminal justice are worth isolating for a careful legal analysis for a number of reasons. They are likely to soon become pervasive. They are also that rare instrumentality of state power in respect to which normative intuitions remain inchoate and hence malleable. They represent a qualitative change from the crude evaluative tools embodied in present bail and sentencing practices. These build on imprecise measures of recidivism risk, fail to account for immediate or downstream costs, and cannot be calibrated with the precision of emerging tools. The precision enabled by the algorithmic turn pries open a substantively new domain of policy-design possibilities. Finally, an analysis of algorithmic tools has more general lessons for our equal protection jurisprudence—or at least so I shall argue.

Two distinct analytic frameworks in use now could be used to evaluate the racial effects of machine-learning tools in criminal justice. The first derives from constitutional law. The second is found in the computer science literature on algorithm design.[39] Neither, in my view, is up to the task. The constitutional law of racial inequality directs attention to trivial or irrelevant design margins; it is at times counterproductive. In contrast, technical discussions of algorithmic fairness have yielded a dazzling array of parameters that capture different elements of an algorithm's operation. But as the debate between ProPublica and Northpointe shows, the computer science literature has generated no clear consensus about which parameter *matters*. This Article fills the gap left by the irrelevance of constitutional law and the undertheorization of computer science. It offers a novel, normatively grounded, and empirically pertinent framework for thinking about racial equity in this emerging technological context.[40]

38. I use the term prediction not because all of these instruments aim at the future. Rather, the term captures the possibility that one data set will be used to generate an instrument for drawing inferences about a different sample of data. It is a prediction in the sense of being an out-of-sample estimate.

39. I will not work through all of the relevant computer science literature here. For a brief survey that touches on some of the questions analyzed here, see Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 682–90 (2017).

40. To the extent that algorithmic tools are more generally replacing diffuse human discretion, my reconceptualization of equality norms may have more general application.

Consider first the current constitutional framework for the regulation of race effects in policing. The doctrine, in rough paraphrase, has two main prongs. One concern in the jurisprudence turns on the use of "racial stereotypes or animus" held by individual actors.[41] A focus on animus or stereotypes, though, doesn't easily translate into contexts in which an algorithm blends data streams to estimate unknown parameter values. At best, a concern with intent captures a subset of deeply problematic cases in which data inputs are tainted. Worse, while these cases are likely to be common in practice, it is not clear that contemporary doctrine is up to the task of flagging them. Second, equal protection doctrine is also concerned with the use of racial classifications. But in the emergent context of algorithmic criminal justice, where decision rules are computed endogenously from historical data and then applied without being broadcast to the public, the expressive or distortive harms of racial classifications may well not be present. An algorithm's use of racial data is unlikely to stigmatize or otherwise impose any harm putatively linked to the use of suspect classifications. Eliminating such criteria, moreover, can leave actual outcomes unchanged. Worse, it can generate needless public safety–related costs. This is because algorithmic use of a proscribed criterion, such as race, might in some instances improve the quality of predictions. Thinking about equal protection jurisprudence in relation to algorithmic criminal justice therefore suggests that the former is not a coherent or morally acute metric. This mismatch is likely to have wider significance as algorithms are increasingly substituted for human judgment in criminal justice and beyond.

If constitutional law provides no creditable guidance, what of the burgeoning computer science scholarship on "algorithmic fairness" and "algorithmic discrimination," terms to date used to cover a number of different means of evaluating predictive tools?[42] At a very high level of abstraction, the technical literature usefully distinguishes between two different ways in which race effects might emerge in algorithmic criminal justice. The first is the use of racially tainted historical data to build an algorithm. For example, a policing algorithm used to predict who will be involved in crime, such as the SSL, might employ data

41. Peña-Rodriguez v. Colorado, 137 S. Ct. 855, 869 (2017); *see also* Aziz Z. Huq, *What is Discriminatory Intent?*, 103 CORNELL L. REV. 1211, 1225–31 (2018) [hereinafter Huq, *Discriminatory Intent*] (analyzing the central role of intent in the context of the Fourteenth Amendment).

42. For a survey of the relevant work, see *infra* Part II.

gathered by police, such as records of past street stops or past arrests. If the pattern of this historical policing activity is informed by racial considerations, then the algorithm's predictions will be accordingly skewed. Fixing this first problem of polluted training data is straightforward in theory but often quite difficult in practice. As several legal scholars have noted, algorithms can in theory always be constructed without tainted training data.[43] Whatever considerable difficulties this might present in terms of implementation, it raises no great theoretical impediment.

But the second way in which a racial problem can arise from the use of algorithmic tools does present a theoretical obstacle.[44] It turns on the possibility that an algorithm will generate patterns of error that are systematically skewed between racial groups. As the debate between ProPublica and Northpointe illustrates, however, there is more than one way of measuring errors and more than one way of thinking about racial skewing. Indeed, the computer science literature has generated a plethora of possible metrics. Simplifying this literature by stripping away redundant and irrelevant conceptual trappings, I suggest that an analysis of racial equity might focus on one of four different parameters.

First, one might simply look at whether equal fractions of each racial group are labeled as risky, such that they will be subject to additional policing or detention. Where risk is measured as a continuous variable, this would mean looking at whether the average risk scores of different racial groups varied. *Second*, one might ask whether the same classification rule is being used to assign racial groups to the high-risk category. This condition is satisfied if the same numerical risk score is used as a cutoff for all groups. *Third*, one might separate each racial group and then look at the rate of false positives conditional on being categorized as high risk. This is the parameter that Northpointe stressed. And *fourth*, one might separate each racial group and ask how frequently false positives are conditional on being

43. *See, e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023, 1039 (2017) (suggesting that algorithmic discrimination may be addressed with more transparency about inputs and outputs); Kate Crawford, *Artificial Intelligence's White Guy Problem*, N.Y. TIMES (June 25, 2016), https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html [https://perma.cc/HJ2D-TUG4]; Kroll et al., *supra* note 39, at 680 ("[A]lgorithms that include some type of machine learning can lead to discriminatory results if the algorithms are trained on historical examples that reflect past prejudice or implicit bias . . . .").

44. *Accountable Algorithms* recognizes that "machine learning models can build in discrimination through choices in how models are constructed." Kroll et al., *supra* note 39, at 681. This is not, however, the central focus of their wide-ranging and useful analysis.

in fact a low-risk person. This is the parameter ProPublica underscored.

Each of these metrics tracks a subtly different conception of nondiscrimination. So which fits best a normatively relevant conception of racial equity? The question is complicated by two considerations. First, there is an irreconcilable tension in practice between the first and second criteria. If the average risk score of two racial groups diverge, it is not possible to use the same classification rule and also to ensure that an equal fraction of each group is categorized as high risk. That is, the same risk threshold applied to different populations yields different results. Second, computer science scholars (in collaboration with legal scholars, including myself) have developed in the past two years an impossibility result concerning the third and fourth metrics. Under most empirically plausible conditions, a risk instrument cannot satisfy both the third and the fourth criterion. That is, if the proportion of false positives as a fraction of all positives is equalized between races, then the ratio of low-risk individuals subject to coercion will diverge between the two groups. There is hence an irreconcilable tension (in many feasible states of the world) between having equally accurate predictions of high risk and equalizing the rates of false positives within the pool of nonrecidivist suspects.

To prioritize between these conceptions of racial equity, it is necessary to give an account of the normative stakes of racial equity in criminal law. In the ordinary course, we might look to constitutional law to this end. But we have already seen that constitutional law does not provide a fit or tractable frame for analysis. I thus return to first principles. In my view, the primary reason for concern with racial equity in the algorithmic criminal justice context is that efforts to suppress crime entrench wider social patterns of racial stratification. In important part, stratification effects arise because of the asymmetrical spillovers from criminal justice for minority but not majority populations. A parameter for measuring racial equity, therefore, should track this causal effect of criminal justice on racial stratification.

An algorithm that recommends coercion for a member of the subordinated racial group at the margin when it is not justified in terms of benefits to that racial group will likely increase racial stratification. When coercion of the marginal minority group member is unjustified, it imposes a net burden on the minority group, thus compounding social stratification. Further, if the majority group does not benefit from the policy, or if its net gain is less than the costs imposed on the minority group, that policy is also socially inefficient.[45] I suspect that governments often overestimate the crime suppression benefits of coercive actions while underestimating their costs. Racial equity is therefore served in the first instance today by ratcheting coercion down to socially optimal levels[46] and then by selecting for criminal justice tools that do not burden minority groups.

In designing an algorithm, this intuition must be translated into instructions for the classification protocol. As a rough first cut, this might be done differently for serious and less serious crimes. For serious violent and property crimes, the most important costs and benefits of crime (and crime prevention) accrue directly to the perpetrator and the victim. Spillovers are small by comparison. In these conditions, a single, socially optimal classification rule will advance racial equity and satisfy an efficiency criterion. Rates of false positives, underscored by ProPublica and Northpointe, are less relevant. For less serious crimes and misdemeanors, however, empirical studies identify large spillover costs asymmetrically imposed on minority but not majority communities. At the margin, these spillovers mean that coercion of the minority is both less likely to be efficient and more likely to generate racial stratification. Accordingly, a bifurcated classification rule using different risk thresholds for differently stratified racial groups is appropriate to account for asymmetrical spillovers.

Plural risk thresholds may be socially efficient and racially just, but they confront practical and legal hurdles. First, evaluating algorithmic tools in light of social externalities will require much more information about downstream costs than is presently available. Governments have been woefully deficient in collecting such data; existing risk assessment instruments embody information about recidivism risk but include neither the direct nor the indirect costs of criminal justice coercion.[47]

---

45. Only if the gains to a majority group exceed the costs to a minority group is there a tension between efficiency and racial equity. As I explain below, I think it is plausible to prioritize equality norms in many of these conflicts.

46. In using the term "social efficiency," I mean to capture a static (and in my view naïve) account of welfare that looks only to proximate costs and benefits. It is my view that racial stratification is plausibly described as an "inefficient" equilibrium to the extent that it dissipates large amounts of human capital while inflicting onerous psychological and stigmatic burdens. But since my view is not orthodox, I do not insist on it here and instead use "efficiency" in its more common sense.

47. *See* Christopher Slobogin, *Principles of Risk Assessment: Sentencing and Policing*, 15 OHIO ST. J. CRIM. L. 583, 583 (2018) (proposing certain guiding principles for better risk assessment but not touching on the issue of coercion).

This is a large epistemic void that scholars can fill. It is at least possible that other big-data tools will be important in this regard. Second, the use of racially bifurcated thresholds would raise constitutional concerns akin to those engendered by affirmative action programs. But to the extent current doctrine mandates an outcome that is both socially inefficient and also racially iniquitous, it is the doctrine that is indefensible.

Some limitations on my analysis in this Article should be flagged up front. First, I should again underscore that the costs and benefits of algorithmic tools vary depending on where in the criminal justice process they are deployed. My aim here is to set out a general framework; it is not to pass judgment on any particular computational tool. Second, this Article does not address the integration of algorithmic outputs into individualized suspicion determinations under the Fourth Amendment[48] or the issues related to procedural due process rights from the Fifth and Fourteenth Amendments.[49] These constitutional rules engage different elements of algorithm design. For example, an important recent article develops a concept of "procedural regularity" to ensure that algorithmic decisions are "made using consistently applied standards and practices."[50] This is a meaningful concern. But it is distinct from racial equity. I also do not address the statutory standard supplied by Title VII of the Civil Rights Act of 1964. This has been a topic in other valuable recent work on algorithmic

48. *See, e.g.*, Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1254 (2017) (arguing that for predictions to be used as a basis for searches under the Fourth Amendment, they have to be "intelligible," in the sense of being amenable to explanation); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 383–84 (2015); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 929 (2016) (developing a "framework" for integrating machine-learning technologies into Fourth Amendment analysis). Judicial consideration of this issue has been limited. *Cf.* Commonwealth v. Smith, 709 S.E.2d 139, 143 (Va. 2011) (relying on constructive knowledge doctrine to allow officer use of a predictive algorithm in a *Terry* stop).

49. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256–57 (2008) (criticizing the "crudeness" of then-extant algorithms).

50. Kroll et al., *supra* note 39, at 637–38; *see* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward A Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 109 (2014) (arguing for "procedural data due process [to] regulate the fairness of Big Data's analytical processes with regard to how they use personal data . . . in any adjudicative process").

justice.[51] Nor do I address algorithms' use outside the criminal justice context.[52]

Finally, my conclusions diverge from those of one prominent article that examines the racial effects of a larger class of "evidence-based" predictive instruments and condemns those instruments in gross. It argues that they elicit "overt discrimination based on demographics and socioeconomic status."[53] Its legal analysis is premised on the dubious proposition that "[c]urrent" constitutional law "calls into serious question the variables related to socioeconomic status, such as employment status, education, income, dependence on government assistance, and job skills."[54] I am not convinced this is an accurate statement of current law. My analysis thus proceeds on the basis of different doctrinal predicates. Moreover, the earlier article does not explicate carefully both the costs and benefits of algorithmic criminal justice.[55] A more meticulous approach is needed that disaggregates possible technological approaches and normative effects.

51. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 694 (2016) (examining "[l]iability under Title VII for discriminatory data mining [which] will depend on the particular mechanism by which the inequitable outcomes are generated"); *see also* Kroll et al., *supra* note 39, at 692–95 ("Algorithmic decisionmaking blurs the definitions of disparate treatment and disparate impact [under Title VII] and poses a number of open questions.").

52. In addition, there is a small body of insightful popular literature about the distributive effects of algorithmic instruments more generally. *See generally* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018) (discussing the impact of automated systems on poor people in America); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 203–06 (2016) (decrying the regressive tendencies of big-data technologies generally).

53. Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 806 (2014). One other article contains the assertion that "if racial and ethnic variables significantly improved the predictive validity of risk-needs models, then including them would appear to be narrowly tailored to the government's compelling interests." Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 259 (2015). Hamilton equates narrow tailoring with minimal efficacy. She fails to meaningfully grapple with existing precedent. And she is opaque as to what kind of racial effects might have legal or normative significance. Her analysis is thus quite limited. Finally, a brief 2016 article suggests that the application of certain algorithmic tools in a sentencing context might violate the Bill of Attainder Clause. Gregory Cui, *Evidence-Based Sentencing and the Taint of Dangerousness*, 125 YALE L.J. FORUM 315, 317 (2016).

54. Starr, *supra* note 53, at 830. Starr also argues that evidence-based methods do worse in sheer accuracy terms than readily available alternatives such as clinical assessments. *Id.* at 842–62. This is also orthogonal to my analysis here.

55. Starr notes that "[t]here appears to be a general consensus that using race would be unconstitutional," *id.* at 812, but this assertion is not based on a comprehensive appreciation of the ways in which racial effects might be embedded in, or emerge from, algorithmic instruments.

The Article unfolds in three steps. Part I defines algorithmic criminal justice and illustrates it by isolating discrete clusters of related instruments now employed in criminal justice or likely soon to be used. I also supply nontechnical exposition of the relevant technologies. Part II explores the legal criteria of racial equity with special attention to the Equal Protection Clause. It identifies deficiencies in that framework as it applies to algorithmic criminal justice. Part III then turns to the nascent computer science literature on technical standards of fairness for algorithmic criminal justice. I begin by articulating a normative account of racial equity concerns in criminal justice. I then work through the various metrics identified in the literature to measure racial equity, as well as the tensions between those metrics. Finally, I set forth my own account of racial equity and explain how it can be operationalized—both in theory and in practice.

I. ALGORITHMIC CRIMINAL JUSTICE: SCOPE AND OPERATION

Predictive criminal justice was old when Captain Renault told his men in *Casablanca* to "round up the usual suspects."[56] The meaningful use of "criminal justice determinations that do not rest simply on probabilities but on statistical correlations between group traits and group criminal offending rates" can be traced back to the beginning of the twentieth century.[57] The resulting profusion of predictive instruments extends well beyond the algorithmic criminal justice instruments to be considered here. For example, an array of evidence-based interventions from interviews to actuarial scoring have long been employed in the sentencing context.[58]

56. CASABLANCA (Warner Bros. 1942); *see also* Jurek v. Texas, 428 U.S. 262, 275 (1976) ("[P]rediction of future criminal conduct is an essential element in many of the decisions rendered throughout our criminal justice system.").

57. BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE 18 (2007). Prediction has become an entrenched part of criminal justice:

> Criminal justice actors often predict which defendants are going to commit an additional crime in determining whether to arrest defendants, to release them on bail, or to release them on parole, or in determining their sentence. This prediction is often based not only on individual evaluation, but also on a group's criminality and past behavior.

Shima Baradaran, *Race, Prediction, and Discretion*, 81 GEO. WASH. L. REV. 157, 176–77 (2013). *See also* Richard Berk, *Forecasting Methods in Crime and Justice*, 4 ANN. REV. L. & SOC. SCI. 219, 221–23 (2008) (setting out the history of formal crime prediction models).

58. Cecelia Klingele, *The Promises and Perils of Evidence-Based Corrections*, 91 NOTRE DAME L. REV. 537, 539 (2015) (discussing "the use of actuarial risk and need assessment instruments, motivational interviewing and counseling techniques, deterrence-based sanction

To sharpen the ensuing analysis, it is useful to define with some precision a discrete domain of practices as "algorithmic criminal justice." This Part offers such a definition and then fleshes out that concept with a series of examples from the policing, bail, and post-conviction (parole and probation) contexts. Where salient, I offer capsule accounts of relevant technologies central to my analysis.

A. A Definition of Algorithmic Criminal Justice

Algorithmic criminal justice, as I define the term, is the application of an automated protocol to a large volume of data to classify new subjects in terms of the probability of expected criminal activity and in relation to the application of state coercion. This definition has three elements. Once explicated, those elements provide a justification for treating this domain as a distinct object of legal and normative inquiry.

First, my definition requires an *automated* protocol, or algorithm, that routinizes a decision—here, about state coercion.[59] In contrast to such a structured decision-making context, American criminal justice is replete with instances in which officials such as police officers, sentencing judges, parole boards, or probation officers exercise partially structured discretion to determine the legality of coercing a particular person. Even where a written protocol is used, as in the sentencing context, substantial residual discretion remains.[60] In a larger domain of cases, though, criminal justice actors are unbounded by either protocol or clear rules. For example, the Fourth Amendment imposes thresholds of reasonable "articulable suspicion" for certain street stops,[61] and "probable cause" for certain arrests.[62] The Supreme Court has resisted efforts to formalize these concepts into "technical"[63]

programs, and incentives to probationers and parolees for successful compliance with court orders," with attention to their effects on aggregate incarceration levels).

59. THOMAS H. CORMEN, CHARLES E. LEISERSON, RONALD L. RIVEST & CLIFFORD STEIN, INTRODUCTION TO ALGORITHMS 5 (3d ed. 2009) (defining an algorithm as "any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as outcome" (emphases omitted)); *see also* MARTIN ERWIG, ONCE UPON AN ALGORITHM: HOW STORIES EXPLAIN COMPUTING 26–27 (2017) (offering an illuminating conceptual account of algorithms); Reuben Binns, *Algorithmic Accountability and Public Reason*, PHIL. & TECH. 1, 3 (2017) (describing algorithms in terms of whether a system will "take in certain inputs and produce certain outputs by computational means").

60. For an analysis of the scope of discretion in the federal context at present, see Kevin R. Reitz, *"Risk Discretion" at Sentencing*, 30 FED. SENT'G. REP. 68 (2017).

61. Terry v. Ohio, 392 U.S. 1, 22, 31 (1968).

62. Brinegar v. United States, 338 U.S. 160, 174 (1949).

63. *Id.* at 175.

rules and instead has preferred "practical, common-sense judgment."[64] Algorithmic criminal justice represents a categorical rejection of such ad hoc, situated judgments as an instrument of regulation.

Second, automation is required because of the sheer *volume* of data used by these tools. Law enforcement agencies increasingly have access to pools of data that are "vast, fast, disparate, and digital."[65] Colloquially, the instruments at issue here rely on "big data" as that term is used in computational science.[66] The Los Angeles Police Department, for example, has supplemented traditional law enforcement databases of persons arrested or convicted of crimes with information about all contacts, of any sort, with police, social services, health services, and child welfare services.[67] This data is integrated with data from "dragnet surveillance tools," closed-circuit television ("CCTV") cameras used to acquire and track license plate numbers, and "privately collected data."[68] Because the ensuing massive data pools cannot be sorted by hand, they are only useful because of advances in processing power and computational software. The IC Realtime Company, for instance, offers an application called "Ella," which can recognize and execute natural language queries for CCTV footage.[69] Such changes in the speed and accuracy of queries effect a step change in the quality of surveillance-based evidence available to police.

Third, these algorithmic instruments make *out-of-sample predictions* about new actors' likely criminal conduct. It is true that algorithmic instruments can also be applied to extant pools of big data in order to identify historical crimes. For example, the Securities and Exchange Commission analyzes large volumes of trading to identify

64. Illinois v. Gates, 462 U.S. 213, 244 (1983). The Court has stressed police expertise rather than formal rules. United States v. Cortez, 449 U.S. 411, 418 (1981) ("[A] trained officer draws inferences and makes deductions . . . that might well elude an untrained person.").

65. Sarah Brayne, *Big Data Surveillance: The Case of Policing*, 82 AM. SOC. REV. 977, 980 (2017).

66. DAWN E. HOLMES, BIG DATA: A VERY SHORT INTRODUCTION 15–16 (2017) (characterizing big data as "huge amounts of data that has not been collected with any specific questions in mind and is often unstructured" and that is characterized by "volume, variety, and velocity").

67. Brayne, *supra* note 65, at 995.

68. *Id.* at 992–95.

69. James Vincent, *Artificial Intelligence is Going to Supercharge Surveillance*, VERGE (Jan. 23, 2018, 10:54 AM), https://www.theverge.com/2018/1/23/16907238/artificial-intelligence-surveillance-cameras-security [https://perma.cc/U4SN-VFL2].

investors who might be engaged in insider trading.[70] Pattern analysis of this kind can raise questions of racial effects. But it does so in different ways from out-of-sample prediction methods. The instruments I'm focused on here are generally calibrated using one pool of data and then applied to new data as a means of identifying or predicting crime that was previously unknown and that, typically, has not yet occurred. For example, a parole board might have information on historical patterns of reoffending. It supplies that data to a machine-learning tool, which in turn generates a test for forecasting recidivism by suspects yet to interact with the criminal justice system.[71]

So defined, algorithmic criminal justice tools are inductive rather than deductive. They lack opportunities for verification via the collation of other indicia of lawbreaking. Algorithmic criminal justice, moreover, claims no insight into the *causes* of crime or criminality.[72] It is just an arrow pointing at crime's likely next incidence.

B. The Operation of Algorithmic Criminal Justice

I have already discussed two instances of algorithmic criminal justice, the COMPAS algorithm and the SSL. These examples, though, do not provide a good measure of the scope and effects of algorithmic criminal justice's operation. New technologies of machine learning (and in particular the subspecies of deep learning) are likely to dominate algorithmic criminal justice in the future. As a result, both COMPAS and the SSL algorithms are likely soon to be relics. Newer tools will combine powerful computational instruments with large volumes of data to enable prediction of a kind that is qualitatively distinct from historical antecedents.[73] A survey of the potential uses of

---

70. Mary Jo White, Chair, Sec. and Exch. Comm'n, Keynote Address at the 41st Annual Securities Regulation Institute (Jan. 27, 2014), http://www.sec.gov/News/Speech/Detail/Speech/1370540677500 [https://perma.cc/M7YV-33PR] (describing the SEC's NEAT program, which can identify and analyze insider trading activity around times of major corporate events).

71. Richard Berk, *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recividism*, 13 J. EXPERIMENTAL CRIMINOLOGY 193, 195 (2017).

72. *Cf.* Usama Fayyad, *The Digital Physics of Data Mining*, 44 COMM. ACM 62, 62, 64 (2001) ("Even with all these techniques [for data mining algorithms], we have taken only the first unsteady steps toward addressing such difficult problems as understanding and exploiting the meaning of information hidden from our perception in the higher dimension.").

