THE YALE LAW JOURNAL

SANDRA G. MAYSON

Bias In, Bias Out

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

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

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

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

II. PREDICTION AS A MIRROR

A. The Premise of Prediction

There is a simple reason why it is impossible to achieve equality by every metric when base rates differ: prediction functions like a mirror. The premise of prediction is that, absent intervention, history will repeat itself. So what prediction does is identify patterns in past data and offer them as projections about future events. If there is racial disparity in the data, there will be racial disparity in prediction too. It is possible to replace one form of disparity with another, but impossible to eliminate it altogether.

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

The deep problem, in other words, is not algorithmic methodology. Any form of prediction that relies on data about the past will produce racial disparity if the past data shows the event that we aspire to predict—the target variable—occurring with unequal frequency across racial groups. And if an algorithm's forecasts are correct at equal rates across racial lines, as were the COMPAS forecasts in Broward County,[112] any disparity in prediction reflects disparity in the data. To understand and redress disparity in prediction, it is therefore necessary to understand how and when racial disparity arises in the data that we look to as a representation of *past* crime.

B. Racial Disparity in Past-Crime Data

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

C. *Two Possible Sources of Disparity*

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

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

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

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

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

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

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

1. *Disparate Law Enforcement Practice?*

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

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

[126] Part IV considers this possibility.

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

[128] *See, e.g.*, MODEL PENAL CODE § 1.02(2), Reporters' Note 31 (AM. LAW INST., Proposed Final Draft 2017) (noting that racial disparities in sentencing that arise from racial skew in law enforcement "are largest for crimes at the low end of the seriousness scale—especially drug offenses," and collecting sources); Lauren Nichol Gase et al., *Understanding Racial and Ethnic Disparities in Arrest: The Role of Individual, Home, School, and Community Characteristics*, 8 RACE & SOC. PROBS. 296, 304-08 (2016) (finding "that racial/ethnic differences in arrest were not explained by differences in individual-level delinquent behaviors," but were explained by "neighborhood racial composition"); Kristian Lum & William Isaac, *To Predict and Serve?*, 13

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

To the extent that racial disparities in past-arrest rates derive from disparate law enforcement practice, that distortion makes "future arrest" a racially skewed proxy for "future crime."

a

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

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

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

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

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

2. *Disparate Rates of Crime Commission?*

The second possible explanation for racial disparity in past-arrest rates is a difference in the underlying incidence of crime. This possibility arises because crime is the product of complex social and economic determinants that, in a race- and class-stratified society, may also correlate with demographic traits. Where that is so, the incidence of a given type of crime may vary among demographic groups. A number of recent studies have found, for instance, that contemporary white and Hispanic college students use illicit drugs at significantly higher rates than African American and Asian students.[133] White men have committed the vast majority of mass shootings in the United States during the last thirty years.[134] Nationwide firearm homicide rates have been higher in recent decades in black communities than in white ones, but the degree of disparity varies by

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

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

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

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

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

The point is that *if* underlying offense rates do vary by race in the data on which a given algorithm is built, racial disparity in prediction is unavoidable. The reason, once again, is that prediction functions as a mirror. If the black population in the relevant data is statistically riskier with respect to the designated crime category, risk-assessment tools will reflect as much. If the mirror is modified to ignore this statistical fact, that very blindness will have disparate racial

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

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

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

[138] Skeem & Lowenkamp, *supra* note 38, at 690 (opining that arrest for a "violent offense" is a "valid criterion" free from racial skew in law enforcement); *see also* Alex R. Piquero et al., *A Systematic Review of Age, Sex, Ethnicity, and Race as Predictors of Violent Recidivism*, 59 INT'L J. OFFENDER THERAPY & COMP. CRIMINOLOGY 5, 17 (2015) (finding "that age, sex, and race . . . were significantly related to violent recidivism").

impact: in treating the black and white groups subject to assessment as statistically identical, the tools will "miss" more of the designated crimes committed by black individuals – crimes that, because most crime is intraracial, will disproportionately befall communities of color.[139] No matter how we alter the data or algorithm, then, inequality in commission rates for the crime(s) we undertake to predict will produce inequality in prediction.

It is important, in considering this possibility, to recognize what any such difference in crime commission rates would and would not signify. Differential crime rates do not signify a difference across racial groups in individuals' innate "propensity" to commit crime.[140] They signify social and economic divides. Where the incidence of crimes of poverty and desperation varies by race, it is because society has segregated communities of color and starved them of resources and opportunity.[141] Where race and gender differences exist in the rate of high-stakes financial crime, it is because white men retain control of the levers of high-stakes finance.[142] Crime rates are a manifestation of deeper forces; racial variance in crime rates, where it exists, manifests the enduring social and economic inequality produced by centuries of racial subordination.