73. *See* JERRY KAPLAN, ARTIFICIAL INTELLIGENCE: WHAT EVERYONE NEEDS TO KNOW 39 (2016) (pointing to "improvements in computing speed and memory, the transition from physically to electronically stored data, easier access (mainly due to the Internet), and low-cost high-resolution digital sensors" as the technological predicates of machine learning).

these new instruments allows a fuller sense of the scope of algorithmic prediction tools and the effects that they will have on criminal justice.

1. *Machine Learning and Deep Learning.* A machine-learning algorithm solves a "learning problem . . . of improving some measure of performance when executing some task, through some type of training experience."[74] The basic task a supervised machine-learning algorithm must perform can be framed as follows: The algorithm is prompted to define a function f(x) which produces an output y for any given input x. In other words, it classifies x in terms of y.[75] Its outputs take the form of a sorting of x onto categories of y.[76] The resulting classifications are correlational rather than causal in nature.[77] Hence, its performance is measured in terms of how well it captures the relation of x to y.[78]

To begin, a *supervised* machine-learning algorithm is assigned a set of "training" data labeled in terms of y so it can develop a model, represented by the mathematical function f(x), that best represents the relationship between features of each observation in the training data and the known classification y. This function f(x) is then applied to a new "test set" of data.[79] The algorithm predicts how to classify this new data by applying f(x) to generate predictions of y.[80] Such supervised tools are but one kind of machine learning. There is also a species of

74. Jordan & Mitchell, *supra* note 34, at 255.

75. *Id.* This process can also be described in terms of a "classifier," rather than a function, that examines inputs with "feature values" and outputs a class variable. Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, 55 COMM. ACM 78, 79, 82 (2012) (emphases omitted) ("A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class.").

76. PETER FLACH, MACHINE LEARNING: THE ART AND SCIENCE OF ALGORITHMS THAT MAKE SENSE OF DATA 14 (2012) (noting that "multi-class classification" is "a machine learning task in its own right").

77. Consider in this regard recommendation algorithms employed by consumer-facing companies such as Amazon and Netflix. *Cf.* KAPLAN, *supra* note 73, at 32 (arguing that machine-learning algorithms operate like "incredibly skilled mimics, finding correlations and responding to novel inputs as if to say, 'This reminds me of . . . ,' and in doing so imitate successful strategies gleaned from a large collection of examples").

78. *See* Jordan & Mitchell, *supra* note 34, at 255–57 (noting that performance can be defined in terms of accuracy, with false positive and false negative rates being assigned a variety of weights).

79. *See* ALPAYDIN, *supra* note 32, at 40 (describing the use of training and validation data); HOLMES, *supra* note 66, at 24 (discussing classification and distinguishing training and test sets of data).

80. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 693–97 (3d ed. 2010).

unsupervised machine-learning algorithms. These begin with unlabeled training data and tend to be tasked with the development of classifications based on the data's immanent structure.[81]

No machine-learning algorithm is given ex ante a functional form f(x) that defines the relationship between observations and classifications. Rather, the algorithm employs one of a wide number of procedures to ascertain f(x) through a process called "feature selection."[82] The latter includes decision trees, decision forests, logistic regression, support vector machines, neural networks, kernel machines, and Bayesian classifiers.[83] Each of these tools identifies a mathematical criterion for selection, f(x), by testing many potential criteria using training data. By sorting though many different possible f(x)s on the basis of its training data using one of these methods, the algorithm homes in upon an f(x) that optimizes the accuracy of its performance metric. Many people encounter this kind of machine learning in interactions with Siri, Alexa, or other virtual assistants.[84]

Deep learning is a subset of machine learning whereby the algorithm is made up of "multiple levels of representation," each of which transforms the raw data into a slightly more abstract form.[85] Given enough layers of transformation, the algorithm can perform very complex functions. It can, for instance, play the Chinese game Go or recognize specific images from representational input.[86] What

81. *See* FLACH, *supra* note 76, at 14–17.

82. *See* Kroll et al., *supra* note 39, at 681 (describing feature selection as concerning the "choices about which data models should consider"); *see also* Avrim L. Blum & Pat Langley, *Selection of Relevant Features and Examples in Machine Learning*, 97 ARTIFICIAL INTELLIGENCE 245, 250–53 (1997) (decomposing feature selection into a nested sequence of analytic tasks); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 700–01 (2017) (describing feature selection).

83. *See* David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 1, 3 (2006) (documenting these instruments and contending that, in "real-world conditions," simpler instruments often perform better).

84. *See 9 Applications of Machine Learning From Day-to-Day Life*, MEDIUM (July 30, 2017), https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0/ [https://perma.cc/GNE6-X6W5].

85. LeCun et al., *supra* note 35, at 436, 438 ("A deep-learning architecture is a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input-output mappings."); *see generally* ALPAYDIN, *supra* note 32, at 85–109 (describing neural networks). For a nontechnical account of back propagation, the key element of deep learning, see James Somers, *Is AI Riding a One-Trick Pony?*, 120 MIT TECH. REV. 29, 31 (2017).

86. *See, e.g.*, Tom Simonite, *This More Powerful Version of AlphaGo Learns on Its Own*, WIRED (Oct. 18, 2017, 1:00 PM), https://www.wired.com/story/this-more-powerful-version-of-alphago-learns-on-its-own [https://perma.cc/L38N-D94H] (describing a program that not only

distinguishes deep learning is that its "layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure."[87] The most well-known forms of deep-learning tools are based on "neural networks," which are very loosely inspired by patterns observed in the human brain.[88] Deep-learning instruments are especially apt for unsupervised tasks, with no specification of features and little "manual interference," such that designers "just wait and let the learning algorithm discover all that is necessary by itself."[89] The utility to police of an instrument that can extract speech or visual patterns from large quantities of audio-visual inputs (e.g., CCTV footage, cellphone call content) is self-evident.[90]

2. *The Impact of Machine Learning on Criminal Justice.* Adoption of machine learning within the criminal justice system changes the scale, reach, and operation of state power. Consider each of these parameters in turn.

First, these tools dramatically inflate the state's ability to acquire otherwise inaccessible information.[91] For instance, police in London and in South Wales now track individuals' locations and movements over days and weeks by applying machine-learning tools to thousands of hours of CCTV footage.[92] Machine-learning tools also facilitate predictions that would be far less precise if based solely upon more familiar regression analyses.[93]

---

wins board games but "showcases an approach to teaching machines new tricks that makes them less reliant on humans").

87. LeCun et al., *supra* note 35, at 436.

88. *See* Jürgen Schmidhuber, *Deep Learning in Neural Networks: An Overview*, 61 NEURAL NETWORKS 85, 86–87 (2015).

89. ALPAYDIN, *supra* note 32, at 309.

90. *See* Maryam M. Najafabadi et al., *Deep Learning Applications and Challenges in Big Data Analytics*, 1 J. BIG DATA 1, 11–13 (2015) (describing uses of deep-learning tools). Deep learning has also been used to play "games of perfect information," such as chess and Go. *See, e.g.*, David Silver et al., *Mastering the Game of Go with Deep Neural Networks and Tree Search*, 529 NATURE 484, 490 (2016).

91. For recognition of this general point, see *United States v. Garcia*, 474 F.3d 994, 998 (7th Cir. 2007) ("Technological progress poses a threat to privacy by enabling an extent of surveillance that in earlier times would have been prohibitively expensive," thereby "giving the police access to surveillance techniques that are ever cheaper and ever more effective").

92. David Bond, *CCTV Watchdog Warns UK Police over Use of Facial Recognition*, FIN. TIMES (Oct. 29, 2017), https://www.ft.com/content/ab60f9f2-bb26-11e7-8c12-5661783e5589 [https://perma.cc/3HC2-9TBF].

93. *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ziad Obermeyer, *Prediction Policy Problems*, 105 AM. ECON. REV. 491, 493–94 (2015) (comparing a machine-learning prediction to a regression analysis in forecasting whether elderly osteoarthritis patients will live

Second, machine-learning instruments sever the connection between the human operator and the function f(x) used to solve the classification problem. Unstructured human discretion, which once infused the criminal justice system, is displaced by an algorithmically structured logic not wrought by any human hand. As a result, it will often not be possible to speak of the intent or the anticipated consequences of a classification protocol. Rather, the algorithm will "sift through vast numbers of variables, looking for combinations that reliably predict outcomes," "handling enormous numbers of predictors—sometimes, remarkably, more predictors than observations—and combining them in nonlinear and highly interactive ways,"[94] hence generating utterly unexpected outcomes.

Moreover, to the extent that the design of a machine-learning process involves the intentional crafting and selection of training data, feature sets, or the like, there will often be no way to directly ascertain the role of designers' racial sentiments (if any)[95] and no easy way to indirectly infer intentionality from the instrument's results.[96] There is no such thing as code that bespeaks racial animus. Design choices that might be molded by racial animus also cannot be reverse engineered to cast light on background human motivations. And it is difficult to know how to disentangle the effect of background differences in criminality and discriminating designer intent when evaluating the outputs of an algorithm. As a result, the effects of, and evidence for, human intentions—a central element of legal and constitutional analysis—are likely to be elusive.

Third, algorithmic tools can be as sticky or stickier than the forms of human discretion. Hence, whereas it is always a possibility that human agents will observe the unintended effects of human action, machine decision-making can be opaque and hence resistant to change. Algorithmic systems can thus be "stuck in time until engineers dive in

long enough to benefit from joint replacement surgeries).

94. Ziad Obermeyer & Ezekiel J. Emanuel, *Predicting the Future—Big Data, Machine Learning, and Clinical Medicine*, 13 NEW ENG. J. MED. 1216, 1217 (2016) (citation omitted).

95. *See* Barocas & Selbst, *supra* note 51, at 710 ("The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions.").

96. *See* Tal Z. Zarsky, *Transparent Predictions*, 2013 U. ILL. L. REV. 1503, 1519–20 (2013) (noting how predictions can be generated in processes "which [are] not explainable in human language," such that "[i]t would be difficult for the government to provide a detailed response when asked why an individual was singled out to receive differentiated treatment by an automated recommendation system").

to change them."[97] Indeed, it will often not be clear to a human operator that an algorithmic criminal justice tool needs reconsideration. That human operator necessarily sees only a limited and unrepresentative tranche of case outcomes. She must also grapple with the sheer technological complexity of algorithmic tools. Hence, algorithmic errors are often liable to prove more durable than human errors.

Fourth, the consequences of switching between unstructured human discretion and algorithmically structured prediction can often be unexpected. This happens even when a semistructured instrument is altered. For example, in the wake of the Supreme Court's decision eliminating the mandatory character of the Federal Sentencing Guidelines, studies found "significantly increased racial disparities after controlling for extensive offender and crime characteristics."[98] This was not, one hopes, the Court's ambition.

Fifth, the emerging crop of algorithmic tools are potentially very different from risk assessment tools currently employed in bail and sentencing. Current instruments rely on a relatively small number of variables—two leading models use 12 and 20 parameters, respectively—and fixed classification rules to generate recidivism risks.[99] These instruments focus solely on recidivism risk. They make no effort to estimate either the direct or the remote costs of coercive action. In contrast, tools such as COMPAS include recommended cutoff points that at least imply an evaluation of aggregate social costs. There is no reason, moreover, that an algorithm could not be trained with data that reflects both the costs and the benefits of coercive action, broadly understood. This does not appear to be standard practice yet.

To summarize, the operation and the effects of predictions offered by algorithmic criminal justice are qualitatively distinct from the unstructured and semistructured forms of human discretion that have until now dominated the criminal justice system. Not all such tools use machine learning or deep learning. But it is only a question of time before these powerful instruments crowd out simpler models. Indeed,

97. O'NEIL, *supra* note 52, at 204.

98. Crystal S. Yang, *Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing*, 44 J. LEGAL STUD. 75, 77 (2015); *see also* Max M. Schanzenbach & Emerson H. Tiller, *Strategic Judging Under the U.S. Sentencing Guidelines: Positive Political Theory and Evidence*, 23 J.L. ECON. & ORG. 24, 54 (2007) (presenting a similar finding).

99. Slobogin, *supra* note 47, at 584–86 (explaining the OxRec and VRAG assessment tools); *see also* Lauryn P. Gouldin, *Disentangling Flight Risk from Dangerousness*, 2016 BYU L. REV. 837, 869–71 (describing the PSA tool).

it is striking that both the COMPAS algorithm and the SSL instrument described in the Introduction have been criticized on the basis of their weak predictive power[100]: A likely, if not inevitable, consequence of such critiques is the adoption of new, more powerful computational tools to achieve the same end. In any event, a phase shift in the quality of criminal justice action can already be observed across the spectrum of criminal justice functionalities. Even if machine-learning and deep-learning tools are not now omnipresent, they will be soon.[101]

C. Algorithmic Criminal Justice on the Ground

Algorithmic tools are used now in three main criminal justice contexts: policing, bail decisions, and post-conviction matters. This section provides a capsule summary of the ways in which predictive instruments are operationalized across those three distinct domains.

1. *Policing.* In the policing context, algorithmic tools are employed to make predictions about both places and people.[102] Place-focused tools aggregate "real-time" information on the frequency and geographic location of crimes to "determine staffing needs or allocate resources" as between different regions.[103] Consonant with a focus on the location of crime, police departments across the country have

100. *See* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, SCIENCE ADVANCES (Jan. 17, 2018), http://advances.sciencemag.org/content/4/1/eaao5580 [https://perma.cc/9WC8-CDES] (finding that the COMPAS algorithm performs no better than people with no experience with the criminal justice system in making recidivism predictions); Jessica Saunders, Priscillia Hunt & John S. Hollywood, *Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot*, 12 J. EXPERIMENTAL CRIMINOLOGY 347, 363 (2016) (finding that "while using arrestee social networks improved the identification of future homicide victims, the number was still too low in the pilot to make a meaningful impact on crime").

101. One reason for this, of course, is the promotion of algorithmic implements by the companies that manufacture them and stand to gain financially from their adoption. *See* Elizabeth E. Joh, *The Undue Influence of Surveillance Technology Companies on Policing*, 92 N.Y.U. L. REV. 101, 114–20 (2017) (describing mechanisms of private influence on public adoption of computational technologies in the criminal justice sector).

102. *See* RAND CORP., PREDICTIVE POLICING: FORECASTING CRIME FOR LAW ENFORCEMENT 2 (2013), https://www.rand.org/content/dam/rand/pubs/research_briefs/RB9700/RB9735/RAND_RB9735.pdf [https://perma.cc/U8GG-MSJ7]; *see also* JENNIFER BACHNER, IBM CTR. FOR THE BUS. OF GOV'T, PREDICTIVE POLICING: PREVENTING CRIME WITH DATA AND ANALYTICS 14 (2013), http://www.businessofgovernment.org/sites/default/files/Predictive%20Policing.pdf [https://perma.cc/9CS8-5X37] ("The fundamental notion underlying the theory and practice of predictive policing is that we can make probabilistic inferences about future criminal activity based on existing data.").

103. Andrew Guthrie Ferguson, *Crime Mapping and the Fourth Amendment: Redrawing "High-Crime Areas*," 63 HASTINGS L.J. 179, 182 (2011).

increasingly adopted the Compstat, or Crime Control Strategy Meeting, structure first developed in New York. Under Compstat, precinct commanders are subject to biweekly questioning by senior departmental leadership in a "data-saturated environment" about how they are responding to crime trends.[104] While Compstat itself does not necessarily incorporate algorithmic tools, its focus on data-driven predictions of crime's geographic dispersion invites the use of algorithmic tools. Further, a number of criminologists have identified promise in a place-based prediction approach involving "the application of police interventions at very small geographic units of analysis," or hot spots.[105] A number of randomized, controlled experiments have found evidence that such place-focused tools are effective in suppressing crime.[106]

Consistent with these developments, influential jurisdictions have adopted machine-learning tools to facilitate place-based policing.[107] One of the earliest adopters, starting in 2015, was the New York Police Department. This force embarked on a two-year pilot program using HunchLab, an algorithm developed by the Philadelphia-based Azavea company.[108] According to Azavea's web site, HunchLab's "ensemble machine learning" algorithm uses "temporal cycles" (day of week, seasonality); "weather"; "risk terrain modeling" (locations of bars, bus stops, etc.); "socioeconomic indicators"; historic crime levels; and near-

104. James J. Willis, Stephen D. Mastrofski & David Weisburd, *Making Sense of COMPSTAT: A Theory-Based Analysis of Organizational Change in Three Police Departments*, 41 LAW & SOC'Y REV. 147, 148 (2007); *see also* David L. Carter & Jeremy G. Carter, *Intelligence-Led Policing: Conceptual and Functional Considerations for Public Policy*, 20 CRIM. JUST. POL'Y REV. 310, 320–22 (2009) (comparing intelligence-led policing with CompStat).

105. ANTHONY A. BRAGA & DAVID L. WEISBURD, POLICING PROBLEM PLACES: CRIME HOT SPOTS AND EFFECTIVE PREVENTION 9 (2010). More generally, proactive policing of various kinds—not necessarily involving stops—is also associated with crime-control effects. *See* Charis E. Kubrin, Steven F. Messner, Glenn Deane, Kelly McGeever & Thomas D. Stucky, *Proactive Policing and Robbery Rates Across U.S. Cities*, 48 CRIMINOLOGY 57, 62–63 (2010).

106. *See, e.g.*, Anthony A. Braga & Brenda J. Bond, *Policing Crime and Disorder Hot Spots: A Randomized Controlled Trial*, 46 CRIMINOLOGY 577 (2008) (Lowell, Massachusetts); Anthony A. Braga, David L. Weisburd, Elin J. Waring & Lorraine Green Mazerolle, *Problem-Oriented Policing in Violent Crime Places: A Randomized Control Experiment*, 37 CRIMINOLOGY 541 (1999) (Jersey City, New Jersey).

107. *See, e.g.*, J. Brian Charles, *How Police in One City Are Using Tech to Fight Gangs*, GOVERNING (Apr. 11, 2018), http://www.governing.com/topics/public-justice-safety/gov-gang-violence-predictive-policing-high-point-lc.html?r [https://perma.cc/2TVR-8UJC] (describing the use of ONESolution predictive software by the High Point, North Carolina, police department).

108. *See* Laura Nahmias & Miranda Neubauer, *NYPD Testing Crime-Forecast Software*, POLITICO (July 8, 2015, 5:52 AM), https://www.politico.com/states/new-york/city-hall/story/2015/07/nypd-testing-crime-forecast-software-090820 [https://perma.cc/3G49-UP9B].

repeat patterns as a means of predicting individual crime expectations across the jurisdiction.[109] Other cities, such as Los Angeles, have adopted a system created by the PredPol company. PredPol produces a propriety algorithm based on a "near-repeat" machine-learning model. This assumes that if a crime occurs at a given location, the immediate surroundings are at increased risk for future crime.[110] First developed by anthropologist Jeffrey Brantingham and mathematician Andrea Bertozzi, the PredPol model is an extrapolation of an algorithm used to predict the distribution of earthquake shocks.[111] One randomized, controlled study observed the use of a machine-learning tool derived from models of epidemic aftershocks to implement hot-spot policing; it found that the instrument predicted crime well and led to a 7.4 percent reduction in crime volume as a function of patrol time.[112]

In the last five years, however, authorities have begun to supplement place-focused tools with person-focused tools. Chicago, for example, started to build a database of alleged gang members in order to draw inferences about their propensity to commit violent crimes.[113] That city's SSL predicts the likelihood of an individual becoming a homicide victim using an analysis of that person's known social network—in particular, by counting the number of first-degree co-arrest links and the number of second-degree co-arrest links with previous homicide victims.[114] Names generated by the SSL algorithm were disseminated to district commanders, who had discretion about what interventions to apply.[115] The algorithm, however, identified less

109. AZAVEA, HUNCHLAB: UNDER THE HOOD 12 (2015), https://cdn.azavea.com/pdfs/hunchlab/HunchLab-Under-the-Hood.pdf [https://perma.cc/45YT-6F69].

110. *See* Brayne, *supra* note 65, at 989–90; *see generally How PredPol Works: Predictive Policing*, PREDPOL, http://www.predpol.com/how-predictive-policing-works [https://perma.cc/8J3A-GHLN] (providing a predictably rosy overview of the algorithm's uses).

111. Aaron Shapiro, *Reform Predictive Policing*, 541 NATURE 458, 459 (2017); *see also* Elizabeth E. Joh, *Policing by Numbers: Big Data and the Fourth Amendment*, 89 WASH. L. REV. 35, 44–45 (2014) (describing PredPol's use in Santa Cruz, California).

112. G. O. Mohler et al., *Randomized Controlled Field Trials of Predictive Policing*, 110 J. AM. STAT. ASS'N 1399, 1407 (2015).

113. John Buntin, *Social Media Transforms the Way Chicago Fights Gang Violenc*e, GOVERNING (Oct. 2013), http://www.governing.com/topics/public-justice-safety/gov-social-media-transforms-chicago-policing.html [https://perma.cc/ALA4-FSGL] (describing how data was acquired for social media analysis).

114. Saunders et al., *supra* note 100, at 354.

115. *Id.* at 354–55; *cf.* Mark Guarino, *Can Math Stop Murder?*, CHRISTIAN SCI. MONITOR (July 20, 2014), https://www.csmonitor.com/USA/2014/0720/Can-math-stop-murder [https://perma.cc/25P7-SCFH] (discussing predictive policing techniques in Chicago, including

than one percent of the pool of eventual homicide victims and yielded no identifiable crime-control gains.[116] In a similar vein, some jurisdictions use machine-learning tools to mine social services records for predictions of child abuse.[117]

A related use of deep-learning tools involves facial recognition algorithms that can search for dangerous persons in a specific place at a particular time. This emerging use is not a matter of out-of-sample prediction; it is a matching exercise based on new data. As such, it falls at the periphery of my analysis. For instance, the Metropolitan Police of London combine dense CCTV with facial recognition instruments in monitoring certain public events, although not (yet) for purposes such as terrorism and serious crime prevention.[118] In May 2017, the deployment of facial recognition algorithms to real-time CCTV inputs generated the first arrest of its kind for British police.[119]

The situation in the United States is less clear. As of 2016, at least five metropolitan police departments—including Chicago's, Dallas's, and Los Angeles's—claimed to use, or expressed interest in buying, a facial recognition algorithm to comb public CCTV data.[120] Facial images have been made available by the Federal Bureau of Investigation since 2011.[121] In 2017, Orlando, Florida, and Washington County, Oregon, were identified as purchasers of Amazon's "Rekognition" tool, which uses "artificial intelligence" to scan and identify up to a hundred faces in a single CCTV shot.[122] Nevertheless,

sending officers to the houses of suspected gang leaders).

116. Saunders et al., *supra* note 100, at 363; *see id.* at 365 (noting that those included on the SSL list were in fact less likely to be a victim of a shooting, although that difference was not statistically significant).

117. Rachel Courtland, *The Bias Detectives: The Researchers Striving to Make Science Fair*, 558 NATURE 357, 358 (2018).

118. *See* Mark Townsend, *Police To Use Facial-Recognition Cameras at Cenotaph Service*, GUARDIAN (Nov. 11, 2017, 7:05 PM), https://www.theguardian.com/technology/2017/nov/12/metropolitan-police-to-use-facial-recognition-technology-remembrance-sunday-cenotaph [https://perma.cc/PF2B-KZHS].

119. Cara McGoogan, *British Police Arrest Suspect Spotted with Facial Recognition Technology*, TELEGRAPH (June 7, 2017, 4:31 PM), http://www.telegraph.co.uk/technology/2017/06/07/british-police-arrest-suspect-spotted-facial-recognition-technology [https://perma.cc/Q5H7-W4QF].

120. CLARE GARVIE, ALVARO BEDOYA & JONATHAN FRANKLE, CTR. ON PRIVACY & TECH. AT GEORGETOWN LAW, THE PERPETUAL LINE-UP: UNREGULATED POLICE FACE RECOGNITION IN AMERICA 2 (2016), https://assets.documentcloud.org/documents/3145908/The-Perpetual-Line-Up-By-The-Center-on-Privacy.pdf [https://perma.cc/YN64-FAMU].

121. U.S. GOV'T ACCOUNTABILITY OFFICE, GAO-16-267, FACE RECOGNITION TECHNOLOGY: FBI SHOULD BETTER ENSURE PRIVACY AND ACCURACY 7, 15 (2016).

122. Matt Cagle & Nicole Ozer, *Amazon Teams Up with Government To Deploy Dangerous*

real-time application of facial recognition technologies to CCTV data still appears rare, in particular because of technological barriers. It is telling that between June and September 2017, the National Institute for Science and Technology offered a prize for facial recognition technology. The winner of the contest, NTechLab, created an algorithm with a rate of 0.22 false nonmatches for every 0.001 false matches.[123] And in 2018, as noted above, the IC Realtime company introduced a commercially available algorithm called Ella that can recognize and respond to natural language queries to search large quantities of video footage for specific images.[124]

2. *Bail.* The second use of algorithmic tools is in the pretrial context of arraignment hearings, in which judges determine whether defendants are to be detained pending criminal trial or released having posted a money bail or otherwise. Pretrial detainees comprise roughly 60 percent of the jail population, and between 2005 and 2013 some 450,000 people were incarcerated awaiting trial on any given day.[125] Pretrial detention decisions impose considerable costs on individuals in relation to employment, health outcomes, and childcare costs.[126] One study, for example, estimates a lower-bound net cost of detention for the marginal individual of $55,385 and an upper-bound net cost of $101,223.[127] At the same time, "[r]elatively little is known with regard to charge characteristics and case dispositions" for that pretrial

---

New Facial Recognition Technology, ACLU (May 22, 2018, 10:00 AM), https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazon-teams-government-deploy-dangerous-new?redirect=blog/privacy-technology/surveillance-technologies/amazon-teams-law-enforcement-deploy-dangerous-new [https://perma.cc/D5LY-MK68].

123. PATRICK GROTHER, MEI NGAN, KAYEE HANAOKA, CHRIS BOEHNEN & LARS ERICSON, NAT'L INST. STANDARDS & TECH., NISTIR 8197, THE 2017 IARPA FACE RECOGNITION PRIZE CHALLENGE (FRPC) 2 (2017).

124. Vincent, supra note 69.

125. Jaeok Kim, Preeti Chauhan, Olive Lu, Meredith Patten & Sandra Susan Smith, *Unpacking Pretrial Detention: An Examination of Patterns and Predictors of Readmissions*, 29 CRIM. JUST. POL'Y REV. 663, 664 (2018); *see also* ROY WALMSLEY, INT'L CTR. FOR PRISON STUDIES, WORLD PRE-TRIAL/REMAND IMPRISONMENT LIST 1 (2d ed. 2014), http://www.prisonstudies.org/sites/default/files/resources/downloads/world_pre-trial_imprisonment_list_2nd_edition_1.pdf [https://perma.cc/K9MX-4NAD] (describing the "number of prisoners held in pre-trial detention and other forms of remand imprisonment in 211 independent countries and dependent territories").

126. For a discussion of the costs of pretrial detention, see Sandra G. Mayson, *Dangerous Defendants*, 127 YALE L.J. 490, 546–57 (2018).

127. Crystal S. Yang, *Toward an Optimal Bail System*, 92 N.Y.U. L. REV. 1399, 1436 (2017).

detention population.[128] But studies in a range of jurisdictions find evidence of racial disparities in bail decisions. Predictably, if dismayingly, black and Latino defendants receive systematically less favorable treatment.[129]

Much of the impetus for recent bail reform has hinged on the oft-criticized effect of wealth upon access to pretrial release.[130] Algorithmic criminal justice does not necessarily respond to this problem, except to the extent it enables a reduction of pretrial detention generally without imposing any cost on crime-related outcomes.[131] Rather, such tools are an obvious fit in a context where magistrates are forced to make predictive decisions about the risk of violence, criminality, or flight on the basis of relatively cursory information. Already, two simple algorithms, the Public Safety Assessment ("PSA") and the Canadian Level of Service Inventory Revised ("LSI-R"), use information ranging from criminal history to personality patterns and age to offer recidivism predictions.[132] The latter instrument, however, is administered by professionals through interviews—it involves no computational element.[133] More sophisticated algorithmic instruments are now starting to be introduced into courtrooms to inform bond determinations in jurisdictions across the country.[134]

128. Kim et al., *supra* note 125, at 667.

129. *See* Stephen Demuth & Darrell Steffensmeier, *The Impact of Gender and Race-Ethnicity in the Pretrial Release Process*, 51 SOC. PROBS. 222, 234 (2004) ("Black and especially Hispanic defendants receive less favorable pretrial treatment than white defendants."); Traci Schlesinger, *Racial and Ethnic Disparity in Pretrial Criminal Processing*, 22 JUST. Q. 170, 175–76, 187 (2005) (presenting the same result for hold rates within a given county using the nationally representative State Court Processing Statistics); Yang, *supra* note 127, at 1466–67 (finding "compelling evidence that bail judges in these jurisdictions treat defendants of different races differently in setting bail"). *But see* Frank McIntyre & Shima Baradaran, *Race, Prediction, and Pretrial Detention*, 10 J. EMPIRICAL LEGAL STUD. 741, 769 (2013) (rejecting any finding of racial disparities after having controlled for the probability of re-arrest).

130. *See* Nick Pinto, *The Bail Trap*, N.Y. TIMES MAG. (Aug. 13, 2015), http://www.nytimes.com/2015/08/16/magazine/the-bail-trap.html?_r=1 [https://perma.cc/AST3-9UBB] ("[In New York City,] only 15 percent of defendants are able to come up with the money to avoid jail.").

131. *See* Aviva Shen & CityLab, *New Orleans' Great Bail-Reform Experiment*, ATLANTIC (Oct. 19, 2017), https://www.theatlantic.com/politics/archive/2017/10/new-orleans-great-bail-reform-experiment/544964 [https://perma.cc/7UG9-3LHK] (finding that both the pretrial detention and the pretrial crime rate had fallen using the tool).

132. *Algorithms in the Criminal Justice System*, ELECTRONIC PRIVACY INFO. CTR., https://epic.org/algorithmic-transparency/crim-justice [https://perma.cc/4E4A-GXDL].

133. Alexander M. Holsinger, *Implementation of Actuarial Risk/Need Assessment and Its Effect on Community Supervision Revocations*, 15 JUST. RES. & POL'Y 95, 98–99 (2013).

134. Ellora Thadaney Israni, *When an Algorithm Helps Send You to Prison*, N.Y. TIMES (Oct. 26, 2017), https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html

Numerous jurisdictions give judges access to the COMPAS system in the pretrial arraignment context.[135] But there is a surprising paucity of public information about the manner of its implementation and its effects on the rates of pretrial release or on the composition of the pretrial detainee population. Two studies, one conducted in New York City and the other in an unnamed large American city, compared the predictive accuracy of different machine-learning algorithms with that of judges. Both found that the computational method generated less misranking of criminal defendants and less crime.[136] These studies, however, focus narrowly on the important question of gains to public safety that would result from a move from human to machine prediction. The studies appear to assume that jurisdictions will respond to algorithmic criminal justice instruments by using less pretrial incarceration to obtain the same levels of deterrence. It is not clear, though, why this assumption is warranted. These studies are silent as to the possibility or magnitude of racial effects–a striking omission given the large empirical literature documenting racial disparities in bail decisions.[137]

3. *Sentencing.* A recent survey of state sentencing practice comments that it is "improbable" that any convicted felon, whether an adult or juvenile, would be sentenced today without the aid of some sort of actuarial risk instrument, albeit not necessarily one that employs

---

[https://perma.cc/6ZQ2-NTAB]; *see also* Richard Berk & Jordan Hyatt, *Machine Learning Forecasts of Risk to Inform Sentencing Decisions*, 27 FED. SENT'G REP. 222, 223 (2015) (explaining advantages of machine-learning tools over the LSI-R); Richard F. Lowden, *Risk Assessment Algorithms: The Answer to an Inequitable Bail System?*, 19 N.C. J.L. & TECH. 221, 230–31 (2018) (listing jurisdictions that have adopted algorithmic tools).

135. *See* Angwin et al., *supra* note 16 ("As often happens with risk assessment tools, many jurisdictions have adopted Northpointe's [COMPAS] software before rigorously testing whether it works."). It is not wholly clear how much weight judges give to the COMPAS scores, or whether there is even a uniform practice.

136. *See* Richard A. Berk, Susan B. Sorenson & Geoffrey Barnes, *Forecasting Domestic Violence: A Machine Learning Approach To Help Inform Arraignment Decisions*, 13 J. EMPIRICAL LEGAL STUD. 94, 110 (2016) (finding that the release rate of 20 percent repeat offenders in a pool of domestic violent defendants could be dropped to a 10 percent rate through a move from judicial to machine-led determinations); Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237, 237–38 (2018) [hereinafter Kleinberg et al., *Human Decisions*] (finding that large decreases in offending rates could be achieved by moving from judicial to machine predictions in the bail context for violent crimes).

137. *See supra* note 129 and accompanying text. One study to address potential racial effects is Kleinberg et al., *Human Decisions*, *supra* note 136, at 237.

algorithmic means.[138] In some jurisdictions, such as Pennsylvania, New Hampshire, Arkansas, and Vermont, state law even affirmatively mandates the use of predictive instruments in the sentencing phase.[139] Such instruments have emerged as part of a "full-on embrace of practices that promise to reduce the risk of reoffending by convicted persons . . . ."[140] The federal FIRST STEP[141] Act also embraces algorithmic risk assessment tools for determining early prison release.[142]

In 2015, more than 60 risk assessment tools were used in sentencing contexts.[143] Risk assessments typically evaluate where within a statutorily calibrated sentencing range an offender's sentence should lie, accounting for "utilitarian crime-control grounds."[144] Some jurisdictions, such as Virginia, use a noncomputational "actuarial" instrument calibrated by age, felony record, offense type, employment, and gender, to sort nonviolent, low-risk offenders to alternative punishments such as probation, jail time, and restitution.[145] In other jurisdictions, computational instruments such as COMPAS are used

138. *See* Zachary Hamilton et al., *Designed To Fit: The Development and Validation of the STRONG-R Recidivism Risk Assessment*, 43 CRIM. JUST. & BEHAV. 230, 231 (2016).

139. *See* ARK. CODE ANN. § 16-93-615(a)(1)(B) (2016) ("The determination . . . shall be made by reviewing information such as the result of the risk-needs assessment to inform the decision of whether to release a person on parole by quantifying that person's risk to reoffend, and if parole is granted, this information shall be used to set conditions for supervision."); N.H. REV. STAT. ANN. § 504-A:15(I) (2017) (requiring that "[e]very person placed on probation or parole . . . be assessed by the department of corrections, using a valid and objective risk assessment tool, to determine that person's risk of recidivating" and that the results be used to determine the length of active supervision); 42 PA. CONS. STAT. § 2154.7 (2018) (adopting a "risk assessment instrument" that may aid in the determination of "the relative risk that an offender will reoffend and be a threat to public safety"); VT. STAT. ANN. tit. 13, § 7554c(a)(1) (2017) ("The objective of a pretrial risk assessment is to provide information to the court for the purpose of determining whether a person presents a risk of nonappearance or a risk of re-offense so the court can make an appropriate order concerning bail and conditions of pretrial release.").

140. Klingele, *supra* note 58, at 551–52.

141. Formerly incarcerated Reenter Society Transformed Safely Transitioning Every Person (FIRST STEP) Act, Pub. L. No. 115-391, 132 Stat. 5194 (2018) (to be codified at scattered sections of 18 U.S.C.

142. Bärí A. Williams, *The Prison-Reforming First Step Act Has a Critical Software Bug*, FAST COMPANY, Dec. 21, 2018, https://www.fastcompany.com/90284823/the-first-step-acts-reliance-on-algorithms-is-a-misstep [https://perma.cc/RTG8-KU2Q].

143. Anna Maria Barry-Jester, Ben Casselman & Dana Goldstein, *Should Prison Sentences Be Based on Crimes That Haven't Been Committed Yet?*, FIVETHIRTYEIGHT (Aug. 4, 2015), http://fivethirtyeight.com/features/prison-reform-risk-assessment [https://perma.cc/YU6N-AHCA].

144. *See* John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CLINICAL PSYCHOL. 489, 493–94 (2016).

145. *Id.* at 495.

for that same purpose.[146] Such instruments have only recently attracted judicial attention, including a high-profile constitutional challenge to Wisconsin's algorithm.[147]

Finally, I have located only one well-detailed example of a machine-learning algorithm being employed in the parole context. In 2010, the Pennsylvania Board of Probation and Parole started developing a machine-learning protocol using random forests to generate forecasts of recidivism to assist members of the Board in making discrete parole decisions.[148] When subject to performance evaluation seven years later, the algorithm was found to have reduced re-arrests for both nonviolent and violent crime.[149]

D. The Emerging Evidence of Race Effects

Race, its effects, and its legacies loom large in criminal justice.[150] To date, however, consideration of the racial effects (if any) of algorithmic criminal justice has been piecemeal. This section briefly surveys existing studies of algorithmic criminal justice systems that touch on questions of race. This survey hints at real reasons for closely analyzing the racial effects of algorithmic tools on criminal justice.

1. *Policing and the Problem of Tainted Training Data.* In the policing context, the unthinking use of algorithmic instruments will reinforce historical race-based patterns of policing.[151] This may occur because algorithmic predictions will vary depending on the quality of the training data used to construct the predictive function. For example, if the training data systematically omits data about certain subsets of a population—if it has what Kate Crawford calls "black

---

146. Angwin et al., *supra* note 16.

147. *See* State v. Loomis, 881 N.W.2d 749, 753 (Wis. 2016) (upholding the use of the COMPAS tool in sentencing in Wisconsin); *infra* notes 172–73 and accompanying text (discussing *Loomis*).

148. Richard Berk, *An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism*, 13 J. EXPERIMENTAL CRIMINOLOGY 193, 195 (2017) [hereinafter Berk, *Impact Assessment*]. For a clear and nontechnical explanation of random forests methods, see D. Richard Cutler et al., *Random Forests for Classification in Ecology*, 88 ECOLOGY 2783, 2784–85 (2007).

149. Berk, *Impact Assessment*, *supra* note 148, at 212–13.

150. *See supra* notes 10, 12.

151. Brayne, *supra* note 65, at 997 (arguing that "data-driven surveillance practices may be implicated in the reproduction of inequality . . . by deepening the surveillance of individuals already under suspicion; [by] widening the criminal justice dragnet unequally; and leading people to avoid 'surveilling' institutions that are fundamental to social integration").

holes"[152]—it will generate results that fail to account for some population. Such gaps can be a function of poor relations between law enforcement and certain communities. For example, imagine a jurisdiction that allocates patrol resources based on historical reports of crime. Neighborhoods characterized by poor relations with police might underreport crime, such that they receive fewer policing resources in the future.

Note, however, that algorithmic tools might also be used to compensate for asymmetrical data gaps. For example, the Shotspotter system records shots fired in urban environments. It can thus reveal neighborhoods in which residents do not report shootings to police.[153] This has at least the potential to mitigate historical enforcement gaps. To conclude that algorithmic instruments will either necessarily undermine, or necessarily perpetuate, historical imbalances in the allocation of criminal justice resources seems premature. They can do both. It just depends on how carefully training data is selected and on how the algorithm is then designed.

Policy distortions might also arise if historical data of police activity, deployed as training data for an algorithmic tool, is infected by the racial presumptions and stereotypes of the past officials. This kind of measurement effort has been found in studies concerning health services to "create decision and allocation biases."[154] A concern here is a variant on a worry common in medical research that "race is such a dominant category in the cognitive field that the 'interim solution' [of using race as a proxy for some other trait of interest] can leave its own indelible mark . . . ."[155] That is, race is such a freighted category that,

152. Kate Crawford, *The Anxieties of Big Data*, NEW INQUIRY (May 30, 2014), https://thenewinquiry.com/the-anxieties-of-big-data [https://perma.cc/25L7-4XUM].

153. *See* Sarah Griffiths, *Fighting a Losing Battle? AI ShotSpotter Computer Used To Track Gunfire Reveals Far More Shots Are Fired Than Are Ever Reported*, DAILY MAIL (Apr. 19, 2016, 9:10 AM), http://www.dailymail.co.uk/sciencetech/article-3547719/Fighting-losing-battle-AI-ShotSpotter-computer-used-track-gunfire-reveals-far-shots-fired-reported.html [https://perma.cc/YPJ3-XGLX]. Another example is a predictive model that could be used by hospitals to forecast readmissions of patients with pneumonia risk, which corrects for a counterintuitive pattern in the training data caused by the admission of asthmatic patients directly to intensive care. *See* Rich Caruana et al., *Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission*, 21 PROC. ASS'N FOR COMPUTING MACHINERY SPECIAL INT. GROUP ON KNOWLEDGE DISCOVERY AND DATA MINING INT'L CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING 1721, 1725–27 (2015).

154. Sendhil Mullainathan & Ziad Obermeyer, *Does Machine Learning Automate Moral Hazard and Error?*, 107 AM. ECON. REV.: PAPERS & PROC. 476, 478 (2017).

155. Troy Duster, *Race and Reification in Science*, 307 SCIENCE 1050, 1050 (2005); *see also* Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, & Marshall H. Chin,

once deployed, it cannot be taken back. Race effects can arise if data collected as a byproduct of police activity does "not pertain to future instances of crime" but rather to "instances of crime that become[] known to police."[156] That is, if police activity is predicted by race, then subsequent policing (and hence the costs of policing) will be unevenly allocated by race. The result is greater black exposure to arrest and incarceration.[157] Again, it is worth flagging the possibility of technical solutions. The computer science literature demonstrates that such effects can be buffered by incorporating an element of randomization into the algorithm.[158]

How forceful, as an empirical matter, are these concerns? One study of PredPol's algorithm suggests ground for concern. According to that study, when the algorithm used police data to generate predictions of narcotics crimes in Oakland, the algorithm recommended that twice as much policing resources be directed to black areas as white areas, despite that narcotics offenses were reasonably equally spread across both white and black areas.[159] A second study, also focused on the PredPol algorithm, identified the possibility of "runaway feedback loops," by which police are repeatedly sent back to the same neighborhood in a way that reinforces and exacerbates initial distortions in the training data.[160] Third, a recent

Ensuring Fairness in Machine Learning to Advance Health Equity, ANN. INTERNAL MED. 866, 870 (2018) (enumerating potential sources of bias in medical delivery).

156. Kristian Lum & William Isaac, *To Predict and Serve?*, SIGNIFICANCE, Oct. 2016, at 16 (emphasis omitted).

157. For findings that race, rather than criminality alone, was an indirect factor in deployment in one city (Seattle), see Katherine Beckett, Kris Nyrop, Lori Pfingst & Melissa Bowen, *Drug Use, Drug Possession Arrests, and the Question of Race: Lessons from Seattle*, 52 SOC. PROBS. 419, 435 (2005). *See also* Huq, *Disparate Policing*, *supra* note 5, at 2429–40 (discussing effects of such disproportionate allocations of policing resources).

158. Kroll et al., *supra* note 39, at 682–83. *But see* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 195 (2017) ("[W]hen a model produces biased outcomes due to the processes generating the input values, merely tweaking the distribution of data inputs will not solve the problem.").

159. Lum & Isaac, *supra* note 156, at 18 fig.2. The background estimate of the geographic distribution of narcotics offenses derives from separate national data. A survey of Los Angeles residents by the advocacy group Stop LAPD Spying Coalition similarly found police contact highly concentrated in less than two percent of the population and suggested that this flowed from the use of algorithmically derived "Chronic Offender Bulletins." Maha Ahmed, *Aided by Palantir, The LAPD Uses Predictive Policing to Monitor Specific People and Neighborhoods*, INTERCEPT (May 11, 2018, 9:15 AM), https://theintercept.com/2018/05/11/predictive-policing-surveillance-los-angeles/ [https://perma.cc/9H9Q-Z2Q4].

160. DANIELLE ENSIGN, SORELLE A. FRIEDLER, SCOTT NEVILLE, CARLOS SCHEIDEGGER, SURESH VENKATASUBRAMANIAN, RUNAWAY FEEDBACK LOOPS IN PREDICTIVE POLICING 11–12 (Sorelle A. Friedler & Christo Wilson eds., 2017), https://arxiv.org/abs/1706.09847

qualitative study of predictive policing in Los Angeles concluded that PredPol increases surveillance of low-income minority residents who are already under surveillance, widens the surveillance dragnet unequally, and drives members of the aforementioned communities to "avoid[] surveilling institutions."[161] These studies suggest that PredPol and similar technologies indeed distort the optimal allocation of policing resources. In part, this is because coercive resources will be inefficiently allocated. It may also happen because the individuals being regulated react to PredPol-driven interventions in different ways. For example, some engage in more avoidance behavior than others, leaving police resources concentrated on a small minority. Error rates across the population as a result of such variable responsiveness to new police intervention will consequently be uneven.[162]

Companies marketing algorithmic criminal justice instruments have evinced varying levels of concern about this possibility of racial effects as a result of biased training data. On the one hand, PredPol's manufacturer advertises its exclusion of "drug related offenses and traffic citation data from its predictions to remove officer bias."[163] Similarly, HunchLab underscores its reliance on non-crime-related data as a way of shielding predictions from the influence of potentially flawed past exercises of officer discretion.[164] On the other hand, the Sentencing Commission of Pennsylvania has incorporated arrest data into its sentencing algorithm, despite there being good reason to think that police discretion as to when and whom to arrest may have racial distortions.[165] How easy the problem is to fix without abandoning algorithmic prediction depends, of course, on the availability of unbiased substitute training data.

Finally, concerns about the polluting effect of historical training data are not limited to predictive algorithms. Studies of facial recognition technologies also suggest racial disparities in accuracy

[https://perma.cc/JF4D-MWN2].

161. Brayne, *supra* note 65, at 999.

162. For an account of the difference between prediction problems and causal inference problems, and the risks of confusing the two, see Athey, *supra* note 33.

163. *Machine Learning and Policing*, PREDPOL BLOG (July 19, 2017, 10:00 AM), http://blog.predpol.com/machine-learning-and-policing [https://perma.cc/39PV-XXE6].

164. *See* AZAVEA, *supra* note 109, at 12 ("Our belief is that the use of non-crime data sets as variables within a crime prediction system is important, because variables based solely upon crime data become skewed as predictions are used operationally.").

165. Barry-Jester et al., *supra* note 143.

rates. One 2012 study tested three commercial algorithms on mug shots from Pinellas County, Florida. African Americans were between five and ten percent less likely to be successfully identified—that is, more likely to be falsely rejected—than other demographic groups. It identified a similar decline for females relative to males and for younger subjects relative to older subjects.[166] A measure of caution, though, should be used in evaluating these studies. Much has changed in the domain of machine and deep learning since 2012.[167] It cannot be assumed that limitations on computational instruments that existed then still hinder analogous tools today.

2. *Bail/Sentencing Predictions and the Problem of Distorting Feature Selection.* The problems with algorithmic criminal justice do not begin and end with a concern about tainted historical training data. Attention to the bail and sentencing context suggests that even when there is no allegation of tainted training data, algorithmic criminal justice can generate concerns related to racial equity as a consequence of feature selection decisions. Even if these concerns focus on arguably unanticipated results, they might nonetheless have empirically consequential magnitudes.

Perhaps the highest profile debate concerning the racial effects of algorithmic instruments in criminal justice, though, has focused on the COMPAS algorithm. To recapitulate the facts: Analyzing COMPAS data from Broward County, Florida, ProPublica observed that the algorithm was "likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants," and to mislabel white defendants as "low risk more often than black defendants."[168] That is, conditional on being a nonrisky type, the COMPAS algorithm is more likely to overstate the risk presented by a black person than a white person. Northpointe's

166. Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge & Anil K. Jain, *Face Recognition Performance: Role of Demographic Information*, 7 IEEE TRANSACTIONS ON INFO. FORENSICS AND SECURITY 1789, 1797 tbl.II, 1800 (2012); *cf.* P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad & Alice J. O'Toole, *An Other-Race Effect for Face Recognition Algorithms*, 8 ACM TRANSACTIONS ON APPLIED PERCEPTION 14:1, 14:5 (2011) (finding race-specific bias in algorithms primarily designed for East Asian and Caucasian facial identification).

167. A particularly vivid illustration of this is the dramatic increase in the quality of machine translation tools. *See* Gideon Lewis-Kraus, *The Great A.I. Awakening*, N.Y. TIMES MAG. (Dec., 14, 2016), https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html [https://perma.cc/KUM9-XPJW] (describing the increase in quality of machine-learning translation).