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

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

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

[140] The notion that differential crime rates signal a difference in innate criminal propensity has been a central justification for racist ideology and practices. *See generally, e.g.*, KENNEDY, *supra* note 123, at 12-17 (analyzing race relations in the administration of criminal justice); RUSSELL-BROWN, *supra* note 123 (discussing race, crime, and law, beginning with slavery in the United States).

[141] *See, e.g.*, MODEL PENAL CODE § 1.02(2) cmt. k (AM. LAW INST., Proposed Final Draft 2017) ("Serious crime rates, and victimization rates, are highest in America's most disadvantaged communities, which overwhelmingly are minority communities."); *id.* (citing sources on "the multiple causes of high crime rates in disadvantaged communities," along with research demonstrating that "the 'underclass' status of a community is associated with high crime rates among those who live there, regardless of race and ethnicity"); MEHRSA BARADARAN, THE COLOR OF MONEY: BLACK BANKS AND THE RACIAL WEALTH GAP (2017); KENNEDY, *supra* note 123. This is not to disclaim all individual responsibility for criminal acts. But individual responsibility for particular acts does not amount to group responsibility for group crime rates.

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

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

Distortion can take many forms. In the criminal justice context, the choice of a proxy target variable with racial skew (i.e., "any arrest" as a proxy for "commission of serious crime") may be the most important.[143] But racial distortion can also result if the data are systematically less reliable for one racial group than for another. This problem can arise if the data are simply more limited for one racial group.[144]

s

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

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

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

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

⋆ ⋆ ⋆

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

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

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

III. NO EASY FIXES

As the risk-assessment-and-race debate accelerates, critics increasingly argue for three strategies to promote racial equity in prediction. The first is the exclusion of both race and factors heavily correlated with race as input variables.[149] The second is "algorithmic affirmative action": some intervention in the design of a predictive algorithm to equalize its outputs, by one or more of the metrics enumerated above.[150] In particular, advocates have urged intervention to ensure an equal rate of adverse predictions across racial groups (statistical parity),[151] or equal error rates among those in each racial group who have the same outcome (parity in false-positive and false-negative rates).[152] The discussion here will use the term "algorithmic affirmative action" to refer to these proposals collectively, acknowledging that this shorthand is reductive. Lastly, critics argue that if algorithms cannot be made race neutral, the criminal justice system should reject algorithmic methods altogether.[153]

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

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

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

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

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

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

A. Regulating Input Variables[155]

Input variables are often cited as the primary concern in the quest for racial equity in risk assessment. It is an almost-universal orthodoxy, in fact, that race must be excluded as an input to prediction.[156] Many people extend this principle to variables that correlate with race in a given locale, like zip code.[157] The underlying concern is that the use of such factors will produce higher risk scores for black defendants and thereby compound historical racial oppression.

This focus on input variables, however, is not an effective approach to achieving racial equity.[158] The most basic reason is that excluding race and race proxies might actually hurt black defendants. In this context, as elsewhere, being blind to race can mean being blind to racism. As Justice Sotomayor replied to Chief Justice Roberts, the "way to stop discriminating on the basis of race" is not to ignore race, but rather to apply law and develop policy "with eyes open to the unfortunate effects of centuries of racial discrimination."[159]

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

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

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

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

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

[159] Chief Justice Roberts, writing for the plurality in *Parents Involved in Community Schools v. Seattle School District No. 1*, declared that "[t]he way to stop discrimination on the basis of race is to stop discriminating on the basis of race." 551 U.S. 701, 748 (2007) (plurality opinion). Justice Sotomayor rejoined, seven years later, that "[t]he way to stop discrimination on the

A simple example illustrates. When I worked in New Orleans as a public defender, the significance of arrest there varied by race. If a black man had three arrests in his past, it suggested only that he had been living in New Orleans. Black men were arrested all the time for trivial things. If a white man, however, had three past arrests, it suggested that he was really bad news! White men were hardly ever arrested; three past arrests indicated a highly unusual tendency to attract law enforcement attention.[160] A race-blind algorithm would not observe this difference. It would treat the two men as posing an identical risk. The algorithm could not consider the arrests in the context of disparate policing patterns and recognize that arrests were a much less significant indicator of risk for a black man than for a white man.[161] It would perpetuate the historical inequality by overestimating the black man's relative riskiness and underestimating the relative riskiness of the white man.

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

In statistical terms, the problem is that, as a result of disparate law enforcement practices, race might moderate the predictive value of certain variables (or the algorithm as a whole), such that the algorithm overestimates risk for black people relative to white people.[162] A few risk-assessment-tool developers have

basis of race is to speak openly and candidly on the subject of race, and to apply the Constitution with eyes open to the unfortunate effects of centuries of racial discrimination." Schuette v. Coal. to Defend Affirmative Action, 134 S. Ct. 1623, 1676 (2014) (Sotomayor, J., dissenting).