168. Angwin et al., *supra* note 16.

response did not focus on this measure of false positives (or, correlatively, the measure of false negatives that list in favor of whites). Instead, it identified the pool of individuals assigned a certain risk score as the relevant pool of comparators and showed that, within that pool, white and black defendants were equally likely to recidivate.[169] The ratio it emphasized, that is, takes as a denominator the group identified as high risk within each racial group and then asks how many of those identifications are erroneous. This is the rate of false positives conditional on being identified as a risky type. The resulting debate might well be understood not in terms of whether the COMPAS algorithm is racially discriminatory—after all, there is no dispute that the algorithm did not include race as a feature—but rather what kind of racial effects count in a normative or legal evaluation of its performance.[170]

COMPAS's use in criminal sentencing has been challenged on various constitutional grounds. But its race-related effects remain untested in court. The most extensive judicial treatment of COMPAS, offered by the Wisconsin Supreme Court in *State v. Loomis*,[171] involved a criminal defendant's due process claim that he was entitled to know how the algorithm calculated his risk score.[172] Rejecting a challenge to the way in which the algorithm accounts for a suspect's gender, the Wisconsin Court noted "concerns regarding how a COMPAS assessment's risk factors correlate with race."[173] Unfortunately, the Court did not connect that observation with either a legal theory pursuant to which such correlations might be objectionable or, alternatively, a normative basis for concern notwithstanding legality. Otherwise, commentators have noted that actuarial sentencing tools, whether algorithmic or not, might have more or less disparate racial impact or "inequitable social consequences."[174] But precisely what these "consequences" might be remains unclear.

---

169. Dieterich et al., *supra* note 20, at 2–3.

170. *See* Feller et al., *supra* note 22 (observing that algorithmic fairness should also entail an inquiry into the effects of that algorithm).

171. State v. Loomis, 881 N.W.2d 749 (Wisc. 2016).

172. *Id.* at 761–62 (holding that because the algorithm employed only publicly available data, or data that a defendant has supplied, the defendant could have denied or explained any information that was employed to develop his risk score). It is worth noting that the court's analysis here misses the force of the defendant's argument. The latter seemed to object not so much to the nondisclosure of information about his own circumstances but to the manner in which that information was evaluated and weighted by the COMPAS algorithm.

173. *Id.* at 763.

174. Monahan & Skeem, *supra* note 144, at 507. For an empirical study that renders these

Nevertheless, the COMPAS debate suggests that concerns about racial equity can persist *even if the inputs to the algorithm are not tainted by any historical bias*. Part of my aim here, particularly in Part III, is to explain how this can be so. For now, it suffices to say that earlier commentators who have suggested that algorithmic bias can be addressed exclusively through "a transparency of inputs and outputs" may have captured only one part of a larger normative picture.[175]

3. *Conclusion: An Incomplete Evidentiary Record.* Race interacts with algorithmic criminal justice tools in one of three ways. First, racial animus or stereotypical thinking can infect and distort training data. Second, race may be a feature used for classification. Third, the classification rule may have predictable effects that seem asymmetrical between racial groups. Scholars' thinking about and responses to the racial effects of algorithmic criminal justice instruments have been ad hoc and unsystematic. We have, at best, fragments of a broader account of how such effects arise and their consequences. Hence, empirically understanding the manner in which algorithmic tools redistribute coercive outcomes should remain an important focus of research. Still, even with limited evidence in hand, it seems reasonable to think about the appropriate normative framework for evaluating these instruments' racial effects—especially given the long and troubled interaction between criminal justice policy and widely held beliefs about racial differences in culture and behavior.

The conceptual tools for that investigation are plainly wanting at the moment. There is no general agreement on the ways in which racial effects might count against the adoption or continued use of an algorithm. Insufficient attention, moreover, has been paid to the difference between tainted training data and problematic feature selection. There is also no general understanding of what it means to say that feature selection is flawed. Nor is there any consideration of how different kinds of racial effects might be weighed against each other. The field is ripe, in short, for more careful theorization of what it precisely means to talk about racial equity in algorithmic criminal justice.

---

concepts with more precision, see Jennifer L. Skeem & Christopher T. Lowenkamp, *Risk, Race, and Recidivism: Predictive Bias and Disparate Impact*, 54 CRIMINOLOGY 680, 702–03 (2016) (analyzing the relation of the Post Conviction Risk Assessment tool and future arrests, finding that scores tracked the same level of recidivism within each group).

175. Chander, *supra* note 43, at 1039 (emphasis omitted).

II. Equal Protection and Algorithmic Justice

But is such theorization needed? The Equal Protection Clause of the Fourteenth Amendment, after all, purports to provide a general norm regulating the state's use of race. Perhaps constitutional equality jurisprudence provides the needful criterion for evaluating the race effects of algorithmic criminal justice.

Or perhaps not. I describe and apply in this Part conventional doctrinal norms under the Equal Protection Clause. The core takeaway is that the dominant intent- and classification-focused calibration is ill suited to the forms and dynamics of algorithmic criminal justice tools. To be sure, one might choose to apply the litmus tests supplied in the jurisprudence. But given that these focus on qualities of state action that are irrelevant, or barely relevant, to the way that algorithms in practice work, it is hard to see why one would do so. If there is a lesson here, it is about the woeful inadequacy of our constitutional equality norms for the contemporary world.

A. What Equal Protection Protects

Equal protection doctrine imposes two fundamental prohibitions on governmental action touching on race.[176] One concerns formal racial classifications. The other pertains to racialized intentions. In contrast, the Court has either rejected or ignored concerns about the illegitimate nature or delegitimizing consequences of raw racial disparities in criminal justice.

Almost since its inception, constitutional equal protection has been understood to prohibit most laws containing an explicit racial classification as well as laws that assign rights or burdens based on racial classifications.[177] The first major judicial interpretation of the Equal Protection Clause, *Strauder v. West Virginia*,[178] concerned a state statute limiting jury service to "white male persons . . . twenty-one years of age."[179] Invalidating the conviction of an African American man by an all-white jury, the Court explained that the statute's want of

176. Concerns about racial equity in criminal law need not be expressed in terms of equal protection jurisprudence. Many cases formally concerning due process arose in the context of discriminatory law enforcement and are plausibly understood in terms of the Court's desire to constrain the latter's discretion. My concern in this Part is the formal doctrinal specification of equality, not its potential jurisprudential substitutes.

177. A leading exception is *Plessy v. Ferguson*, 163 U.S. 537, 550 (1896).

178. Strauder v. West Virginia, 100 U.S. 303 (1879).

179. *Id.* at 305 (citation omitted).

facial equality violated the Constitution's guarantee of "immunity from inequality of legal protection."[180] Racial classifications today are not per se invalid. Rather, they now trigger searching judicial review of their tailoring and means-ends rationality, an inquiry known as "strict scrutiny."[181]

Notoriously, strict scrutiny is not necessarily fatal.[182] In *Fisher v. University of Texas at Austin*,[183] for example, the Court upheld the University of Texas at Austin's admission program, even though it accounted for race as one element of a "Personal Achievement Index," or PAI.[184] The latter satisfied strict scrutiny because the university "articulated concrete and precise goals" in relation to educational diversity, relied on "both statistical and anecdotal" evidence" of a need for affirmative action, and engaged in ongoing deliberation about admissions protocols.[185] Precisely how *Fisher* calibrated strict scrutiny, though, is difficult to say. Educational diversity is not easily reduced to "concrete and precise" terms. Nothing the Court said illuminated how it tested the means-ends rationality behind the university's actions.[186] Yet in other contexts, it has construed strict scrutiny to work a near-categorical prohibition on similarly race-conscious government action.[187] For instance, in an earlier capital habeas case, the Court made the errant suggestion that race is "totally irrelevant to the sentencing

180. *Id.* at 310.

181. *See* Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (imposing such scrutiny whenever "the government distributes burdens or benefits on the basis of individual racial classifications"); *see also* Gratz v. Bollinger, 539 U.S. 244, 270 (2003) (describing the use of such classifications as "pernicious" (citation omitted)); Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 235 (1995) ("Federal racial classifications, like those of a State, must serve a compelling governmental interest, and must be narrowly tailored to further that interest.").

182. Grutter v. Bollinger, 539 U.S. 306, 326 (2003) ("Strict scrutiny is not 'strict in theory, but fatal in fact.'").

183. Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198 (2016).

184. On the construction of the PAI, see *Fisher v. University of Texas at Austin*, 570 U.S. 297, 304 (2013).

185. *Fisher*, 136 S. Ct. at 2211–12.

186. David A. Strauss, Fisher v. University of Texas *and the Conservative Case for Affirmative Action*, 2016 SUP. CT. REV. 1, 16 ("The central problem is that judgments about the kind and degree of diversity that a student body should have . . . are simply not susceptible to precise metrics.").

187. *See* Richard H. Fallon, Jr., *Strict Judicial Scrutiny*, 54 UCLA L. REV. 1267, 1302 (2007) ("According to one interpretation, strict scrutiny embodies a nearly categorical prohibition against infringements of fundamental rights, regardless of the government's motivation, but subject to rare exceptions when the government can demonstrate that infringements are necessary to avoid highly serious, even catastrophic harms.").

process."[188] Such evanescent dicta, however, are probably too frail to support any firm conclusion.

Second, the Equal Protection Clause's regulation of racial considerations extends to instances in which the state harms an individual because of "a racially discriminatory purpose."[189] This requires litigants to "show both that the passive enforcement system had a discriminatory effect and that it was motivated by a discriminatory purpose."[190] The Court has not defined with precision what counts as a "racially discriminatory purpose."[191] But at a minimum, it seems to include naked, taste-based aversion to a group based exclusively on race.[192] So the Court recently explained that evidence that a juror relied on "racial stereotypes or animus to convict a criminal defendant" would be sufficient to warrant reversal of that conviction on Sixth Amendment grounds.[193] Even here, the doctrine is not without ambiguity. It is not clear, for instance, whether a state actor shown to have made a decision based on racial animus could plausibly respond that their action could nonetheless be upheld because it survived strict scrutiny. Analytically, it is hard to see how a measure based on an invidious stereotype could ever be closely fitted to a

---

188. Zant v. Stephens, 462 U.S. 862, 885 (1983).

189. Washington v. Davis, 426 U.S. 229, 240 (1976) (holding that "the basic equal protection principle that the invidious quality of a law claimed to be racially discriminatory must ultimately be traced to a racially discriminatory purpose"). Racial intent must be the but-for cause of an action. Pers. Adm'r. v. Feeney, 442 U.S. 256, 279 (1979) (finding proof of discriminatory purpose requires showing that government decision-maker "selected or reaffirmed a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group").

190. Wayte v. United States, 470 U.S. 598, 608 (1985).

191. *See* Huq, *Discriminatory Intent*, *supra* note 41, at 21–36 (describing five different theories of discriminatory purpose in the case law); *accord* David A. Strauss, *Discriminatory Intent and the Taming of* Brown, 56 U CHI. L. REV. 935, 947 (1989) (noting that even canonical cases such as *Brown v. Board of Education* did not clarify "which conception of discrimination [the Court] embraced, or how far the principle of [Equal Protection] extended").

192. This is what economists call taste-based discrimination. GARY S. BECKER, THE ECONOMICS OF DISCRIMINATION 14–15 (2d ed. 1971) (modeling taste-based discrimination as a "discrimination coefficient," which "acts as a bridge between money and net costs. Suppose an employer were faced with the money wage rate π of a particular factor; he is assumed to act as if $\pi(1 + d_i)$ were the net wage rate, with d_i as his [discrimination coefficient] against this factor" (emphasis omitted)).

193. Peña-Rodriguez v. Colorado, 137 S. Ct. 855, 869 (2017); *see also* Foster v. Chatman, 136 S. Ct. 1737, 1747–55 (2016) (holding that the Georgia Supreme Court had made a "clearly erroneous" decision when it found that prosecution use of preemptory strikes in a capital case was *not* animated by a discriminatory purpose in the face of lurid evidence to the contrary).

compelling state interest. So it may be that the question does not arise because it has little operational importance.[194]

Race can infiltrate the mind in forms other than animus. For example, a rational reliance on race as a statistically accurate proxy for some other policy-salient quality is analytically distinct from taste-based discrimination.[195] The Court has not been clear on whether such statistical discrimination triggers constitutional concerns. On the one hand, in the 2007 case *Johnson v. California*,[196] a majority of the Justices held that strict scrutiny applied to an unwritten California prison policy of racially segregating prisoners for up to sixty days each time they enter a new correctional facility with the aim of mitigating violence between gangs of different races.[197] On the other hand, lower federal courts routinely shake off challenges to race-specific suspect descriptions. The Supreme Court has consistently and repeatedly declined to intervene to prevent the latter practice.[198] All that can safely be said is that, at least in some instances, statistical discrimination will be subject to close judicial scrutiny, and sometimes it won't be. The cut-point between those domains remains to be defined.

Moreover, it seems likely that not all racial animus in the criminal justice system is crisply articulated in the Queen's English. Instead, we might expect overt racial labelings to be the exception, with race more commonly embedded in "tacit," unspoken understandings.[199] The only evidence of the latter's operation may be downstream differential

194. I am not sure, however, that this conclusion would be warranted. Consider, for example, an action shown to be tainted by racial animus but that could be defended as narrowly tailored given different motivational premises and additional evidentiary support. The so-called travel ban might have this character. For extended discussion, see Aziz Z. Huq, *Article II and Antidiscrimination Norms*, 117 MICH. L. REV. (forthcoming 2019).

195. Kasper Lippert-Rasmussen, *"We Are All Different": Statistical Discrimination and the Right To Be Treated as an Individual*, 15 J. ETHICS 47, 54 (2011) (providing a formal definition of such rational discrimination). This is what economists call statistical discrimination; *see* Kenneth J. Arrow, *The Theory of Discrimination*, *in* DISCRIMINATION IN LABOR MARKETS 3, 24–27 (Orley Ashenfelter & Albert Rees eds., 1973) ("Skin color and sex are cheap sources of information. Therefore prejudices (in the literal sense of pre-judgments, judgments made in advance of the evidence) about such differentia can be easily implemented.").

196. Johnson v. California, 543 U.S. 499 (2005).

197. *Id.* at 503.

198. *See, e.g.*, Monroe v. City of Charlottesville, 579 F.3d 380 (4th Cir. 2009), *cert. denied*, 130 S. Ct. 1740 (2010) (denying certiorari in a case exempting race-based suspect selection from equal protection scrutiny); Brown v. City of Oneanta, 221 F.3d 329 (2d Cir. 1999), *cert. denied*, 534 U.S. 816 (2001) (same).

199. On the notion of tacit understandings, see Michael Polanyi, *The Logic of Tacit Inference*, 41 PHILOSOPHY 1, 2–3 (1966).

effects on suspects and defendants of different races. In contrast to its strict superintendence of overt classifications, however, the Court has rejected the argument that a constitutional violation can be made out by a showing of disparate racial impact. In *McCleskey v. Kemp*,[200] the Court rejected an equal protection challenge to Georgia's capital punishment system based on econometric evidence of racial disparities.[201] Lower courts have extended that holding to the distinct context of statistical evidence about the role of race in a single decision-maker's actions over time (e.g., a single district attorney over a number of years).[202]

Oddly, both the Court's embrace of the racial-intent rule and its repudiation of a disparate-treatment rule have been justified by the need to maintain the criminal justice system in good working order. In *McCleskey*, Justice Powell's majority opinion expressed alarm that the defendant's challenge would "throw[] into serious question the principles that underlie our entire criminal justice system."[203] In Powell's view, it was inconceivable that the Constitution would "require that a State eliminate any demonstrable disparity that correlates with a potentially irrelevant factor in order to operate a criminal justice system . . . ."[204] On the other hand, the Court has explained decisions enforcing closer invigilation of race's role in the jury deliberation context as "necessary to prevent a systemic loss of confidence in jury verdicts, a confidence that is a central premise of the Sixth Amendment trial right."[205] So the Court appears to believe that the legitimacy of a criminal justice system simultaneously requires keen alertness to concerns of racial justice and also a willful blindness to such concerns.

Stated in summary form, then, current constitutional jurisprudence compels judges to maintain the stability of the criminal justice system by ignoring racial disparities, by isolating racial

---

200. McCleskey v. Kemp, 481 U.S. 279 (1987).

201. *Id.* at 292–93.

202. John H. Blume, Theodore Eisenberg & Sheri Lynn Johnson, *Post-*McCleskey *Racial Discrimination Claims in Capital Cases*, 83 CORNELL L. REV. 1771, 1794 (1998) (collecting cases); *see also* Chavez v. Ill. State Police, 251 F.3d. 612, 645, 648 (7th Cir. 2001) (finding no discriminatory purpose despite statistical showing of racial disparities in traffic stops). *But cf.* Reva B. Siegel, *Blind Justice: Why the Court Refused To Accept Statistical Evidence of Discriminatory Purpose in* McCleskey v. Kemp*-and Some Pathways for Change*, 112 NW. U. L. REV. 1269, 1288 (2018) (flagging limits to *McClesky*'s scope).

203. *McCleskey*, 481 U.S. at 315.

204. *Id.* at 319.

205. Peña-Rodriguez v. Colorado, 137 S. Ct. 855, 869 (2017).

classifications, and by extirpating (some) racial animus. It is a doctrinal status quo that poorly fits emergent algorithmic realities.

B. How Equal Protection Fails to Speak in Algorithmic Terms

Equal protection doctrine is sharply criticized by those who perceive it to embody a judicial failure to account for the diffusion and impact of racial effects in society, let alone our highly racially stratified criminal justice system.[206] I set these concerns aside here (although I share them) and take the doctrine seriously on its own terms. Even then, I find reasons to doubt that the current doctrine can respond effectively to the questions of race raised by algorithmic criminal justice. The concerns of constitutional law simply do not map onto the ways in which race impinges on algorithmic criminal justice. The result is a gap between legal criteria and their objects.

Crucially, the two main doctrinal touchstones of bad intent and bad classifications provide scant traction for the analysis of algorithmic criminal justice. Both hinge on concepts that translate poorly, if at all, to the algorithmic context and are not easily adapted for application to that end. A focus on racial animus will almost never be fruitful. A focus on classification leads to perverse and unjustified results. The replacement of unstructured discretion with algorithmic precision, therefore, thoroughly destabilizes how equal protection doctrine works on the ground. The resulting mismatches compel my conclusion that a new framework is needed for thinking about the pertinent racial equity questions.

1. *The Trouble with Intent.* Taking intent as a touchstone of equal protection directs attention to questions at best tangential to the potential role of race in algorithmic criminal justice. To be sure, problematic intent might enter into algorithmic design in different ways, one of which is easily accounted for in doctrinal terms. But, in general, intent will rarely be the crux of the matter.

To begin, I suspect that the notion of machine intentionality is sufficiently counterintuitive to find no place in constitutional law. Speculation about a future of "superintelligent" artificial intelligences

206. Recent critiques include Ian Haney-López, *Intentional Blindness*, 87 N.Y.U. L. REV. 1779, 1828 (2012) (arguing that the Court has "split equal protection into the separate domains . . . one governing affirmative action and the other discrimination against non-Whites" in a move that has made it systematically easier for white plaintiffs to prevail) and Russell K. Robinson, *Unequal Protection*, 68 STAN. L. REV. 151, 154 (2016) (contending that "the Supreme Court has steadily diminished the vigor of the Equal Protection Clause in most respects").

aside,[207] the transformation of training data into new schemes of classification by machine learning or deep learning does not obviously map onto familiar forms of human intentionality. The most advanced artificial intelligences can now pass the Turing test[208] and defeat (human) world champions at Go.[209] But even these machines do not obviously possess the sort of psychological interiority commonly thought to be a necessary predicate to intentionality.[210] Talk of machine intentionality, therefore, is either premature or a badly poised metaphor. It is better to treat the algorithm itself as irrelevant to the constitutional analysis so far as intentionality is concerned.

Bracketing the machine-learning tool as agent, however, there are two possible ways in which intention might enter the picture. First, an algorithm's designer might be motivated by either an animosity toward a racial group, or else a prior belief that race correlates with criminality, and then deliberately design the algorithm on that basis. Barocas and Selbst call this "masking."[211] Masking might occur through either a choice to use polluted training data or the deliberate selection of some features but not others on racial grounds. For instance, it is well understood that when employers ignore credit score information, they tend to search for proxies that have the inadvertent effect of deepening racial disparities.[212] A discriminatory algorithm designer will leverage

207. *Cf.* NICK BOSTRAM, *SUPERINTELLIGENCE*: PATHS, DANGERS, STRATEGIES 52 (2014) (defining superintelligence as "intellects that greatly outperform the best human minds across many very general cognitive domains").

208. In June 2014, an artificial intelligence passed the Turing test, arguably for the first time. Kevin Warwick & Huma Shah, *Can Machines Think? A Report on Turing Test Experiments at the Royal Society*, 28 J. EXP. & THEO. ARTIFICIAL INTELLIGENCE 989, 990 (2016). The Turing test involves human judgments about natural language conversations between a computer and a machine; a machine passes the test if the human observer is unable to distinguish human from machine. *Id.*

209. Silver et al., *supra* note 90, at 490.

210. Accounts of this interiority vary. In one influential definition, intentions are "conduct-controlling pro-attitudes, ones which we are disposed to retain without reconsideration, and which play a significant role as inputs into reasoning . . . ." MICHAEL E. BRATMAN, INTENTION, PLANS, AND PRACTICAL REASON 20 (1987). In another view, when *S* is doing *A* intentionally, *S* knows that she is doing *A*. G.E.M. ANSCOMBE, INTENTION 11–15 (2d ed. 1963). Machines lack attitudes or self-knowledge in the relevant senses.

211. Barocas & Selbst, *supra* note 51, at 692 ("[D]ecision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes.").

212. *See* Robert Clifford & Daniel Shoag, *"No More Credit Check Score": Employer Credit Check Bans and Signal Substitution* 4 (Fed. Reserve Bank of Boston, Research Dep't Working Papers, No. 16-10, 2016), https://www.bostonfed.org/-/media/Documents/Workingpapers/PDF/wp1610.pdf [https://perma.cc/SQU6-6BVQ].

such knowledge to fashion instruments that yield the disparate racial effects they believe to be warranted *a priori*. Without knowing the full spectrum of features that could, conceivably, have been included in the training data—which can be "enormous"[213]—it will be difficult or impossible to diagnose this kind of conduct absent direct evidence of discriminatory intent.[214] It will, moreover, be especially difficult to show that, but for race, a specific feature would or would not have been included, as the doctrine requires.[215] A basic principle of "feature selection" instructs that one should keep the important features and discard the unimportant ones.[216] To the extent that masking occurs, therefore, it seems clear that the litigation process would rarely yield evidence of such intentional manipulation of the algorithm's design.

Another reason to set aside the masking phenomenon, however, is the fact that it does not appear to be a significant one in practice. Part of the reason for this is that racial animus has a performative, interpersonal aspect. Racial discrimination commonly entails an effort by one group to "produce esteem for itself by lowering the status of another group,"[217] correlatively producing a "set of . . . privileges[] and benefits" of superordinate group membership.[218] Masking is a form of discrimination that involves no interpersonal interaction and no esteem-affirming performance.

But polluted data may be used *faute de mieux* or due to ignorance. Imagine a jurisdiction where African Americans were targeted for frequent and unjustified police contact, such that the pool of arrestees and convicted criminals may be biased by an underrepresentation of nonblack individuals.[219] Or consider a jurisdiction in which black

213. Athey, *supra* note 33, at 483.

214. Training data will often have so many potential features that inferring the reason for the inclusion of some and exclusion of others will often not be feasible. *Id.* at 483.

215. Pers. Adm'r v. Feeney, 442 U.S. 256, 279 (1979) (noting that proof of discriminatory purpose requires showing that a government decision-maker "selected or reaffirmed a particular course of action at least in part 'because of,' not merely 'in spite of,' its adverse effects upon an identifiable group").

216. ALPAYDIN, *supra* note 32, at 110.

217. Richard H. McAdams, *Cooperation and Conflict: The Economics of Group Status Production and Race Discrimination*, 108 HARV. L. REV. 1003, 1044 (1995). For a similar idea, see George A. Akerlof, *Discriminatory, Status-Based Wages Among Tradition-Oriented, Stochastically Trading Coconut Producers*, 93 J. POL. ECON. 265, 265 (1985).

218. Cheryl I. Harris, *Whiteness as Property*, 106 HARV. L. REV. 1707, 1713 (1993). For a seminal account of this concept, see DAVID R. ROEDIGER, THE WAGES OF WHITENESS: RACE AND THE MAKING OF THE AMERICAN WORKING CLASS (1991).

219. For findings of such disparities, see, for example, *Floyd v. City of New York*, 959 F. Supp. 2d 540, 573–74 (S.D.N.Y. 2013) (reporting statistical racial disparities in arrest rates by New York

neighborhoods are underserved by police responses to emergency calls, which might, in contrast, generate data on the distribution of crime with a black (or grey) hole in respect to African American neighborhoods.[220] The two hypotheticals can even be combined: A jurisdiction might underserve black neighborhoods by understaffing responses to 911 calls at the same time as concentrating a disproportionate amount of street policing resources on the same neighborhoods.[221] An algorithm trained on police-generated data from any of these jurisdictions is likely to allocate resources in ways that reflect and perhaps entrench disparities in how policing resources are allocated. But police might adopt the algorithm without considering racial effects or with an honest but erroneous belief that their training data is untainted.

The relevant intent in these examples, though, differs in two important ways from canonical instances of impermissible intent in equal protection case law. First, in the absence of an express policy, the use of racial preferences by officials in activities that produce training data will generally be highly decentralized and uncoordinated. Policing decisions, and to a lesser extent bail determinations and sentencing, are scattered rather than centralized forms of state action. Individual officers or magistrates have a large degree of discretion in consequence of their sheer numerosity and the difficulty of monitoring their decisions. It is hardly clear how a court could or would make a determination of joint intent when confronted with an extensive multitude ungoverned by formal decision-making procedures. Constitutional doctrine has not developed an intellectual toolkit for aggregating a large number of dispersed individual motives so as to ascertain whether a but-for standard of intentionality has been met by a collectivity.

A similar problem arises in the legislative context, where many individuals bring to bear potentially diverse motives in order to shape

police officers) and *Boston Police Commissioner Announces Field Interrogation and Observation (FIO) Study Results*, BOS. POLICE DEP'T NEWS (Oct. 8, 2014), http://bpdnews.com/news/2014/10/8/boston-police-commissioner-announces-field-interrogation-and-observation-fio-study-results [https://perma.cc/7M4L-R3EG] (noting a similar disparity in Boston).

220. *See, e.g.*, Cent. Austin Neighborhood Ass'n v. City of Chicago, 1 N.E.3d 976, 979 (Ill. App. Ct. 2013) (describing allegations of longer response times to 911 calls in minority neighborhoods in Chicago).

221. As appears to be the case with Chicago. *See id.*; Aamer Madhani, *Chicago Police and ACLU Agree to Stop-and-Frisk Safeguards*, USA TODAY, (Aug. 7, 2015, 9:14 AM), http://www.usatoday.com/story/news/2015/08/07/chicago-police-agree-reform-stop-and-frisk/31277041 [https://perma.cc/4PUT-V7S9].

singular institutional acts with the force of law. Equal protection law has struggled with how to conceptualize the concept of intent in the legislative context so unsuccessfully that one influential commentator has advocated wholesale retreat from judicial accounting for legislators' subjective intent. In his view, the task of principled aggregation is simply too hard for judges.[222] Unlike legislatures, a plurality of geographically and temporally diffused cohorts of officials (whether police or magistrates) lack any stable procedures or mechanisms for eliciting and formalizing a singular intent.[223] Their ability to form a coherent, let alone legally relevant, intent may seriously be doubted.

Even if such an intentionality could be derived from a diffuse haze of discrete policing decisions or detention-related judgments, it is not clear whether the mere incorporation by reference of such historical judgments into new, forward-looking algorithmic tools would trigger equal protection concerns. Even if historical intent can be inferred successfully, there remains a question of whether reliance on flawed historical data *counts* as a constitutionally relevant form of intent. It is certainly possible for bad intent to endure over time. Indeed, the Court has invalidated state laws enacted to preserve "white supremacy" many decades before litigation began; in so doing, it rejected the notion that "events occurring in the [intervening] years [could have] legitimated the provision."[224] But there are no equal protection cases in which the Court has considered outcomes resulting from concededly discriminatory official action that in turn was adopted by a new and different actor as the rationale for forward-looking policy.[225] In short,

222. Richard H. Fallon, Jr., *Constitutionally Forbidden Legislative Intent*, 130 HARV. L. REV. 523, 533–34 (2016) ("[U]ltimate determinations of constitutional validity should always depend on the content and effects of challenged legislation, not the subjective intentions of the enacting legislature . . . .").

223. *Cf.* CHRISTIAN LIST & PHILIP PETTIT, GROUP AGENCY: THE POSSIBILITY, DESIGN, AND STATUS OF CORPORATE AGENTS 81 (2012) ("[A] group's performance as an agent depends on how it is organized: on its rules and procedures for forming its propositional attitudes . . . ."). In most cases, the groups relevant to algorithmic criminal justice have no such rules or procedures.

224. Hunter v. Underwood, 471 U.S. 222, 229, 233 (1985).

225. The closest analog of which I am aware arises under the Fair Housing Act, where there can be a question whether a municipal decision on, say, taxes or zoning causes a pattern of residential racial segregation. *Cf.* Tex. Dep't. of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2523 (2015) (demanding a showing of "robust causality"). But the question here is not one of causation; it is a question of whether the intentions of the original police or magistrate ought to be imputed to the algorithm, given their influence on the training data.

there is simply no way of knowing whether a "relay-race" theory of bad intent would pass muster in constitutional law.

Perhaps the closest analog to this problem of governmental reliance on flawed data arises in the Fourth Amendment context. In that domain, the Court has declined to treat the flaws in a first-moving official's behavior as infecting a second, subsequently acting official's decision to depend on that first officer. For example, when a police officer relies on a recalled warrant mistakenly distributed by another police force, the latter's mistake of law is not imputed to the arresting officer such that evidence must be excluded.[226] The analogy is inexact.[227] But the Fourth Amendment's stingy treatment of imputed fault suggests that an intent-focused equal protection lens will have limited traction in the algorithmic criminal justice context.

Still, that theoretical problem may be precisely that—theoretical. Even if flawed training data were identified, it seems unlikely that its tainted nature could suffice to establish a constitutional concern in practice. Any moderately competent municipality found using flawed data would hardly concede that it was doing so intentionally. Rather, it would be far more likely to defend its decision as the best option given historically shaped constraints. Because a constitutional violation cannot be shown unless the state relied on race as a ground of decision, as opposed to acting in spite of race,[228] this defense would likely succeed. As a practical matter, therefore, the narrow definition of intent in equal protection doctrine would likely insulate racially tainted training data from legal attack.

This means that none of the pathways for integrating intent into the equal protection analysis of algorithmic criminal justice are likely to prove fruitful. None of them are well suited for a consideration of the ways in which race in practice interacts with algorithmic criminal justice. Equal protection doctrine was designed to police the dispersed, open-ended discretionary judgments of street-level officials. It does a

---

226. Herring v. United States, 555 U.S. 135, 140 (2009); *see also* Arizona v. Evans, 514 U.S. 1, 14–15 (1995) (acknowledging the same result for errors by a judicial administrator). The Court, however, is willing to impute another officer's knowledge of information salient to the legality of a search when doing so renders a search lawful. Whiteley v. Warden, 401 U.S. 560, 568 (1971). Although these positions can be squared, it is striking that imputation is available only when it expands state authority.

227. The availability of exclusion in Fourth Amendment cases is said to turn on the deterrent effect of that remedy. *Herring*, 555 U.S. at 144. The consequential focus on deterrence is absent when one is concerned with attributions of intentionality.

228. Pers. Adm'r v. Feeney, 442 U.S. 256, 279 (1979).

poor job when applied to the very different context of algorithm design and application. It is therefore necessary to consider the logic of anticlassification as an alternative lens.

2. *The Trouble with Classification.* The anticlassification strand of equal protection doctrine prohibits the government from "classify[ing] people either overtly or surreptitiously on the basis of a forbidden category" such as race.[229] At first blush, it seems a natural fit: Algorithms work by applying categories to training data (when defining features) and then generating novel classification rules to apply to test data. A rule to the effect that race or ethnicity could not be used either as a feature or as an element of a classifier, absent narrow tailoring to a compelling state interest, would seem to be a natural fit. Such a rule, however, would be unmoored from the justifications for an anticlassificatory rule. It would also engender results that contradict the assumed purposes of the rule.

The anticlassification account of equal protection is premised on two main justifications. First, it is motivated by a concern that the state's use of racial classifications will facilitate or amplify private discrimination.[230] This worry is premised on an empirical claim that a "perception . . . fostered by [government]" of differences between racial groups "can only exacerbate rather than reduce racial prejudice."[231] The foundation of this empirical claim is hardly clear. Why would the communicative effect of state racial classifications entail a legitimation of private animus? The causal link here is not obvious.[232] One interpretation of the Court's argument might start with the Court's claim that race is "'in most circumstances irrelevant' to any constitutionally acceptable legislative purpose."[233] Read

229. Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003).

230. Anderson v. Martin, 375 U.S. 399, 402 (1964) (holding that a Louisiana statute, which mandated the designation of a candidate's race on election ballots, violated the Equal Protection Clause because it enlisted the power of the state to enforce private racial prejudices).

231. Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 229 (1995) (emphasis omitted).

232. It is not clear why the reasonable observer would draw an inference about *a racial group*, though, instead of an inference that the *government* was unjustified and irrational in its action. In any event, the claim that racial identity is not salient in a context where racial preferences retain a powerful hold is a deeply dubious one. For an estimate of the prevalence of racial animus using an innovative empirical method, see Seth Stephens-Davidowitz, *The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data*, 118 J. PUB. ECON. 26, 26–28 (2014) (using Google data to estimate the prevalence and geographic variation of antiblack sentiment).

233. McLaughlin v. Florida, 379 U.S. 184, 192 (1964) (citing Hirabayashi v. United States, 320 U.S. 81, 100 (1943)). Notice that the Court's argument here is crucially ambiguous. It could be

sympathetically, the Court appears to be saying that because race is irrelevant to the vindication of legitimate government ends, the observation that the state is treating race nevertheless as salient has the effect of propagating a false popular belief in racial hierarchies.[234] The second possible interpretation of an anticlassification rule turns on a nonconsequentialist, deontological intuition. That is, according to some Justices, it is a moral axiom that the state must treat all persons as individuals, and such individualization precludes any taking account of their race.[235] This moral demand for individuation entails demanding judicial scrutiny for all racial classifications.

There are, to be sure, reasons for skepticism about these moral and theoretical premises of the anticlassification principle.[236] But even

that an individual's *race* is irrelevant to many legitimate state ends but that the persistence of *racism* as an ambient social phenomenon is relevant to how the state can achieve those ends. The Court's formulation elides this difference and therefore misses the possibility that racism may be salient to the state's means-end rationality, even if race per se is not.

234. Fletcher A. Blanchard, Christian S. Crandall, John C. Brigham & Leigh Ann Vaughn, *Condemning and Condoning Racism: A Social Context Approach to Interracial Settings*, 79 J. APPLIED PSYCHOL. 993, 993, 995–96 (1994) (demonstrating that cues from other people that racial discrimination is permissible or impermissible affect whether a person will condemn a racist remark, and that students hearing others condemn racism led to antiracist opinions while hearing others condone racism weakened antiracist opinions); Christian S. Crandall, Amy Eshleman & Laurie O'Brien, *Social Norms and the Expression and Suppression of Prejudice: The Struggle for Internalization*, 82 J. PERSONALITY & SOC. PSYCHOL. 359, 359 (2002) (finding that "[t]he public expression of prejudice toward 105 social groups was very highly correlated with social approval of that expression. Participants closely adhere to social norms when expressing prejudice, evaluating scenarios of discrimination, and reacting to hostile jokes"); Katie M. Duchscherer & John F. Dovidio, *When Memes Are Mean: Appraisals of and Objections to Stereotypic Memes*, 2 TRANSLATIONAL ISSUES PSYCHOL. SCI. 335, 341 (2016) (describing an online experiment involving memes about Asian stereotypes in which "seeing another person object to the meme increased the likelihood that White participants would object . . . but only when the race of the person was unstated, and not when the person was Asian").

235. Missouri v. Jenkins, 515 U.S. 70, 120–21 (1995) (Thomas, J., concurring) ("At the heart of this interpretation of the Equal Protection Clause lies the principle that the government must treat citizens as individuals, and not as members of racial, ethnic, or religious groups."). The same position is articulated with respect to gender in DAVID MILLER, PRINCIPLES OF SOCIAL JUSTICE 168–69 (1999) (arguing that to treat a woman on the basis of "information that relates to the whole group or class" to which she belongs is "to fail to treat her respectfully as an individual, and potentially to commit an injustice"). This argument does not rest on empirical evidence of the stigmatizing consequences of race-based or gender-based action. Rather, it applies whether or not the classified individuals perceive themselves as aggrieved.

236. For devastating critiques of the idea of colorblindness, see ELIZABETH ANDERSON, THE IMPERATIVE OF INTEGRATION 155–79 (2010) (describing the concept as "confused" and "[i]ncoherent"); Reva B. Siegel, *The Racial Rhetorics of Colorblind Constitutionalism: The Case of* Hopwood v. Texas, *in* RACE AND REPRESENTATION: AFFIRMATIVE ACTION 29 (Robert Post & Michael Rogin eds., 1998); Reva B. Siegel, *Discrimination in the Eyes of the Law: How "Color Blindness" Discourse Disrupts and Rationalizes Social Stratification*, 88 CALIF. L. REV. 77, 81–83 (2000). For an originalist critique of anticlassification rules as interpretations of the Equal

bracketing those hesitations, and taking those justifications at face value, there is still no reason to think that the logic of anticlassification strongly militates against the use of race either as a feature or as an element of a classifier by machine-learning tools. To the contrary, as a matter of either precedent or logic, equal protection law can accommodate racially sensitive algorithmic criminal justice.

Consider the first concern about the communicative effect of racial classifications. It is not clear that an algorithmic classifier is the sort of racial criterion that courts perceive to be objectionable. Rather, it is somewhat akin to the explicit use of race in criminal suspect identifications, which has to date elicited scant constitutional concern.[237] Suspect descriptions instead operate as given elements of the regulatory backdrop. Courts have not been wholly clear about why such suspect descriptions do not elicit careful scrutiny. One possible explanation is that judges believe suspect descriptions to be based on extrinsic facts, rather than airy suppositions about racial types, and so not the kind of generalizations that trigger anticlassificatory concerns. This logic might be extended to the algorithmic context. Race-based feature selections would then trigger no more constitutional concern than race-based suspect descriptions. The argument would be that a classifier based on training data is akin to a suspect description of a familiar sort, insofar as both are predicated on historical facts about crime.[238] Indeed, an advocate of algorithmic criminal justice might note that human observers are more likely than a machine to err in their deployment of race as a signal of criminality than an algorithm.[239] They

Protection Clause, in favor of a "duty-to-protect" view, see Christopher R. Green, *The Original Sense of the (Equal) Protection Clause: Pre-Enactment History*, 19 GEO. MASON U. C.R.L.J. 1, 3 (2008).

237. For a collection of cases, see R. Richard Banks, *Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse*, 48 UCLA L. REV. 1075, 1095–96 (2001). Note that this is not a function of the inclusion of other considerations. Classifications that include race as one among many elements can run afoul of the Equal Protection Clause. *See* Balkin & Siegel, *supra* note 229, at 16–17 (noting conflicting precedent on this point).

238. Are algorithms different because the historical data upon which they are based is not specifically linked to a particular crime? Consider the decision in *Brown v. City of Oneonta*, for example, which declined to impose constitutional tort liability when a description of a black male suspect provoked Oneonta police to stop more than two hundred "non-white persons," including women, encountered on the streets. Brown v. City of Oneonta, 235 F.3d 769, 779 (2d Cir. 2000) (Calabresi, J., dissenting from denial of rehearing en banc) (citation omitted). Although the stops were, in a trivial sense, based on a historical fact, the connection between that fact and the subsequent police actions was very strained. *Id*. The same might be said of algorithmic tools.

239. *See, e.g.*, Kleinberg et al., *Human Decisions*, *supra* note 136, at 277 tbl.VII (making precisely this argument).

might further contend that it is perverse to object to efforts to mitigate the effects of race on criminal justice outcomes through the substitution of machine for human judgments.

A second reason to think that an anticlassificatory logic does not work well in this domain would focus upon the absence of any communicative effect from algorithmic criminal justice. Many of the algorithms discussed in Part I are sheltered from disclosure by trade secrets law and hence are not disclosed presently to the public.[240] Even if they were to be disclosed in the course of litigation, it would likely be under the auspices of a protective order. To the extent that anticlassification rules rest on a concern about the communicative effects of state action, the use of an algorithmic tool that is wholly opaque should mitigate those concerns. More generally, the Supreme Court has been more accommodating of the conceded state use of race when it is somewhat obscured from public view.[241] A state actor that relies upon an algorithmic tool, but that muffles the precise content of that tool from the public through trade secrets law or otherwise, might mitigate the most powerful challenges on equal protection grounds. Stated more positively, the much-maligned algorithmic quality of "opacity" has the benefit of dampening troublesome communicative effects for racial classification. Advocacy of transparency has the perverse effect of courting the expressive harms that equal protection tries to minimize.

A related, if somewhat subtler, question arises if race is employed as a feature of the training data—that is, for each discrete observation (individual) in the training data, race is recorded—but race plays no role in the labels used to describe the classification task or in the tools used to identify an appropriate function. Does that approach have a constitutionally impermissible communicative effect?

Northpointe omitted race from the training data used for COMPAS.[242] But this appears to reflect corporate risk aversion, not an effort at legal compliance. Current law does not address whether the availability of race as an input into the deliberative process that results

---

240. *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 1349–53 (2018) (collecting examples and arguing for more transparency). More generally, algorithms used by commercial actors are also "secret." Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 10–11 (2014).

241. *See* Strauss, *supra* note 186, at 24 (noting "the Court's insistence on nontransparency" in affirmative action cases).

242. Angwin et al., *supra* note 16.

in state action violates the Equal Protection Clause on anticlassification grounds. To be sure, there is language in earlier precedent that suggests that any racial trace in official deliberation raises a constitutional problem.[243] But the weight of precedential evidence (as well as common sense) suggests that the mere fact that a decision-maker can *observe* the race of subjects does not mean that resulting action is therefore invalid. As a practical matter, many front-line state officials encounter suspects, defendants, and citizens and thereby directly perceive their interlocutors' race.[244] Similarly, the federal judiciary must—and indeed does—routinely recognize the race of litigants in order to reach judgments on statutory and constitutional discrimination claims, even when it is not strictly necessary.[245] Finally, recent affirmative action jurisprudence implies (without expressly stating) that the bare fact of racial awareness is not sufficient to state a constitutional violation. The University of Texas, whose admission policy was reviewed and upheld by the Court in 2016, considered race as part of its PAI, and this alone did not suffice to generate a constitutional problem.[246] In short, it seems quite plausible that an algorithmic criminal justice tool can use race as a feature in training data without triggering constitutional concern.

What of the argument against the state's use of racial classifications based on its putative obligation to treat individuals as individuals rather than as members of groups? The moral logic of individuation trains on "intentional uses" of racial classifications, not merely coincidental or happenstance entanglements with race.[247] That logic might seem to have traction here since algorithmic criminal justice entails a decision-maker relying on group membership rather than accounting for all relevant characteristics of an individual.

243. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007) ("The way to stop discrimination on the basis of race is to stop discriminating on the basis of race.").

244. *See* Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over* Brown, 117 HARV. L. REV. 1470, 1471 (2004) ("For a half-century now, the Constitution has prohibited state action that classifies on the basis of race, yet as Americans have debated the implications of that principle, few have thought it barred collecting racial data.").

245. Justin Driver, *Recognizing Race*, 112 COLUM. L. REV. 404, 408 (2012) (documenting courts' "unsettled and unsettling approach" to the recognition of litigants' racial identities).

246. *See* Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2206, 2212 (2016) ("[R]ace is given weight as a subfactor within the PAI.").

247. ANDERSON, *supra* note 236, at 155. Anderson is discussing "racial preferences" here, but her point applies to racial classifications too. *Id.*

But this is not quite right. In the absence of masking,[248] there is no human decision to assign costs or benefits on the basis of a racial classification with algorithmic criminal justice. And race is generally not going to be used as a substitute for more fine-grained traits. In any case, merely withholding race information does not ensure that an algorithm will not point toward race as a salient proxy. Machine-learning algorithms take training data (with or without a race parameter) and use them to generate a new classifier, which can then be applied to test data.[249] The fact that an algorithm is not initially supplied with an impermissible ground of decision as a feature of training data does not mean that it will not end up tracking that criterion in its classifier. Machine-learning tools are powerful and useful precisely because they can detect regularities in a data set that would not manifest in the absence of computational tools. Although machine-learning tools can be designed to be "private," in the sense of eschewing reliance on certain traits,[250] they can also "help to pinpoint reliable proxies" for traits even without information about the distribution of such traits in the population.[251] If race emerges as part of the classifier, this is not an intentional action in any meaningful sense—and yet it is still a classification on the basis of race.

Even if that happens, the official deploying the algorithm cannot be faulted for failing to engage in sufficient individuation: She supplies granular training data, selects among different computational tools, and then applies these tools to the specific facts about the individual being classified.[252] Even if the training data includes race information, the official has not designated race as a salient trait in any meaningful way. A decision-making process in which no human actor has elected to employ race as a criterion of action is not fairly characterized as an instance in which "the government distributes burdens or benefits *on the basis of* individual racial classifications."[253] The argument against

248. *See supra* notes 211–14 and accompanying text.

249. *See supra* notes 80–82 and accompanying text.

250. Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, 9 FOUND. & TRENDS IN THEORETICAL COMPUTER SCI. 211, 216–18 (2014) (developing a related concept of differential privacy).

251. Barocas & Selbst, *supra* note 51, at 692–93.

252. Kroll et al., *supra* note 39, at 682 (noting that in machine learning, "decision rules evolve on the fly—they are not specified directly, but are inferred from the data").

253. Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (emphasis added). What if the government actor designing the algorithm fails to prevent the algorithm from homing in on race? I read the doctrine not to problematize such culpable omissions.

algorithmic criminal justice from the moral demand for individuation, therefore, fails.

There is one final argument for the inapplicability of anticlassification logic here. Race is commonly thought to be already highly correlated with socioeconomic characteristics related to criminogenic and victimization distributions. It might hence be reasonably anticipated that many algorithmic tools designed to be predictive of criminality will, even absent any race feature in the training data, generate a function that either mimics, or is a good approximation of, racial distributions in the population. Given this, it is possible that "by remaining blind to sensitive attributes, a classification rule can select exactly the opposite of what is intended."[254] That is, the absence of a de facto predictive trait from the training data can generate systematic and serious errors in prediction.

A simple example from outside the machine-learning context illustrates this possibility. Imagine that wearing a particular baseball cap is used by police as a proxy for drug possession (say, because it may signal gang membership). Both blacks and whites wear this cap. For 100 percent of whites, and for zero percent of blacks, the cap is an accurate signal of drug possession. Let us say that police stop all those encountered wearing the cap, and this population is 75 percent white and 25 percent black. Because the cap generates a 75 percent success rate, its categorical (and colorblind) use might be deemed a meritorious criterion. But the efficacy of searches, and the avoidance of needless hassle for minorities, can be increased by limiting the instrument to white suspects.[255] Colorblindness here generates substantial and avoidable social costs. These can be corrected by simply accounting for race.[256]

In the machine-learning context, a fix entails the creation of a predictive tool that assigns individuals from different demographics to different classifications even though they exhibit the same behavioral traits.[257] Lest this seem obviously beyond the legal and moral pale, consider that one study of probation and parole decisions found that,

254. Kroll et al., *supra* note 39, at 686.

255. This example is drawn from Ian Ayres, *Outcome Tests of Racial Disparities in Police Practices*, 4 JUST. RES. & POL'Y 131, 139 (2002).