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

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

[162] This situation arises in every predictive context. In education testing, for instance, it is well established that the correlation between SAT scores and intelligence varies by race and by circumstance. *See, e.g.*, Harold Berlak, *Race and the Achievement Gap*, *in* CRITICAL SOCIAL ISSUES IN AMERICAN EDUCATION: DEMOCRACY AND MEANING IN A GLOBALIZING WORLD 223, 227 (H. Svi Shapiro & David E. Purpel eds., 3d ed. 2005) (discussing the racial achievement gap

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

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

in other standardized tests). A high score achieved by a student who benefited from the best possible primary education and extensive SAT preparation likely means less about her native intelligence than the same score achieved by a student who did not.

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

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

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

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

[167] *See* Kim, *supra* note 43, at 918 ("If the goal is to reduce biased outcomes, then a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies and the risk of omitted variable bias."); Lipton et al., *supra* note 103 (arguing on the basis of statistical examples that a prohibition on race or sex data is counterproductive); Corbett-Davies & Goel, *supra* note 68, at 9 (explaining the "[l]imitations of anti-classification" as a fairness metric).

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

In fact, to achieve any specific form of output equality, it may be necessary to treat race as an input. To equalize false-positive rates across racial groups, for example, it will likely be necessary to have race-specific risk thresholds for each risk class—which is to say that the algorithm will treat people who pose the same risk differently on the basis of race.[169] The same is likely true for equalizing cost ratios across racial groups.[170] To achieve predictive parity, it may be necessary to manipulate the data to cancel out the effect of race on other observable variables,[171] or to assess the predictive import of every input variable contingent on race. As Solon Barocas and Andrew Selbst have noted, algorithmic prediction thus offers a particularly clear window on the conflict between anticlassification and antisubordination conceptions of equality.[172]

Yet neither excluding race and race-correlated factors nor including them can equalize outcomes entirely if the event we have undertaken to predict—the target variable—correlates with race itself. So long as the target variable correlates with race, regulating input data is futile. If the event we have undertaken to predict happens with greater frequency to people of color, a competent algorithm will predict it with greater frequency for people of color. Whatever input data are made available, the facts that correlate with the target variable—and therefore

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

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

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

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

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

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

The larger point is that colorblindness is not a meaningful measure of equality. It can exacerbate rather than mitigate racial disparity in prediction.[176] And even if it does mitigate disparity in prediction, that improvement may come at a cost to accuracy that itself has a racially disparate impact. As long as the target variable correlates with race, predictions will be racially uneven—or they will be so distorted as to be useless. In those circumstances, colorblindness is at best a superficial, and at worst a counterproductive, strategy for racial equity.[177]

B. Equalizing (Some) Outputs

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

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

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

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

[176] *See* Huq, *supra* note 14, at 1100; Kim, *supra* note 43, at 867 ("[I]f the goal is to discourage classification bias, then the law should not forbid the inclusion of race, sex, or other sensitive information as variables, but seek to preserve these variables, and perhaps even include them in some complex models."); Kroll et al., *supra* note 25, at 693-95.

[177] *Cf.* Mayson, *supra* note 155; David A. Strauss, *The Myth of Colorblindness*, 1986 SUP. CT. REV. 99, 114 ("The one option that is not open is the ideal of colorblindness—treating race as if it were, like eye color, a wholly irrelevant characteristic. That is because it is not a wholly irrelevant characteristic. Race correlates with other things . . . .").

rates. The stakes of such interventions depend on whether the disparity they seek to redress is a product of distortion in the data or of a difference in underlying crime rates by race. In either case, though, the interventions fall short.

1. Equalizing Outputs to Remedy Distortion

First, consider algorithmic affirmative action designed to remedy racial distortion in the data vis-à-vis the event we aspire to predict. In the context of criminal justice risk assessment, the gravest concern is that racial disparity in overall arrest rates reflects disparate law enforcement, rather than disparate rates of offending. If this is true, and what we assess is the likelihood of arrest, then risk scores will overstate the risk posed by black men relative to the risk of actual crime commission. The goal of algorithmic affirmative action is to adjust the data to cancel out this racial distortion in arrest rates.[178]

This strategy presumes that the scale of the distortion is known. If so, it should indeed be possible to cancel it out, although there are technical complexities. But it is hardly ever the case that the scale of the distortion is known.[179]

1

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

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

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

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

m-

Stated in more general terms, one might object that we can *never* be confident that our target variable is free from racial distortion.[189] We must rely on the past to predict the future, but we see the past only hazily through the splintered lens of data.[190] We can never know how faithfully the data represent past reality because we have no direct access to past reality.