256. *See* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 918 (2017); Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, 3 INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214, 218–20 (2012) (demonstrating this result formally).

257. For a parallel result using the COMPAS data, see Nabi & Shpitser, *supra* note 20, at 8.

following the decision to omit race from a machine-learning algorithm, the accuracy of recidivism predictions declined "by about 7 percentage points."[258] The procedural purity demanded by an anticlassification rule, in sum, would come at a high price in terms of accuracy in algorithmic application.[259]

3. *The Lessons of Algorithmic Technology for Equal Protection Doctrine.* Current doctrinal approaches to constitutional racial equality arose after the Court had abandoned its early twentieth-century interpretation of the Equal Protection Clause as "a rationality test . . . invoked sporadically to strike down economic regulation."[260] These approaches were configured in the context of judicial efforts to dismantle educational segregation in the Jim Crow South and then during a political backlash to the Civil Rights Movement.[261] It was probably inevitable that the legal conception of racial discrimination as a matter of intention or classification would reflect judicial concern with the discretionary choices of the police officer, school board president, or state legislator—that is, the modal problems presented by mid-century civil rights law.

The institutional context of equal protection, however, has changed. Today, perhaps the sharpest and most controversial questions of racial justice are presented in the criminal justice domain. There, the emergence of algorithmic tools present questions poorly fitted to the doctrinal templates of intention and classification. This loose fit arises because the ways in which race filters into individual officials' discretionary criminal justice decisions are very different from the ways in which it can infuse algorithmic tools. Equal protection, as a result, poses questions that are simply not relevant to the operation of algorithmic criminal justice. It is a superseded legal technology, so far as algorithmic criminal justice goes. As more state power is channeled through algorithmic means, it will become increasingly obsolete.

---

258. Richard Berk, *The Role of Race in Forecasts of Violent Crime*, 1 RACE & SOC. PROBS. 231, 235 (2009).

259. Could such a use of race be justified as a narrowly tailored response to a compelling state interest? It is hard to say how much gain in accuracy would be required to make this claim compelling. For a discussion of how hard it is make this judgment, see *supra* notes 184–88 and accompanying text.

260. Michael Klarman, *An Interpretive History of Modern Equal Protection*, 90 MICH. L. REV. 213, 216 (1991).

261. *Id.* at 217–18; *see also* 3 BRUCE A. ACKERMAN, WE THE PEOPLE: THE CIVIL RIGHTS REVOLUTION 328–37 (2014) (considering the judicial forms of this backlash).

On the one hand, the manner in which algorithmic criminal justice unfolds generally means that are few opportunities for intentional discrimination of the familiar kind. The process of feature selection, to be sure, creates opportunities to use race as an input, to intentionally omit race in order to generate discriminatory patterns, or to choose an insufficient number of variables in ways that mimic the same effect.[262] But this sort of masking will be very hard to discover, much as prosecutorial or judicial animus is hard to identify now. It does not, at least on the basis of current evidence, appear to be a significant problem.

On the other hand, the logic of anticlassification might first seem to provide a firm foundation for regulating algorithmic criminal justice. But that logic turns out also to be a bad fit. The use of race in criminal justice algorithms is akin to the use of race in suspect descriptions. It lacks both the intentionality and the expressive spillovers that render nonindividuation troubling. Just as in the context of race-based suspect descriptions, moreover, it will sometimes be necessary to use race to achieve substantively accurate policy results.

In the dialogue between equal protection and algorithmic criminal justice, I suspect that constitutional law has much to learn and little to teach. A set of tools developed for a regulatory world of dispersed state actors, occasionally motivated by naked animus, cannot be mechanically translated into a world of centralized, computational decision-making. Even after law has made its contribution, therefore, the question of racial equity in algorithmic criminal justice remains open for debate—while the relevance and moral acuity of equality jurisprudence should be viewed as in serious doubt, absent more intensive rethinking.

III. RACIAL EQUITY IN ALGORITHMIC CRIMINAL JUSTICE BEYOND CONSTITUTIONAL LAW

The failure of constitutional law to provide a meaningful benchmark of racial equality is important in its own right. Yet it leaves the study of algorithmic criminal justice unmoored. It means there is no normatively attractive, empirically tractable way of evaluating the race effects of big-data predictive tools. This Part fills that gap. In order to do so, I will start by offering my own account of the normative stakes of racial equity in criminal justice to fill the vacuum left by our deficient

262. *See* Barocas & Selbst, *supra* note 51, at 692; Kroll et al., *supra* note 39, at 681.

constitutional doctrine. My view is that the reason for concern about racial equity in criminal justice generally is that our policing and adjudicative institutions play significant roles in the reproduction and entrenchment of social stratification. In a racially segmented society, when a person's life chances are defined importantly by their race, I believe this to be a moral wrong.

With that normative benchmark in hand, I turn to the extensive computer science literature on the question. That scholarship has developed a series of definitions of what is alternatively defined as algorithmic fairness or algorithmic discrimination. The literature has focused first on precise mathematical formulations of each definition and second on the generation of impossibility theorems—that is, formal proofs that it is not possible to maximize two or more parameters that in some fashion measure the racial effects of an algorithm. Because the computer science literature has been "silent on the choice" between different understandings of fairness,[263] mere specification of alternative conceptions of racial equity is not sufficient for any tractable conclusions about public policy. By applying my account of racial equity in criminal justice to these standards, I aim to make progress on determining which technical conception captures something of normative significance.

Two caveats are useful here. First, for the sake of clarity of exposition, I focus here on a binary between white and black defendants, even though this obscures the more complex racial dynamics of American policing today.[264] A focus on a black-white binary is warranted here as a way of clarifying the fundamental conceptual stakes. It is obviously inadequate as a general account of racial equity in policing, and I do not intend it as such. Moreover, I should emphasize again that my aim here is to offer not a judgment in respect to any specific algorithm but a more general analytic approach.

263. Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art* 29 (May 30, 2017) [hereinafter Berk et al., *Fairness in Criminal Justice*], https://arxiv.org/pdf/1703.09207.pdf [https://perma.cc/MK9Q-VVNG]; *see also* Sorelle A. Friedler, Carlos Scheidegger & Suresh Venkatasubramanian, *On the (Im)possibility of Fairness* 12 (Sept. 23, 2016), https://arxiv.org/pdf/1609.07236.pdf [https://perma.cc/EQ59-3V37] (asserting that "[c]hoice in mechanism [i.e., feature design] must thus be tied to an explicit choice in worldview"—in particular, a choice to prioritize either individual or group fairness).

264. *Cf.* Ramiro Martínez, Jr., *Incorporating Latinos and Immigrants into Policing Research*, 6 CRIMINOLOGY & PUB. POL'Y 57, 57 (2007) (documenting the "lack of research on Latino/as and Latino groups" in relation to the criminal justice system).

Much depends on the particular costs and benefits that in situ flow from a given instrument.

Second, a racial equity analysis of algorithmic criminal justice should not be a comparative one. It is not sufficient, that is, to point to a superseded technology that relies upon flawed human discretion and that already generates large racial effects as a justification for new, slightly less flawed technologies for allocating coercion. The mere fact that the status quo ante is characterized by racial injustice does not legitimatize proposals that preserve or extend some substantial part of that injustice. For example, no one thinks (or should think) the Jim Crow regime laudable merely because it followed slavery. Improvements in the status quo are a necessary but not sufficient condition for racial equity to be satisfied. It seems likely that the shift to algorithmic tools in criminal justice will be an enduring one. At the moment that a new policy is introduced, with potential path-dependent effects that will unfold over many iterations of policy making, it is especially important to understand the conditions under which that policy promotes racial equity: Far better, that is, to embed that principle at a policy's inception than to attend years of damage that cannot ever wholly be unraveled. Each technology ought to be evaluated on its merits and in light of its consequences.

A. The Stakes of Racial Equity in Contemporary American Criminal Justice

Why care about racial equity in criminal law? Without an answer to that question—and we have already seen that constitutional law doesn't provide a convincing one—no analysis of algorithmic criminal justice's racial equity effects gets off the ground. Accordingly, I start by offering my own evaluation of the racial stakes of criminal justice. But in doing so, I do not intend to break new ground here. I rather aim to clearly set forth a distinct normative position respecting racial equity in the criminal justice context.

American criminal justice implicates racial equity concerns because of their dynamic effects on racial stratification. Historical and contemporary empirical evidence suggests that both in the past and the present, criminal justice has been invoked in public discourse and applied in state practice so as to predictably exacerbate the subordinate status of African Americans in general. The dynamic (re)production of iniquitous social stratification—beyond the bare facts of animus and classification—is what should grip our collective conscience.[265]

At a very high level of abstraction, four causal mechanisms link criminal justice institutions to racial stratification. *First*, inherent black criminality has been invoked for more than a hundred years as public justification for more punitive interventions against African Americans and for the withholding of social services from them on moral desert grounds. *Second*, black communities have in practice been both overpoliced (in the sense of subjected to higher rates of coercive interventions) and also underprotected (in the sense of not receiving the same measure of protective legal resources that nonblack communities receive). As a result of this inefficient allocation of policing resources, state coercion has not resulted in lower levels of private coercion for African Americans. *Third*, pivotal actors within the criminal justice system, such as police, prosecutors, judges, and even public defenders, have tended to treat black suspects and defendants more harshly than white ones. Hence, the per capita cost of crime suppression has been greater for blacks than whites. *Fourth*, the spillover effects from disparate policing for black families and communities appear to be larger in magnitude than the spillover effects in white communities, even controlling for the extent of coercion. The net result of these mechanisms is that criminal justice imposes "compounding"[266] disadvantage upon African Americans as a group and works as a brake on individuals' efforts to rise in the social hierarchy. Even if not all African Americans are impeded by this headwind, enough are that we can meaningfully talk of persisting racial stratification to which criminal justice institutions have contributed. These diverse causal pathways underpin the need for careful attention to the manner in which formal criminal justice institutions can undermine the status of African Americans as a group.

265. Racial stratification is objectionable on (at least) two grounds. First, it embodies what Tim Scanlon calls a manifest "failure of equal concern" on the part of the state. T.M. SCANLON, WHY DOES INEQUALITY MATTER? 8 (2018). Second, stratification generates deadweight welfare losses in the form of unused human capital, psychological and social harms, and violence that flows from the latter. Of course, to the extent that such dynamic consequences have normative salience, it is because of a predicate obligation of equal concern toward the disadvantaged.

266. I draw this term from Deborah Hellman, *Indirect Discrimination and the Duty To Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 107 (Hugh Collins & Tarunabh Khaitan eds., 2018). Hellman's use of the term assumes an original act of discrimination; my use does not (although discriminatory acts are woven *across* the operation of criminal justice).

Rather than offering normative and empirical justifications for each element of this position—a task that would require a book rather than an article—I sketch some suggestive evidence for these causal linkages between criminal justice and racial stratification. I start with history, although I do not want to suggest that the state's obligations here rest on its historical responsibility for creating racial stratification in the first instance rather than its role in perpetuating that condition.

At the beginning of the twentieth century, national public discourse about "law and order became racialized, and conviction and incarceration rates for African Americans jumped disproportionately."[267] As the leading historical work by Khalil Gibran Muhammad vividly demonstrates, Progressive-era academics, journalists, and politicians in the North linked crime to African Americans at the same time as they downplayed white ethnic groups as sources of crime. By the early 1940s, Muhammad explains, "'Black' stood as the unmitigated signifier of deviation (and deviance) from the normative category of 'White.'"[268] Concomitant to this rhetorical shift, urban policing and carceral resources were disproportionately allocated to African Americans who were in the process of migrating up from the rural South. In northern cities in particular, police singled out blacks for intense surveillance and coercion.[269] This pushed up the rate of black incarceration and the proportion of the prison population that was black.[270] The black share of that population never subsequently dropped.[271] Racialized mass incarceration, that is, was at its inception a product of a moral panic stoked by northern elites in respect to the growing presence of an African American population that previously had been the South's "problem."

Today, racial disparities characterize both victimization rates and exposure to criminal justice coercion. Black men are more likely than white men to be victims of serious crimes (commonly called index crimes) such as murder.[272] They are also more likely to be arrested and

267. Jeffrey S. Adler, *Less Crime, More Punishment: Violence, Race, and Criminal Justice in Early Twentieth-Century America*, 102 J. AM. HIST. 34, 34 (2015).

268. MUHAMMAD, *supra* note 3, at 13.

269. Christopher Muller, *Northward Migration and the Rise of Racial Disparity in American Incarceration, 1880–1950*, 118 AM. J. SOC. 281, 310 (2012).

270. *Id.*

271. *Id.*

272. BRENDAN O'FLAHERTY, THE ECONOMICS OF RACE IN THE UNITED STATES 333–34 (2015).

incarcerated than white peers.[273] In many urban contexts, blacks and whites also experience widely varying chances of being stopped by police.[274] Moving from the policing to the adjudicative phase of the criminal justice process, common sentencing regimes impose disparate treatment on similarly situated offenders of different races by the use of different penalty structures for behavior closely associated with different racial groups.[275] As a result, one in eight black men in their twenties is in prison or jail on any given day, while some 69 percent of black high school dropouts are imprisoned over their lifetime, compared with just 15 percent of white high school dropouts.[276] For young black men, therefore, prison has thus become a predictable part of life's course.[277]

Note also that the intensive concentration of policing and incarceration resources along racial lines is not a rational, cost-justified response to crime. As I have argued elsewhere, there is evidence that some of the most common forms of policing black communities are inefficacious.[278] Black incarceration rates are also too high to be plausibly justified. One estimate suggests that reducing incarceration rates from 2009 to 1984 levels, and investing the resulting savings in an increased police presence, would lead to a net decline in violent crime nationally of about 130,000 incidents per annum.[279] Therefore, even if racial minorities benefit from the public safety produced by the criminal justice system, it is at a highly disproportionate and unnecessary direct cost.

Is part of this burden, though, justified by higher black crime rates? Even if we assume that "African Americans engage in

---

273. *Id.* at 335–36.

274. Huq, *Disparate Policing*, *supra* note 5, at 2411–12 (summarizing data from Chicago and New York).

275. *See, e.g.*, David A. Sklansky, *Cocaine, Race, and Equal Protection*, 47 STAN. L. REV. 1283, 1303 (1995) (describing the use of racially charged language in the enactment of narcotics statutes that impose different sentences on crack and powder cocaine offenses).

276. *See* Bruce Western & Christopher Wildeman, *The Black Family and Mass Incarceration*, 621 ANNALS AM. ACAD. POL. & SOC. SCI. 221, 231 (2009); Bruce Western & Christopher Muller, *Mass Incarceration, Macrosociology, and the Poor*, 647 ANNALS AM. ACAD. OF POL. & SOC. SCI. 166, 166–67 (2013).

277. For an extended account of these effects, see Kristin Henning, *Boys to Men: The Role of Policing in the Socialization of Black Boys*, *in* POLICING THE BLACK MAN: ARREST, PROSECUTION, AND IMPRISONMENT 57 (Angela J. Davis ed., 2017).

278. Huq, *Disparate Policing*, *supra* note 5, at 2429–40.

279. Philip J. Cook & Jens Ludwig, *Economical Crime Control*, *in* CONTROLLING CRIME: STRATEGIES AND TRADEOFFS 1, 32 n.20 (Philip J. Cook, Jens Ludwig & Justin McCrary eds., 2011).

significantly higher rates of street crime," there is evidence that conditions of "racial segregation and concentrated disadvantage"—environmental conditions that themselves are a function of non-race-neutral policies—explain much of the difference between different racial groups' crime rates.[280] That is, it is not so much that race is causally related to criminality but that African Americans are subject to forms of social and economic stratification and segmentation that conduce to criminality. Paradoxically, these underlying conditions are in an important respect a function of the federal government's decision to shift resources away from building human and social capital to policing crime. The intensification of policing and incarceration since the early 1970s, the historian Elizabeth Hinton has argued, was a conscious, and racially tinged, policy substitute for Great Society programs that could have mitigated those conditions.[281] That substitution could be reversed. As the sociologist Patrick Sharkey has demonstrated, it is precisely the local recreation of social services, and the concomitant creation of social capital, that has been a leading contributor to recent declines in crime. In one empirical study, Sharkey and his colleagues thus estimated 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."[282]

Finally, the direct costs of black incarceration are only part of the distinctive burden imposed by the current criminal justice system on racial minorities. Current crime suppression also imposes considerable collateral costs (or externalities) asymmetrically on racial minorities. To begin with, the immediate cost of encounters with police is racially asymmetric. The black experience of a police stop is reliably correlated with "stigma and stress responses and depressive symptoms"[283] because

280. Callie Harbin Burt, Ronald L. Simons & Frederick X. Gibbons, *Racial Discrimination, Ethnic-Racial Socialization, and Crime: A Micro-Sociological Model of Risk and Resilience*, 77 AM. SOC. REV. 648, 650–52 (2012); *see also* Lauren J. Krivo & Ruth D. Peterson, *The Structural Context of Homicide: Accounting for Racial Differences in Process*, 65 AM. SOC. REV. 547, 556 (2000) (examining marginal effects of social advantage of black and white communities).

281. Elizabeth Hinton, *"A War Within Our Own Boundaries": Lyndon Johnson's Great Society and the Rise of the Carceral State*, 102 J. AM. HIST. 100, 101–02 (2015).

282. Patrick Sharkey, Gerard Torrats-Espinosa & Delaram Takyar, *Community and the Crime Decline: The Causal Effect of Local Nonprofits on Violent Crime*, 82. AM. SOC. REV. 1214, 1234 (2017).

283. Amanda Geller, Jeffrey Fagan, Tom Tyler & Bruce G. Link, *Aggressive Policing and the Mental Health of Young Urban Men*, 104 AM. J. PUB. HEALTH 2321, 2321 (2014). For a powerful account of why these costs accrue distinctly to racial minorities, see Nicholas K. Peart, *Why Is the*

of the historically fraught nature of relations between American police and racial minorities. African Americans are, moreover, commonly subject to policing measures that are not generally employed against white citizens—such as pretextual vehicular stops—and are quite aware that they are objects of disparate treatment based on the presumption of black criminality.[284] They are also quite aware of the stigmatizing connection between race and criminality drawn since the beginning of the twentieth century. Even today, "demography-based suspicion is among the key social facts that define American life in the late twentieth and early twenty-first centuries."[285] Ethnographic studies paint a bleak picture: interactions between police and young black men are marked by distrust and fear, fomenting widespread alienation and disaffection.[286] Against the background of this broadly shared supposition of the relationship between criminality and race, public encounters with police can, even if warranted, humiliate and rob innocent racial minorities of the "ability to present themselves to other groups as the ordinary people they are."[287]

These effects generate further negative spillovers. As Randall Kennedy cogently observed three decades ago, African American men experience a "racial tax" from American criminal justice systems—even if they have no contact with it—because police and citizens are prone to perceive their race as a proxy for criminality and, hence, to configure them as potential criminals rather than potential victims.[288] Recent empirical work has confirmed Kennedy's account of the externalities of criminal justice for minority groups as a whole. African

N.Y.P.D. After Me?, N.Y. TIMES (Dec. 17, 2011), http://www.nytimes.com/2011/12/18/opinion/sunday/young-black-and-frisked-by-the-nypd.html [https://perma.cc/FB4R-7NN2].

284. CHARLES R. EPP, STEVEN MAYNARD-MOODY & DONALD HAIDER-MARKEL, PULLED OVER: HOW POLICE STOPS DEFINE RACE AND CITIZENSHIP 117–18 (2014).

285. WILLIAM J. STUNTZ, THE COLLAPSE OF AMERICAN CRIMINAL JUSTICE 22 (2011).

286. Rod K. Brunson, *"Police Don't Like Black People": African American Young Men's Accumulated Police Experiences*, 6 CRIMINOLOGY & PUB. POL'Y 71, 85 (2007) (finding that street stops that are perceived as unfair create perceptions of the police as unjust among young black men); Jacinta M. Gau & Rod K. Brunson, *Procedural Justice and Order Maintenance Policing: A Study of Inner-City Young Men's Perceptions of Police Legitimacy*, 27 JUST. Q. 255, 266 (2010) (noting that many respondents described contact with police "as demeaning and of inordinate frequency").

287. Paul Bou-Habib, *Racial Profiling and Background Injustice*, 15 J. ETHICS 33, 44 (2010). Bou-Habib addresses profiling, but I am extending his point.

288. KENNEDY, *supra* note 10, at 158–60. The criminal justice system thus creates "shared categories and classification systems through which individuals perceive and make sense of their environment." Michèle Lamont, Stefan Beljean & Matthew Clair, *What Is Missing? Cultural Processes and Causal Pathways to Inequality*, 12 SOCIO-ECON. REV. 573, 574 (2014).

American men hence continue to receive disfavored treatment in a wide array of economic and social contexts that limit important life opportunities.[289] The increased risk of contact with police, and thus incarceration, undermines the economic and social resources available to the larger racial cohort embedded in the same geographic community.[290] One in four black children also experiences parental incarceration—an experience that directly and negatively impacts their health and education outcomes.[291] Most notably, and dismayingly, black parental incarceration is associated with a 49 percent increase in infant mortality, an increase that has no parallel among white families affected by incarceration.[292] So not even children are spared. Rather, a concentration of policing and incarceration within black communities generates distinctive burdens with no parallel for majority racial groups—burdens that diffuse and concatenate across communities and generations. It is on this basis, I think, that it is plausible to characterize the contemporary American criminal justice system as "a systemic and institutional phenomenon that reproduces racial inequality and the presumption of black and brown criminality."[293]

This account of racial equity in criminal justice does not hinge on the presence of discriminatory animus at any specific point in policing or the adjudicative process. Of course, disparate racial treatment happens—probably quite often.[294] But this account of racial equity is forward looking and consequentialist insofar as it is trained on the ways in which systems reproduce practical socioeconomic stratification over time. Moreover, this account suggests that criminal justice institutions are not presently socially efficient. Their footprint could be diminished

289. For an effective summary of the relevant data, see Devah Pager & Hana Shepherd, *The Sociology of Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181 (2008); *see also* DEVAH PAGER, MARKED: RACE, CRIME, AND FINDING WORK IN AN ERA OF MASS INCARCERATION 93–96 (2007) (reporting effects of racialized assumptions of criminality on employment opportunities).

290. *See* Todd R. Clear, *The Effects of High Imprisonment Rates on Communities*, 37 CRIME & JUST. 97, 115–16 (2008) (discussing this effect); *see also* Amy E. Lerman & Vesla M. Weaver, *Staying Out of Sight? Concentrated Policing and Local Political Action*, 651 ANNALS AM. ACAD. POL. & SOC. SCI. 202, 204 (2014) (finding in a study of New York that "witnessing stops that occur with little justification and that feature physical force can make people feel occupied and powerless, and can incentivize disengagement with government").

291. WAKEFIELD & WILDEMAN, *supra* note 8, at 41, 146 (noting that parental incarceration is a mechanism for the intergenerational transmission of inequality).

292. *Id.* at 108.

293. Naomi Murakawa & Katherine Beckett, *The Penology of Racial Innocence: The Erasure of Racism in the Study and Practice of Punishment*, 44 LAW & SOC'Y REV. 695, 701 (2010).

294. *See generally supra* notes 4–6.

in ways that do not create social costs from more crime. At present, however, the inefficiently large costs of criminal justice (which are not justified by sufficient offsetting social benefits) fall disproportionately on racial minorities. Many reforms that increase social efficiency will also further racial equity as a result.

A possible counterargument is that a particular quantum of state coercion will, *ceteris paribus*, be more costly for a member of a white majority than a black minority. That is, whites' greater wealth and more remunerative employment outcomes mean that their economic losses from even transient coercion or incapacitation are likely to be greater than those of African Americans.[295] I am skeptical. I find it troubling to use racial stratification by wealth and income as a lever to discount the costs imposed on African Americans. I also do not accept that the implicit metric at work in this analysis (in effect, the capacity to pay) tracks a normatively attractive species of welfare. Finally, I have already flagged negative externalities to African Americans as a group, and to communities and families, that simply have no parallel for racial majorities. I think it is more likely that black communities and families will want for the social and financial buffers that mitigate the shock of criminal justice contacts. Hence, I think this counterargument is both empirically and normatively flawed.

B. A Racial Equity Principle for (Algorithmic) Criminal Justice

The algorithmic tools described in Part I are new mechanisms to allocate coercion within the criminal justice system. But the introduction of new computational and epistemic technologies does not alter the basic stakes of racial equity. They should be evaluated, that is, as elements of that overall system.

In this light, the key question for racial equity is whether the costs that an algorithmically driven policy imposes upon a minority group outweigh the benefits accruing to that group. If an algorithmic tool generates public security by imposing greater costs (net of benefits) *for blacks as a group*, it raises a racial equity concern. That policy undermines racial equity by deepening the causal effect of the criminal justice system on race-based social stratification.[296] This test is

295. Cedric Herring & Loren Henderson, *Wealth Inequality in Black and White: Cultural and Structural Sources of the Racial Wealth Gap*, 8 RACE & SOC. PROBS. 4, 4–5 (2016).

296. This standard is analytically distinct from disparate impact as conventionally understood, not least because it does not account for the benefits of a policy for those beyond the burdened group. It is an interesting question whether disparate impact, especially as applied to state action,

consequentialist. It focuses on the effects of an algorithm's use.[297] It is also holistic. Unlike older risk assessment tools, it accounts for both the benefits and the costs of intervention. And, to emphasize again, it is quite general: There is no reason not to apply it to criminal justice more generally. I develop the test here nevertheless because I am concerned with algorithmic tools that can develop precise cut-points for using coercion based on analyses of large volumes data.

This standard has a distant kinship to John Rawls's difference principle, which holds that "[a]ll differences in wealth and income . . . should work for the good of the least favored."[298] But the principle offered here operates within a much narrower institutional bore (criminal justice alone) and is justified on much more specific grounds—to ensure that institutions purportedly operating in furtherance of public safety are not doing so in a fashion that exacerbates differences in racial strata.

What, though, of animus? Of course, individual officials do act at times with an invidious state of mind.[299] At present, the institutional process of adjudication and the doctrines structuring inquiries into bad intent ensure that few such instances are ever brought to light, let alone used as a basis for constitutional relief.[300] I am skeptical that the resulting harms are of the same magnitude as the damage that comes from criminal justice's effect on racial stratification. Even if equal protection doctrine were more effective at identifying instances of bad motivation, a criminal justice system purged of animus would still have substantial ramifications for racial stratification. It is the existence of racial stratification, in any case, and the channeling of anxieties about security and difference into racialized forms, that plausibly drive much animus in the first instance. Addressing stratification, in my view, is a more enduring and effective means of regulating animus than the

---

might be reconfigured to approach the standard suggested in the text.

297. Note that it is possible to take the view that there is a nonconsequentialist obligation on the state's part to show equal regard for all its citizens, and to think that my consequentialist metric is a way of honoring that obligation.

298. JOHN RAWLS, *Distributive Justice: Some Addenda*, *in* COLLECTED PAPERS 163 (Samuel Freeman ed., 1999). Rawls formulated the difference principle in a number of different ways. Nothing here rests on those variations, so I ignore them.

299. For evidence of that effect, see CHARLES R. EPP, STEVEN MAYNARD-MOODY & DONALD HAIDER-MARKEL, PULLED OVER: HOW POLICE STOPS DEFINE RACE AND CITIZENSHIP 117–18 (2014).

300. For an extended argument to this effect, see Huq, *Discriminatory Intent*, *supra* note 41, at 21–36.

emaciated and enfeebled investigative doctrinal instruments the Court employs.[301]

There are two ways of analyzing the relevant costs and benefits of an algorithmically allocated coercive measure. The first is to focus solely on the immediate costs and benefits of a coercive intervention and to ignore externalities. As a rough cut, this seems a plausible approach with serious crimes, where externalities are dwarfed by immediate costs and benefits. An alternative approach accounts for both immediate costs and also externalities for different groups. The latter take many forms, including the effect of high incarceration rates on black communities and children as well as the social signification of race as a marker of criminality. But as I argued above, the evidence suggests that these impacts are felt principally by members of racial minorities. It is, moreover, plausible to hypothesize that these spillover costs will largely be experienced by members of the same racial group as the suspect, given persisting patterns of racial residential segregation.[302] Hence, the spillover costs of coercion of minority individuals for the minority group will be greater on a per capita basis than the costs of coercing majority group members. If the costs of coercing minorities are larger while benefits remain static, racial justice will be satisfied by an algorithmic tool that imposes a higher threshold for black suspects than for white suspects. For less serious crimes—again, defined very roughly—these spillover effects may be similar in magnitude to the direct benefits and costs of coercion. Hence, a simplified analysis that ignores spillovers would be inappropriate. Rather, a bifurcated rule with different thresholds for whites and blacks may be necessary to ensure that minority coercion does not exacerbate racial stratification for less serious offenses.

Under either of these approaches, it will often be the case that racial equity and social efficiency (in the sense of ensuring that immediate social benefits exceed immediate social costs) will align. For example, when a majority group does not benefit from a policy, or when its net gain is less than the costs imposed on the minority group—and the latter suffers a net loss—that policy is socially inefficient. Equity and efficiency therefore align.

301. *Id.*

302. Matthew Hall, Kyle Crowder & Amy Spring, *Neighborhood Foreclosures, Racial/Ethnic Transitions, and Residential Segregation*, 80 AM. SOC. REV. 526, 527 (2015) ("[T]he modal experience for blacks (and Hispanics) in U.S. cities is high residential segregation.").

This approach makes certain simplifying assumptions that I believe to be plausible. It assumes that most crime is intraracial, such that costs and benefits do not cross the color line by and large. Obviously, this is not always true. But it does hold as a general matter.[303] Moreover, my analysis assumes away a number of unusual circumstances in which racial equity and social efficiency come apart. Because these circumstances are rare, I do not dwell on them. I mention two here briefly. First, it is possible that a policy benefits both the minority and the majority group, but the former benefit less than the latter. As a result of this gap, the extent of racial stratification increases even as the minority is benefited. The evaluation of such a policy would turn, in my view, on the magnitude of social gain and the extent to which the policy generates stratification. I do not think a general conclusion is appropriate to reject such policies. Rather, I believe the best approach would be not to discontinue the policy but to consider offsetting policies that mitigate its stratifying effect.

Second, net gains from a policy for a majority group may exceed the net cost imposed on a minority group. Imagine, for example, a national security policy that generates significant benefits by imposing crushing burdens on a very small ethnic or religious minority. In this case, there is a tension between efficiency and antidiscrimination. Such conflicts have generated disagreement among scholars.[304] In the crime-control context, I suspect that this will rarely occur given the intragroup nature of much crime. Yet my own view is that gains in net social welfare should generally not be obtained by imposing burdens on minority groups subject to wider dynamics of compounding subordination.[305] In effect, such a policy would yield a regressive wealth

303. *See* Robert M. O'Brien, *The Interracial Nature of Violent Crimes: A Reexamination*, 92 AM. J. SOC. 817, 818–19 (1987) (finding evidence that crime is more intraracial than would be anticipated).

304. *Compare* Louis Kaplow & Steven Shavell, *Should Legal Rules Favor the Poor? Clarifying the Role of Legal Rules and the Income Tax in Redistributing Income*, 29 J. LEGAL STUD. 821, 821–35 (2000) (favoring welfare maximization), *with* Lee Anne Fennell & Richard H. McAdams, *The Distributive Deficit in Law and Economics*, 100 MINN. L. REV. 1051, 1129, 1170 (2016) (doubting this maxim).

305. This view implies that consequences are morally salient but that welfare maximization is not the only measure of such consequences. The basic arrangements of a society are also important and sometimes merit protection or improvement even at the cost of net social welfare. For a different view that turns solely on purpose, and seems unconcerned with consequences, see Jed Rubenfeld, *Affirmative Action*, 107 YALE L.J. 427, 440–41 (1997) ("A law whose express purpose is racial apartheid or expulsion is unconstitutional per se, because racial purification of society is an objective that no legislature can pursue under the Fourteenth Amendment—period.").

transfer from blacks to whites in which the former pay for the security enjoyed by the latter.[306] I would hence prioritize the distribution that resulted from a policy over the sheer quantity of social welfare it yielded, at least in the absence of catastrophic general welfare losses from forbearance. I do not perceive any circumstances in which that latter exception plausibly applies.[307]

C. Benchmarks for Algorithmic Discrimination

A large computer science literature on algorithmic design has generated a plethora of definitions of "algorithmic fairness" and "algorithmic discrimination." One count finds twenty-one definitions.[308] Not all are relevant in the criminal justice context, and not every concept is analytically distinct from all others. My aim in this section is to home in upon a relevant subset of such definitions and to develop a quadripartite taxonomy of potential metrics for gauging racial equity. Stated otherwise, what follows is a synthesis and simplification of a much larger technical literature—a synthesis written with the aim of practical application in mind.

I begin by sketching the four most salient metrics in the literature.[309] These can be summarized as follows: One might *first*

306. *Cf.* Tal Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map To Examine Efficiency and Fairness in Automated and Opaque Decision Making*, 4 SCI., TECH., & HUM. VALUES 118, 123 (2016) (noting the possibility that an algorithm can "enable[] transfers that systematically harm minorities and other protected groups").

307. Minority politicians and police chiefs have at times believed that a disproportionate policing focus on African Americans was warranted in terms of community self-preservation—a belief that the "cumulative impact" of harsh antinarcotics measures has over time shown to be erroneous. JAMES FORMAN, JR., LOCKING UP OUR OWN: CRIME AND PUNISHMENT IN BLACK AMERICA 124–48, 21 (2017) (documenting these calls). If they had been correct—and Forman persuasively suggests that they were wrong on the facts—then this would have justified a less demanding risk threshold for blacks than for whites. *See id.*

308. *See* Arvind Narayan, *Tutorial: 21 Fairness Definitions and Their Politics*, YOUTUBE (Mar. 1, 2018), https://www.youtube.com/watch?v=jIXIuYdnyyk [https://perma.cc/HE3C-GXDU].

309. There are different enumerations of competing definitions of algorithmic fairness in criminal justice in particular. Richard Berk and his co-authors identify six different definitions. Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 13. They do not include one of the definitions I consider. Another paper by Sam Corbett-Davies and colleagues (including me) identifies three definitions that are salient to criminal justice policy. Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, *in* PROC. OF THE 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 798 (2017). In addition, Feldman et al. define fairness as the inability to predict a trait from the execution of an algorithmic function. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and Removing Disparate Impact*, in PROC. OF THE 21ST ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY &

simply look at whether equal fractions of each racial group are labeled as risky, such that they will be subject to additional policing or detention. A similar, although not identical, analysis where risk is measured as a continuous variable without a threshold for coercive action would look for equal average risk scores across different racial groups. *Second*, one might ask whether the same classification rule is being used to assign racial groups to the high-risk category. This condition is satisfied if the same numerical risk score is used as a cutoff for all groups. *Third*, one might separate each racial group and then look at the rate of false positives conditional on being categorized as high risk. And *fourth*, one might separate each racial group and ask how frequently false positives are conditional on being in fact a nonrisky person. In the literature, this has been characterized as defining the population of those within a racial group who in fact will not engage in subsequent criminal conduct and identifying what proportion of that subset were erroneously categorized as warranting coercion.

The four concepts of fairness or nondiscrimination are summarized in Table 1, which pairs each conception to the relevant parameter (or variable) that is to be equalized.

Table 1: Conceptions of Nondiscrimination in Algorithmic Criminal Justice

Conception of Fairness	*Parameter that should be equalized*
Statistical parity	*Proportion of each group subject to coercion*
Single threshold	*Treatment of equally risky persons within each group*
Equally precise coercion	*Proportion of those ranked as risky who are erroneously classified*
Predictive error equality	*Proportion of nonrecidivist persons that are subject to coercion*

DATA MINING 265 (2015).

Table 1 is intended to capture the range of core conceptions of nondiscrimination that should matter in the criminal justice context. It does not, as I have already noted, capture the *full* range of potential conceptions of algorithmic fairness. For instance, one recent survey additionally flags the idea of treatment equality,[310] which looks simply at the ratio of false positives to false negatives for a given racial group. To date, however, the latter concept has not played a large role in debates about racial equity. My analysis does not suggest that it should. Hence, I leave it to one side for present purposes.

Figure 1 below helps clarify these four concepts. It displays the risk ranking assigned by an algorithm—represented as a continuous variable of two groups, white and black. The x-axis represents the risk value assigned to members of the population; the y-axis represents the frequency with which members of the group are assigned to a risk level. For the purposes of this analysis, I assume that the training data used to generate the risk assessments is not flawed. I also assume that it is not biased in ways that result in whites or blacks being subject to disproportionate coercion. I make this assumption so as to enable a narrow focus on the question whether the algorithmic classification rule *standing on its own* presents a question of racial justice.

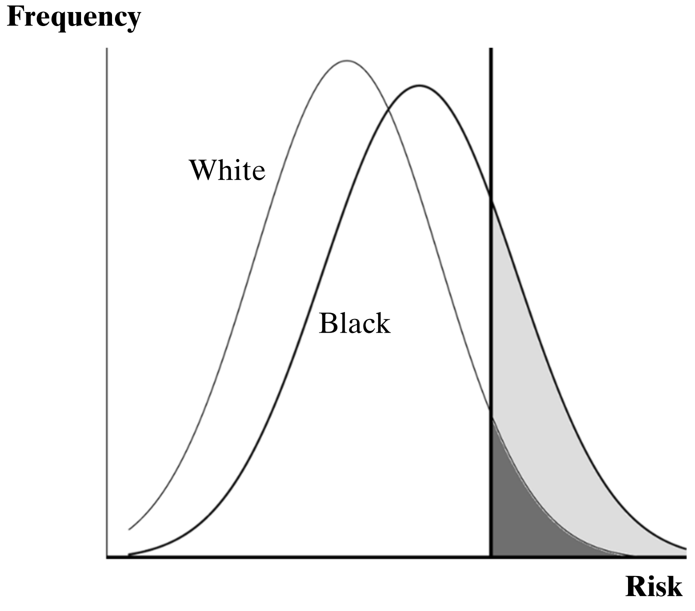
The graphic contains a vertical line to represent the cutoff point for the purposes of allocating coercion. Those who fall to the right of this threshold are subject to the coercive treatment (either a police stop or a detention-related intervention), while those who are to the left of the threshold are not subject to any coercion. The parts of the curve that represent populations that will be coerced (assuming the algorithm's recommendations are followed) are represented with shaded blocks in the graphic. The proportion of the white and the black populations subject to coercion is a function of the area under the respective curve to the right of the threshold.

This form of graphical representation has a number of advantages. It captures the way in which a threshold will distinguish between populations that are themselves quite internally varied in terms of their riskiness. It also reflects some key features of criminal justice algorithms in practice. In particular, it captures the fact that a decision must be made about who the *marginal* person on the risk curve is who should be detained. It also captures the intuition that the risk curves for different racial groups might diverge.[311]

310. Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 14.

311. An alternative used in the literature is a confusion table, which is a two-by-two matrix

Figure 1: Hypothetical risk distributions for white and black populations



that distributes individuals in terms of whether they ultimately committed acts justifying coercion and whether they were in fact coerced. *See, e.g.*, Tom Fawcett, *An Introduction to ROC Analysis,* 27 PATTERN RECOGNITION LETTERS 861, 862 (2006) (describing the use of confusion matrices). Confusion tables, however, do not capture all the information that an algorithm generates, such as the variance in risk values—and rely on knowledge that a decision-maker by construction does not know at the time the relevant decision has to be made, that is, whether a suspect or a defendant in fact will go on to commit a crime or impose a harm on others in the future. Confusion tables hence omit useful information while including information that cannot plausibly inform the decision whether to coerce or not. They are not good instruments for exploring algorithmic fairness, which is a standard that has to be applied at the moment the algorithm is used—not later, once new information about potential states of the world has become available.

Moreover, confusion tables fail to distinguish the average subject of coercion from the marginal subject of coercion. For example, imagine a single decision rule (say, a risk threshold of 10 percent) is applied to both a white and a black population. The white population comprises some with a 1 percent chance of carrying contraband and some with a 75 percent chance. The black population comprises some with a 1 percent chance, and some with a 50 percent chance. A confusion table draws attention to the fact that the proportion of stops that are false positives for the white group will be one-half that for the black group (i.e., 25 percent rather than 50 percent), but the table will not elucidate whether this is a function of (a) a biased decision rule or (b) a neutral and justified decision rule being applied to different distributions in the population. *See* Camelia Simoiu, Sam Corbett-Davies & Sharad Goel, *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS APPLIED STAT. 1193, 1194 (2017) (setting out this example); *see also* Ayres, *supra* note 255, at 131 (discussing the "strengths and weaknesses of using 'outcome tests' to assess racial disparities in police practices"). This confusion, ironically, is avoided by foregoing the use of confusion tables.

In Figure 1, the tails of the curve for the black population are to the right of those for the white population. This means that the algorithm tends to assign general higher risk values to black persons than white persons. If the risk distributions of both populations are equal, no interesting question of racial equity or discrimination would arise: White and black outcomes would not be distinct. This element of the hypothetical is not meant to imply that blacks in fact are more likely to commit crimes than whites. It is rather to present a situation that is plausible and that defines most sharply the questions of racial equity of interest here.

The four conceptions of algorithmic fairness or algorithmic nondiscrimination can be elaborated as follows. First, an algorithmic classifier might exhibit *statistical parity*. This means that an equal proportion of members of each group are subject to coercion. In terms of the graphic, this means that the shaded areas under the white and the black curves to the right of the threshold are equal to each other.[312] This can happen, it is worth noting, even if there is wide variation in the ratio of false positives to true positives for whites and for blacks. Where there is no threshold, one might instead use the average risk score for a given group. A variant on statistical parity is "conditional statistical parity," which requires that, having controlled for a "limited set of 'legitimate' risk factors, an equal proportion of defendants within each race group" are treated as risky.[313] In practice, however, this definition is highly sensitive to what counts as a "legitimate" risk factor. Because my analysis does not assume an answer to the question of what counts as a legitimate risk factor, I put aside here the possibility of conditional statistical parity.

Statistical parity is a clear and simple idea. Indeed, it is employed as part of the prima facie case in disparate impact analysis in employment discrimination law.[314] Under longstanding administrative agency construction, a racial difference in selection rates of "less than four-fifths" is "generally" taken as evidence of "adverse impact."[315] On

312. Corbett-Davies et al., *supra* note 309, at 798; *see also* Dwork et al., *supra* note 256, at 218 (defining statistical parity in terms of the fact that "an individual observed a particular outcome provides no information as to whether the individual is a member of *S* or a member of *T*").

313. Corbett-Davies et al., *supra* note 309, at 798.

314. *See, e.g.*, Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 308 n.14 (1977); Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 339 (1977).

315. Federal guidelines state:

> 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

the other hand, there is no *a priori* reason why state coercion should be equally distributed among racial groups. To be sure, there is some evidence that at least for certain sorts of offenses, such as narcotics crimes, there are "no statistically significant differences" in offending rates for different racial and ethnic groups.[316] But on the assumption that the algorithm's training data are not flawed, the hypothetical would simply not capture such cases.

Second, an algorithmic classifier might be viewed as fair if it treated two people who evinced the same ex ante evidence of risk, but differed by race, in the same way. The computer science literature has distinguished between a single threshold and "multiple race-specific thresholds."[317] A recent paper further offers a formal proof to the effect that the "immediate utility" of a decision rule—defined in terms of the immediate benefits of crime directly suppressed and direct costs of coercion (and ignoring externalities)—is typically optimized by maintaining a single threshold rule for coercion rather than having plural thresholds.[318] That is, a social planner with an algorithmic tool that is trained on unbiased data would select a single risk threshold for both whites and blacks if she wished to optimize over the costs and benefits of crime control. This analysis of social welfare, however, does not answer the question of what necessarily furthers racial equity under all conditions. In particular, it is important to observe that the formal proof of optimality is limited to the immediate effects of an algorithmic tool. Racial stratification is plausibly understood to be a compounding effect of the latter concept rather than something captured by the former.

This conception of fairness in algorithmic criminal justice has not so far attracted a distinctive label. Indeed, some accounts of discrimination in the algorithmic context simply do not cite this kind of fairness, preferring to focus on the relative frequency of false (or true) positives (or negatives) in the two racial groups.[319] In other work, this

> than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

29 C.F.R. § 1607.4(D) (2016); *see also* Ricci v. DeStefano, 557 U.S. 557, 586–87 (2009) (endorsing this four-fifths rule).

316. U.S. DEP'T OF HEALTH AND HUMAN SERVS., RESULTS FROM THE 2013 NATIONAL SURVEY ON DRUG USE AND HEALTH: SUMMARY OF NATIONAL FINDINGS 26 (2013), http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf [https://perma.cc/J46Z-CQ4Y].

317. Corbett-Davies et al., *supra* note 309, at 797.

318. *Id.* at 799–802.

319. *See, e.g.*, Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 13–15 (failing to

conception has been characterized simply as "fairness,"[320] but that nomenclature is too vague to be helpful. I label this definition, therefore, the *single threshold* definition of algorithmic fairness. Graphically, the single threshold definition of fairness is represented by the fact that the vertical line that marks the threshold between coercion and its absence is in the same place for both racial groups. If the vertical thresholds were placed in different locations on the x-axis, there would be a group of individuals between the two thresholds who would present the same evaluated risk but would be treated differently solely on account of their race.

A third conception of algorithmic nondiscrimination examines only the portion of the population that lies to the right of the risk threshold. In Figure 1, this comprises the shaded areas under the curves. These encompass parts of the white and black populations subject to coercion as a consequence of the algorithm's recommendations. Not all of these recommendations, however, will be borne out by future events. In the bail context, for example, some fraction of those subject to state coercion would not have gone on to commit crimes that justified pretrial detention. They will, in other words, be false positives. One way of thinking about nondiscrimination is in terms of the false positive error rate conditional on being assigned state coercion by the algorithm—which can also be stated as $P(nonrecidivist|high\ risk)$. So if a greater fraction of blacks stopped or detained turn out to be innocent in the relevant sense than the same fraction of nonrecidivist whites, then this would violate the third conception of fairness. Or, stated in yet another form, if the proportion of those false positives under the black curve to the right of the risk threshold is greater than the proportion of false positives under the white curve to the right of the threshold, then this conception of equality is violated.[321] This notion is captured by a number of different terms in the computer science literature. A leading group of analysts label it "conditional use accuracy."[322] In my view, it is simplest to label it *equally precise coercion* because this conception is centrally

mention this kind of fairness in a sixfold taxonomy).

320. *See* Dwork et al., *supra* note 256, at 215.

321. This conception is focused not on the absolute number of false positives but rather on the percentage of those subject to coercion within a racial group that would not have gone on to engage in socially undesirable behavior. It would be perverse to define fairness in terms of a parameter that is driven primarily by the relative size of the two groups under study.

322. Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 14.

concerned with the rate at which false positives occur conditional on the fact of being coerced.[323]

Equally precise coercion played a role in the debate over the COMPAS algorithm.[324] Responding to ProPublica's allegations of racial disparity, Northpointe focused on the fact that the rate of error among the black and white groups subject to coercion was the same.[325] In effect, the Northpointe argument was that so long as equally precise coercion obtained, there was no discrimination problem.

The fourth and final conception of fairness in the algorithmic context also focuses on false positives, but from a different angle. Rather than the subset subject to coercion, it focuses on the subset that would not go on to commit a crime or violent act. This subset of nonrecidivating persons is used as a denominator. For a numerator, it asks what fraction of that subpopulation is incorrectly subject to coercion. In the bail context, for example, this means asking whether "among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across race groups."[326] In other words, conditional on being a nonrecidivist (in whatever sense of that term is relevant), the rate of erroneous false positives across racial groups does not vary—or *P(high risk|nonrecidivist)*. This conception of equality is not easy to capture using Figure 1, since the baseline category of nonrecidivists are dispersed on both sides of the risk thresholds. In effect, it comprises a diffuse subset of whites and blacks who in fact would not commit actions that justify coercion. This conception of fairness requires that we look for the proportion of that nonrecidivist subset to the right of the risk threshold. If one racial group's ratio is larger than the other's, there is reason for concern under this theory.