This is a profound objection, and it applies to more than algorithmic methods. It is an objection to prediction itself. All prediction presumes that we can read the past with enough reliability to make useful projections about the future. Perhaps in some contexts we cannot. Maybe our past-crime data are inadequate to serve as the basis for any prediction.[191] Or maybe the answer varies by crime category. But if this is the case, the answer is not to make the data reflect the past as we wish it had been. That merely distorts the mirror so that it neither reflects the data nor any demonstrable reality. The answer is simpler. If past data do not reliably represent the events we want to avoid, we should stop consulting them as a guide for the future.

2. *Equalizing Outputs in the Case of Differential Offending Rates*

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

[188] The correspondence between arrest rates and crime-report rates by race is one fact that scholars sometimes cite as evidence that arrest rates lack racial skew vis-à-vis offending rates. *See, e.g.*, Skeem & Lowenkamp, *supra* note 38, at 690.

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

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

[191] *See* Barocas & Selbst, *supra* note 3, at 682-84; Grant T. Harris & Marnie E. Rice, *Bayes and Base Rates: What Is an Informative Prior for Actuarial Violence Risk Assessment?*, 31 BEHAV. SCI. & L. 103, 121 (2013) ("What is not axiomatic is the straightforward application of assumptions about priors . . . to violence risk assessment—that remains a set of important empirical matters."); Selbst, *supra* note 3, at 140-43.

a. *Practical Problems*

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

[192] Equalizing false-positive rates will result in fewer false positives ("law abiders" mistakenly forecast for rearrest) for the high-base-rate group than the low-base-rate group because there are fewer "law abiders" in the high-base-rate group in the first place.

[193] Sam Corbett-Davies and colleagues, analyzing the same Broward County data that ProPublica did, found that achieving parity in false-positive rates while still optimizing for public safety (and without detaining additional defendants) would result in a 7% increase in violent crime.

C. Rejecting Algorithmic Methods

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

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

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

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

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

[207] *See* MODEL PENAL CODE § 6B.07(1)(c) (AM. LAW INST., Tentative Draft No. 4, 2016) (noting "the danger that the use of criminal-history provisions to increase the severity of sentences may have disparate impacts on racial or ethnic minorities, or other disadvantaged groups"); *id.* § 6B.07(4) (instructing sentencing commissions to "monitor the effects of . . . incorporating offenders' criminal history as a factor relevant to sentencing," giving "particular attention" to whether it "contributes to punishment disparities among racial and ethnic minorities, or other disadvantaged groups"); *id.* § 6B.07 cmt. ("An accumulating body of research indicates

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

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

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

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

[210] *See, e.g.*, Hamilton, *supra* note 30, at 284-85 ("[I]f constitutionally or ethically suspect variables are excised [from risk-assessment tools], it is likely that fact-finders would consider [them] informally anyway, rendering their use less reliable, transparent, and consistent."); Starr, *supra* note 30, at 824 ("There is, to be sure, considerable statistical research suggesting that judges (and prosecutors) *do* on average treat female defendants more leniently than male defendants.").

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

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

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

This is not to overstate the case for algorithms. The evidence for the superior accuracy of actuarial over subjective prediction is not watertight; a great deal depends on the algorithm at issue and the details of its use.[221] There is an urgent need for further research to document the comparative effects of the two methods on the ground.[222] It is also true that there are concerns unique to algorithmic methods. Algorithmic assessment carries a scientific aura, which can produce unwarranted deference or a mistaken impression of objectivity.[223] Some algorithms *are* opaque. Algorithmic systems may be vulnerable to entrenchment because they require specialized skill and resources to alter. Finally, if algorithmic assessment operates on a much larger scale than subjective assessment does, it

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

[221] A thoughtful judge with broad experience may be more effective at assessing risk than a rudimentary algorithm, but a sophisticated algorithm may be more effective than a bad judge; and a good judge operating with the benefit of a good algorithm may be most effective of all. *See* Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES 1 (2018) (finding that untrained human participants performed nearly as well as COMPAS); Green & Chen, *supra* note 209; Starr, *supra* note 30, at 855 (concluding that "the shibboleth that actuarial prediction outperforms clinical prediction is—like the actuarial risk predictions themselves—a generalization that is not true in every case"); Stevenson, *supra* note 6, at 14-19 (surveying existing evidence).

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

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

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

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

Actuarial risk assessment, in other words, has not created the problem of racially disparate prediction, but rather exposed it. Its contribution is to illuminate—in formal, quantitative terms—the way in which prediction replicates and magnifies inequality in the world. More than thirty years ago, Noval Morris and Marc Miller, arguing for a frank reckoning with the costs and benefits of preventive detention, wrote: "We propose to get the dragon out onto the plain."[225] Algorithmic prediction puts the dragon of predictive inequality out on the plain. It is frightful, but at least we can see it. Rejecting the precise mirror of algorithmic prediction in favor of subjective risk assessment does not solve the problem. It merely turns a blind eye.

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

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

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