This conception has attracted a wide variety of labels, including "predictive equality,"[327] "conditional procedure accuracy,"[328] and "equalized odds."[329] Another group of analysts use the label "balance

323. "Precision" is the term used by machine-learning specialists, who perceive the term "accurate" to imply a normative judgment. I am grateful to Sharad Goel at Stanford School of Engineering for discussion of this point.

324. *See supra* notes 14–22 and accompanying text.

325. DIETERICH ET AL., *supra* note 20, at 3.

326. Corbett-Davies et al., *supra* note 309, at 798.

327. *Id.*

328. Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 13–14.

329. Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, *in* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 2 (2016),

for the positive class" for a related concept.[330] Their paper also mentions the concept of "balance for the negative class" to capture the symmetrical idea that "the assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other."[331] Deviating from my own past usage,[332] I will use the label *predictive error equality* here to capture the idea that what is at stake in this fourth definition of nondiscrimination is the notion that the burden placed on the nonrecidivist subset of each racial group should be the same. Predictive error equality is the focus of the ProPublica critique of the COMPAS algorithm: The journalistic organization demonstrated that the proportion of nonrecidivist black defendants recommended for detention by the COMPAS algorithm was substantially higher than the proportion of nonrecidivist white defendants subject to the same recommendation.[333] In effect, ProPublica implicitly leveraged the intuition that what matters with an algorithm is what happens to the nonrecidivist subset. If the treatment of nonrecidivists varies across racial groups, ProPublica's argument went, an algorithm could not be ranked as nondiscriminatory.

D. *Prioritizing Conceptions of Algorithmic Discrimination*

The range of possible ways to operationalize the quality of nondiscrimination in the algorithmic criminal justice context raises the question of how to evaluate and rank the four main competing conceptions. My aim in this section is twofold. First, I point to results in the technical literature that demonstrate the impossibility of pursuing all these conceptions of nondiscrimination simultaneously. Second, I offer my own normative account of which conception to prioritize. This account, detailed above, hinges on the minimization of costs net of benefits for the minority group. Contrary to both Northpointe and ProPublica, this contends that rates of false positives (whatever denominator is used) are not compelling normative

---

http://papers.nips.cc/paper/6373-equality-of-opportunity-in-supervised-learning [https://perma.cc/9N8L-FTBD].

330. Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores* 4, 10 (Nov. 17, 2016), https://arxiv.org/pdf/1609.05807.pdf [https://perma.cc/A4K2-DL2H] (labeling this concept "[c]alibration within groups").

331. *Id.* at 4.

332. Corbett-Davies et al., *supra* note 309, at 798.

333. *See* Angwin et al., *supra* note 16 (finding that "[t]he formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants").

benchmarks. Instead, the analysis should focus on whether a minority risk threshold yields net costs or benefits for that group. Where there are no spillovers, it is likely that the same threshold will obtain for both minority and majority groups. Where there are large and asymmetric spillovers, both social efficiency and racial equity are served by different thresholds.

1. *Conflicts Between Algorithmic Fairness Definitons*. It would seem desirable to satisfy all these definitions of equality. At least at first blush, all capture colorable and important intuitions about the fair allocation of coercion. But matters are not so simple. It turns out that this is not possible in many cases—and not possible under conditions that are reasonably likely to occur in practice—for two reasons.

First, it will generally be the case that statistical parity cannot be achieved using a single threshold. This is readily apparent from Figure 1, which illustrates the case in which the risk distributions of racial groups vary. When this happens, it will always be the case that a single risk threshold will subject different proportions of each group to coercion. Hence, it is not possible—assuming differences in the distributions of risk between the two racial populations—to have both a single threshold and also statistical parity.

Second, it is often also impossible to achieve both equally precise coercion and predictive error equality. This impossibility result holds under two conditions. First, base rates of criminality are different for the two racial groups. Second, there is no function that allows for "perfectly accurate classification" (a condition also known as "separation").[334] Under these conditions, one cannot have both equality in conditional use accuracy and equality in the false negative and false positive rates, where the latter term is simply conditional procedural accuracy.[335] It is for this reason that assessments of the COMPAS algorithm have diverged. On the one hand, the original criticism of the algorithm focused on the difference in the rate of

334. Berk et al., *supra* note 263, at 18–19. For derivations of the same result, see Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153, 157 (2017); Kleinberg et al., *Inherent Trade-Offs*, *supra* note 330, at 5–6. Under certain conditions, it is not possible to equalize conditional procedural accuracy between groups without establishing different thresholds for black and white classifications. Corbett-Davies et al., *supra* note 309, at 802–03.

335. Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 14; Kleinberg et al., *Inherent Trade-Offs*, *supra* note 330, at 5.

conditional procedural errors for blacks and whites.[336] On the other hand, the defenses of Northpointe's instrument focused on the fact that it was calibrated within the categories of risk—that is, the conditional use error rate was equal for both whites and blacks.[337] Neither side recognized that given the possibility of underlying differences in the empirical characteristics of racial groups, and absent separation, these two metrics of algorithmic fairness were bound, mathematically, to diverge under plausible conditions.[338]

A choice therefore must be made about which conception of nondiscrimination to pursue. The computer science literature, while helpful in defining the range of possible conceptions of algorithmic nondiscrimination, is less helpful in evaluating and ranking those definitions.

2. *The Irrelevance of False Positive Rates.* Two of the four definitions of algorithmic nondiscrimination developed above—equally precise coercion and predictive error equality—focus on the rate of false positives. These two definitions differ, however, in terms of their denominator, which is alternatively (1) being coerced, or (2) being a nonrecidivist. False-positive focused definitions not only played a central role in the debate between Northpointe and ProPublica,[339] they have also infiltrated public debate more broadly.[340] A concern with false positives is not without normative appeal. But definitions of nondiscrimination that hinge on false positive rates do not index in any obvious fashion the extent to which an algorithmic instrument exacerbates racial stratification. This section is hence directed at ruling out two of the four possible metrics of racial equity that have attracted the most public attention to date.

For four interrelated reasons, the temptation to focus on false positives should be resisted. *First*, the criminal justice decisions subject to algorithmic resolution are all made in advance of potential adverse

336. *See* Angwin et al., *supra* note 16 (making this cricitism).

337. *See supra* note 20.

338. There are a number of computational fixes, which fall into the categories of pre-, in-, and post-processing. None are a complete fix. *See* Berk et al., *Fairness in Criminal Justice*, *supra* note 263, at 25–29.

339. *See supra* notes 14–22 and accompanying text.

340. *See Perfected in China, a Threat in the West*, ECONOMIST, June 2, 2018, at 11 ("Some sentencing algorithms are more likely to label black defendants than white ones as being at high risk of reoffending."). This reads as a concern with predictive error equality, although this is not wholly free from doubt.

actions. That is, a street stop is conducted by police, bail is denied by a judge, or a sentence is extended before the state knows, or can know, whether a suspect or defendant will in fact commit a criminal act. Officials using an algorithm, therefore, cannot know who is a true positive and who is a false positive among the pool of persons to the right of the vertical threshold illustrated in Figure 1. Even if we assume that an official responsible for applying the algorithm knows the general shape of the distribution (for example, as illustrated in Figure 1), she does not and cannot know whether a particular suspect is in fact going to inflict harm; all she knows is how the algorithm has ranked that person. A test for nondiscrimination that distinguishes false positives from true positives implicitly helps itself to information that is not available to that official. And it is not at all clear why the failure to account for information that the official or algorithm cannot access should be treated as a failure. Provided that the decision rule otherwise achieves valued public goods at the lowest collateral cost, it is not clear why the (ordinarily unknown) distribution of false positives should matter.

Second, the law in practice has a very high tolerance for false positives. In the policing and the pretrial detention contexts in particular, we are willing to tolerate a very high rate of false positives on the ground that the gains to crime suppression offset the costs of those false positives. Hence, in the policing context, a mere showing of "reasonable articulable suspicion," which is far less than probable cause, is enough to warrant a street stop.[341] In the bail context, the standard for detention under federal law is framed in terms of reasonableness and envisages substantial room for error.[342] But disparities in the allocation of state-created goods (or harms) are generally thought to be worrisome if those goods are important. This explains the coverage of housing and employment opportunities by disparate impact regimes.[343] Moreover, if the law takes the view that there is no reason for concern at the prospect of absolutely high levels

341. The initial delineation of rules for a street stop is contained in *Terry v. Ohio*, 392 U.S. 1, 20–22 (1968). The phrase "reasonable, articulable suspicion" was used first in *Brown v. Texas*, 443 U.S. 47, 51 (1979). *See also* Florida v. Royer, 460 U.S. 491, 502 (1983).

342. *See* 18 U.S.C. § 3142(e) (2012) (mandating pretrial detention unless the judge can impose conditions that "reasonably assure the appearance of the person as required and the safety of any other person and the community"); *see also* Stack v. Boyle, 342 U.S. 1, 4 (1951) ("Since the function of bail is limited, the fixing of bail for any individual defendant must be based upon standards relevant to the purpose of assuring the presence of that defendant.").

343. Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2525 (2015) (affirming that disparate impact claims are cognizable under the Fair Housing Act).

of stops or pretrial bail detentions, it is not clear that the law contains the normative resources to establish concern when those resources are allocated in subtly disparate ways—especially if the overall pattern of stops redounds to the net benefit of both society and the subordinated group.[344]

Third, a failure of equally precise coercion or of predictive error equality is a mathematical function of the use of a single threshold for risk for two racial groups with different risk distributions.[345] Given that relationship, it is necessary to choose between unequal rates of false positives and different risk thresholds. Merely pointing to one form of inequality is question begging. If the risk threshold is set at the socially efficient level, moreover, such that it optimizes over immediate costs and benefits for blacks as well as whites,[346] equalizing false positives risks the imposition of unnecessary costs on the minority group. Although not dispositive, it is worth noting that equal protection doctrine does not treat unavoidable disparities generated by the pursuit of a valid governmental interest as cause for concern.[347] At least where the state has no other means of suppressing crime without a violation of equally precise coercion or of predictive error equality, it is not obvious why the ensuing disparities should be treated as fatally problematic.

Finally, and most importantly, if one is concerned with the impact of algorithmic criminal justice on a stratified racial minority, there is no basis for focusing solely on false positives. The negative expressive effects and social harms imposed by criminal justice institutions upon African American communities are not merely triggered by false positives. Directing coercion toward black suspects and defendants even when such coercion is warranted can have an expressive effect on public beliefs about black criminality and more material debilitating effects on communities, families, and children. Indeed, there is no particular reason to believe that any of these spillover costs are less if the person subject to the coercion is in fact a true rather than false positive. Put another way, if one cares about racial stratification, what

344. Note that the mere fact of a violation of equally precise coercion or predictive error equality is not evidence that the net effect of a criminal justice measure is to exacerbate overall racial disparities. There is no empirical equivalence between these terms.

345. Corbett-Davies et al., *supra* note 309, at 799–802.

346. *Id.*; *see supra* note 318 and accompanying text.

347. For a discussion of the current doctrinal position of this element of disparate impact law, see Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV. 1115, 1140 (2016).

should matter is the *absolute* cost of using a coercive tactic against a member of a minority group, net of benefit, for *all* members of that racial group—whether or not they ultimately would have acted in ways that justified coercion. Both kinds of actions have costs; both count for the purposes of racial equity. True, those costs are offset when an algorithm makes a correct prediction, but that is captured better by a focus on the benefits of the coercive measure being allocated.[348]

For these four reasons, I do not think that either equally precise coercion or predictive error equality provides an appropriate metric for thinking about racial equality in this context. Rather, it is desirable in the end to know whether crime control is inflicting more costs than benefits for the minority group as a whole—and not just those who would otherwise not go on to inflict any social harm.

3. *Evaluating the Impact of Algorithmic Criminal Justice on Racial Stratification.* So what does matter? The opening two movements of this Part mapped the effect of criminal justice institutions on racial stratification and charted a general principle of racial equity. Existing criminal justice systems influence the extent of racialized social stratification in society as a whole.[349] Racial equity in criminal justice generally—and in particular in the algorithmic context—should be primarily concerned with mitigating these pernicious effects. It should repudiate the tight linkages that have bound criminal justice to the reproduction of racial hierarchy since the beginning of the twentieth century. Even if the present-day operation of criminal justice institutions cannot undo past harms, at a minimum they should not compound those harms.

The question therefore is which of the available technical benchmarks best captures this pathway between criminal justice and racial stratification. As intimated already, I think that an appropriate benchmark would home in upon the net cost (or benefit) of an algorithmic criminal justice instrument for the racial minority in the socially subordinate position. A measure of costs net of benefits for the racial minority is relevant morally because it captures the extent to which a criminal justice measure depresses the social standing of an

348. I can imagine one more reason for taking normative account of false positives only: One might posit that the ratio of false positives to true positives is a measure of intragroup transfers. The greater the proportion of false positives, that is, the more the burden of crime suppression falls on those members of the minority who are nonrecidivists. This may be a morally relevant quality, but I am not convinced it is a measure of racial equity.

349. *See supra* Part III.A.

already marginalized minority group. In the context of black-white comparisons in America at least, this analysis is simplified by the fact that much violent crime is intraracial. That is, the benefits of a crime suppression measure imposed on blacks are likely to accrue largely to blacks (while the same is true for whites). The analysis would be more complex if we assumed that the racial minority did not capture all or most of the benefits of crime suppression targeting members of that minority.

In my view, there is no one metric developed in the computer science literature or otherwise that captures this concern with racial stratification. Benchmarks that concern the rate of false positives capture in a very loose and partial way the magnitude of unjustified state coercion. But they fail to acknowledge the state's inability to distinguish justified from unjustified exercises of coercion ex ante. Statistical parity does account for the aggregate cost of coercion on a racial minority. But it does so only through a comparative lens; it asks whether the minority is burdened more or less than a majority group. It also fails to consider offsetting benefits for the minority group. Because most crime is intraracial, it fails to account for the possibility that the benefits of crime suppression for blacks outweigh its costs. A comparative measure such as statistical parity is at best considered an evidentiary tool, therefore, rather than a direct measure of racial equity.

An inquiry into racial equity can usefully focus instead on whether the *marginal* decision to impose coercion within the black population can be justified. I present first a simple version of this inquiry that assumes that all costs and benefits are immediate and that there are no spillovers. Consider again Figure 1. Imagine sliding the threshold for coercion for the minority population right, away from the y-axis. At first, the threshold would assign coercion to many people for whom the immediate costs of such coercion outweigh any benefits for the simple reason that their risk of causing harm is so low. At some point in the rightward movement of the threshold, however, the immediate costs of coercion would be balanced by its benefits. When the costs of this marginal decision to coerce are outweighed by its benefits, the threshold has been calibrated such that no net burden is being placed on the minority population, and all coercion generates a net gain for that group. Assuming that most relevant crime is intraracial, this means that the marginal benefits of coercion (for the black community) are greater than the costs of coercion (for the black community). Such a policy leaves that racial group no worse off than it would otherwise be.

For interventions that prevent serious crimes, there is no reason to think that the immediate costs of coercion, or the immediate benefits of crime control, vary between racial groups. Moreover, spillovers can be ignored because such costs are likely to be rounding errors in relation to the costs of murder, sexual assault, armed robbery, and the like. Such a tightly focused analysis might, for example, be appropriate in the analysis of bail decisions where a suspect may go on to commit a serious violent crime. Under these conditions, a single risk threshold calibrated to be socially optimal (in the sense of eliminating cost-unjustified coercion) will satisfy racial equity. It will also be socially efficient.

This goal has likely not been reached in practice. Even assuming that criminal justice decision-makers are applying a single threshold rule (rather than being influenced by animus or racial stereotypes), it is very likely that many present uses of police coercion and detention are unjustified. The benefits of state coercion are likely overestimated, while its costs are underestimated. Consistent with this prediction, current risk assessment tools estimate the benefits of coercion but do not measure costs.[350] Still, the present lack of empirical data on the costs and benefits of many familiar criminal justice institutions, such as street stops and bail denials, means that this intuition is hard to substantiate. But the available data suggests an excess of coercion beyond the socially optimal.[351] When the supernumerary costs of such coercion fall on racial minorities, they intensify racial stratification. Ratcheting back the sheer volume of coercion, therefore, may be a first-order task in reform projects that have racial equity in mind.

This simple analysis of racial equity accounts only for the immediate costs and benefits of coercion. It does not account for the externalities set forth in Part III.A. A more complex model of racial equity would account for all negative spillovers from algorithmically allocated coercion. These externalities are substantially greater for racial minorities than for the racial majority. They are also nontrivial in scale. Where less serious crime is concerned (e.g., public order offenses), it is likely that these externalities are of the same magnitude as the immediate benefits and costs of crime control. Second-order,

---

350. *See* Slobogin, *supra* note 47, at 584–86.

351. *See* Huq, *Disparate Policing*, *supra* note 5, at 2413–29; Note, *Bail Reform and Risk Assessment: The Cautionary Tale of Federal Sentencing*, 131 HARV. L. REV. 1125, 1127–28 (2018) ("The pretrial imprisonment rate in the United States is among the highest in the world—more than four times the world's median pretrial imprisonment rate."); *see also* Mayson, *supra* note 126, at 545–48 (explaining costs and benefits of bail in a way that clarifies its complexity).

downstream costs of coercion therefore cannot be safely ignored as rounding errors in an analysis of the criminal justice system's dynamic effects. The analysis for less serious crime, or for interventions that do not impede serious harms, is hence different from the analysis when serious social harm is directly at stake.

Accounting for the racially asymmetrical distribution of externalities alters the racial equity analysis. It means that the marginal costs of coercion are likely to be greater for the racial minority. Accordingly, the point on the x-axis at which costs are equal to benefits for the minority is to the right of the same break-even point for the majority group. That is, because the operation of criminal justice coercion generates asymmetrical harms to black families and black communities, and exacerbates Kennedy's racial tax, there will be a class of crimes for which a greater benefit will be required to achieve net positive effects for black suspects. And because the costs and benefits of crime are largely intraracial, the same higher risk threshold will be required to achieve social efficacy. Whether the focus is social efficiency or racial equity, this implies that the risk threshold for blacks should be set at a higher level (i.e., farther to the right in Figure 1) than the threshold for whites. Therefore, accounting for both the immediate and spillover costs of crime control when its immediate benefits are small conduces to a bifurcated risk threshold—one rule for the majority, and one for minority. The single vertical line in Figure 1 would bifurcate. The line for blacks would move rightward.

This is akin to common affirmative action schemes, in which otherwise similar black and white persons are treated differently because of the different spillover consequences of their treatment. In the affirmative action context, the existence of a positive diversity benefit (which is another kind of spillover) warrants a less stringent threshold rule for assigning a benefit to the racial minority.[352] In the criminal justice context, similarly, the existence of negative spillovers for black families and communities warrants a more stringent risk threshold for the racial minority. The argument for a bifurcated classification rule is arguably stronger here than the argument for affirmative action: The alleviation of racial stratification, in my view, is a more acute interest than diversity because it directly benefits the most marginalized (which affirmative action may not) and immediately relieves stigmatic and material harms. Alleviating the effect of accumulated disadvantage caused by the historical operation of

352. *See, e.g.*, Fisher v. Univ. of Tex. at Austin, 136 S. Ct. 2198, 2208 (2016).

criminal justice institutions, in other words, is a more compelling goal than crafting a well-rounded university population.

Potentially unlike affirmative action, however, the case for multiple risk thresholds can be made independently on either racial equity or pure social efficiency grounds. So long as a policy's costs (or its benefits) are largely internalized by racial groups, and so long as costs are greater at the margin for the minority group, a socially optimal rule would require different risk thresholds. Where the state adopts a cost-benefit approach to criminal justice policy,[353] an exacting approach to cost-benefit trade-offs in crime control may in some cases generate dual thresholds.[354] In the algorithmic context, it is worth noting that a machine-learning tool, given the necessary data and asked to vindicate social efficiency (understood in a capacious sense that reached both static and dynamic effects), could converge on a bifurcated rule absent race-conscious human decision-making.

However that goal is approached, its achievement imposes large new epistemic burdens on the state. Whereas risk assessment in criminal justice to date has focused narrowly on the costs of crime, a rigorously executed algorithmic method demands data on the costs of crime control. This is a matter not merely of counting state expenditures but also of measuring spillovers. This is a massive task. But its size and difficulty ought not to be a justification for avoidance. The current dearth of information about the spillover costs of criminal justice institutions, particularly for minority communities, is causally related to their stratifying effects. Ignorance of spillovers, coupled to a myopic focus on a small number of high-profile crimes, creates the epistemic background against which actually existing state institutions compound racial stratification. That ignorance is thus a form of "hermeneutical injustice," in which "some significant area of one's social experience [is] obscured from collective understanding owing to persistent and wide-ranging hermeneutical marginalization."[355] Racial inequity cannot be justified by hermeneutic injustice. Precisely how the epistemic gap will be closed is a large question, and I do not take it up

353. A version of cost-benefit analysis is endorsed in Barry Friedman & Maria Ponomarenko, *Democratic Policing*, 90 N.Y.U. L. REV. 1827, 1907 (2015) (encouraging even "small steps" in that direction).

354. It is also possible that a jurisdiction could pursue social efficiency by deploying a nonracial bifurcation in the risk threshold. For instance, it may in some instances be possible to employ socioeconomic stratification to much the same end.

355. MIRANDA FRICKER, EPISTEMIC INJUSTICE: POWER & THE ETHICS OF KNOWING 154 (2007) (emphasis omitted).

here. But it is worth noting that the algorithmic tools mapped here may have a role. Determining how big-data tools can contribute to this epistemic enterprise, indeed, is perhaps the next technological frontier in criminal justice.

At the same time, a multiple threshold rule for different racial groups runs headlong into the anticlassification rule of equal protection doctrine.[356] At a minimum, it would receive strict scrutiny.[357] As a result, a multiple threshold regime would be in serious constitutional jeopardy. Under these conditions, which are hardly empirically implausible, the regime imperiled by our constitutional equality doctrine is the only one that both mitigates racial stratification *and also* maximizes social welfare. Why would we want to place that regime beyond reach? I can think of no good answer. Such a result, in my view, tells us more about our wrongheaded racial equality doctrine than it does about the substance of algorithmic criminal justice.

CONCLUSION

Algorithmic criminal justice, relying first on machine learning and then on deep learning, is only now beginning to impinge on criminal justice institutions. For a much longer time, the latter have been sites for the production of racial stratification. This comes in the form of a policing and carceral apparatus that weighs most heavily on African Americans. It also arises thanks to a racial tax that extends to all members of the group, whether or not they have any connection to criminality.

Given this history, it seems to me important to get algorithmic criminal justice right. Such tools, if fashioned wisely, might be useful in restoring equilibrium and mitigating the burden of racial externalities. Wrongly configured, they may prove subtle levers for preserving or even exacerbating those burdens. Wrongly configured, I also fear, they would be exceedingly hard to dislodge. My aim in this Article has been to demonstrate that constitutional law does not contain effectual tools to meet these problems. It is a mistake, therefore, to contort constitutional doctrine in the hope that it will do service in a context where it is so substantially ill fitted. Far better, in my view, to recognize that the constitutional law of racial equality has almost nothing cogent to say about what counts as a racially just algorithm. It might instead

356. *See supra* notes 229–35 and accompanying text.
357. Johnson v. California, 543 U.S. 499, 505 (2005).

achieve the remarkable doubleheader of impeding both racial equity and social welfare maximization. The doctrine is thus a moral vacuity.

Reformulation of the doctrine, in my view, is desirable but unlikely. In the interim, algorithm designers, local officials, and state legislators should instead ask directly how best to achieve racial equity given the shape of existing criminal justice institutions and the technical tools at their disposal. I have offered an answer to that question that draws on, without quite tracking, existing technical definitions of algorithmic nondiscrimination. I have further stressed that my approach has the distinctive feature of aligning racial equity with social efficiency. My project has been demarcated in terms of algorithmic criminal justice. But it should not escape notice that there is no particular reason to confine the scope of the analysis to algorithmic tools, or even to criminal justice. But those extensions are for another day. For now, a recognition of the potential convergence of equity and efficiency might move us closer to a remedy for the difficult, enduring, and damaging legacy of our racialized criminal justice past.