VIRGINIA LAW REVIEW

VOLUME 106 | JUNE 2020 | NUMBER 4

ARTICLES

MEASURING ALGORITHMIC FAIRNESS

*Deborah Hellman**

Algorithmic decision making is both increasingly common and increasingly controversial. Critics worry that algorithmic tools are not transparent, accountable, or fair. Assessing the fairness of these tools has been especially fraught as it requires that we agree about what fairness is and what it requires. Unfortunately, we do not. The technological literature is now littered with a multitude of measures, each purporting to assess fairness along some dimension. Two types of measures stand out. According to one, algorithmic fairness requires that the score an algorithm produces should be equally accurate for members of legally protected groups—blacks and whites, for example. According to the other, algorithmic fairness requires that the algorithm produce the same percentage of false positives or false negatives for each of the groups at issue. Unfortunately, there is often no way to achieve parity in both these dimensions. This fact has led to a pressing question. Which type of measure should we prioritize and why?

* D. Lurton Massee, Jr. Professor of Law and Roy L. and Rosamond Woodruff Morgan Professor of Law at the University of Virginia School of Law. I would like to thank Charles Barzun, Aloni Cohen, Aziz Huq, Kim Ferzan, Niko Kolodny, Sandy Mayson, Tom Nachbar, Richard Schragger, Andrew Selbst, and the participants in the Caltech 10th Workshop in Decisions, Games, and Logic: Ethics, Statistics, and Fair AI, the Dartmouth Law and Philosophy Workshop, and the computer science department at UVA for comments and critique. In addition, I would like to thank Kristin Glover of the University of Virginia Law Library and Judy Baho for their excellent research assistance. Any errors or confusions are my own.

This Article makes three contributions to the debate about how best to measure algorithmic fairness: one conceptual, one normative, and one legal. Equal predictive accuracy ensures that a score means the same thing for each group at issue. As such, it relates to what one ought to believe about a scored individual. Because questions of fairness usually relate to action, not belief, this measure is ill-suited as a measure of fairness. This is the Article's conceptual contribution. Second, this Article argues that parity in the ratio of false positives to false negatives is a normatively significant measure. While a lack of parity in this dimension is not constitutive of unfairness, this measure provides important reasons to suspect that unfairness exists. This is the Article's normative contribution. Interestingly, improving the accuracy of algorithms overall will lessen this unfairness. Unfortunately, a common assumption that anti-discrimination law prohibits the use of racial and other protected classifications in all contexts is inhibiting those who design algorithms from making them as fair and accurate as possible. This Article's third contribution is to show that the law poses less of a barrier than many assume.

INTRODUCTION

At an event celebrating Martin Luther King, Jr. Day, Representative Alexandria Ocasio-Cortez (D-NY) expressed the concern, shared by many, that algorithmic decision making is biased. "Algorithms are still made by human beings, and those algorithms are still pegged to basic human assumptions," she asserted. "They're just automated. And if you don't fix the bias, then you are automating the bias."[1] The audience inside the room applauded. Outside the room, the reaction was more mixed. "Socialist Rep. Alexandria Ocasio-Cortez . . . claims that algorithms, which are driven by math, are racist," tweeted a writer for the Daily Wire.[2] Math is just math, this commentator contends, and the idea that math can be unfair is crazy.

This controversy is just one of many to challenge the fairness of algorithmic decision making.[3] The use of algorithms, and in particular

[1] Blackout for Human Rights, MLK Now 2019, Riverside Church in the City of N.Y. (Jan. 21, 2019), https://www.trcnyc.org/mlknow2019/ [https://perma.cc/L45Q-SN9T] (interview with Rep. Ocasio-Cortez begins at approximately minute 16, and comments regarding algorithms begin at approximately minute 40); see also Danny Li, AOC Is Right: Algorithms Will Always Be Biased as Long as There's Systemic Racism in This Country, Slate (Feb. 1, 2019, 3:47 PM), https://slate.com/news-and-politics/2019/02/aoc-algorithms-racist-bias.html [https://perma.cc/S97Z-UH2U] (quoting Ocasio-Cortez's comments at the event in New York); Cat Zakrzewski, The Technology 202: Alexandria Ocasio-Cortez Is Using Her Social Media Clout To Tackle Bias in Algorithms, Wash. Post: PowerPost (Jan. 28, 2019), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2019/01/28/the-technology-202-alexandria-ocasio-cortez-is-using-her-social-media-clout-to-tackle-bias-in-algorithms/5c4dfa9b1b326b29c3778cdd/?utm_term=.541cd0827a23 [https://perma.cc/LL4Y-FWDK] (discussing Ocasio-Cortez's comments and reactions to them).

[2] Ryan Saavedra (@RealSaavedra), Twitter (Jan. 22, 2019, 12:27 AM), https://twitter.com/RealSaavedra/status/1087627739861897216 [https://perma.cc/32DD-QK5S]. The coverage of Ocasio-Cortez's comments is mixed. See, e.g., Zakrzewski, supra note 1 (describing conservatives' criticism of and other media outlets' and experts' support of Ocasio-Cortez's comments).

[3] See, e.g., Hiawatha Bray, The Software That Runs Our Lives Can Be Biased—But We Can Fix It, Bos. Globe, Dec. 22, 2017, at B9 (describing a New York City Council member's proposal to audit the city government's computer decision systems for bias); Drew Harwell,

their connection with machine learning and artificial intelligence, has attracted significant attention in the legal literature as well. The issues raised are varied, and include concerns about transparency,[4] accountability,[5] privacy,[6] and fairness.[7] This Article focuses on fairness—the issue raised by Ocasio-Cortez. It focuses on *how* we should assess what makes algorithmic decision making fair. Fairness is a moral concept, and a contested one at that. As a result, we should expect that different people will offer well-reasoned arguments for different conceptions of fairness. And this is precisely what we find.

Amazon's Facial-Recognition Software Has Fraught Accuracy Rate, Study Finds, Wash. Post, Jan. 26, 2019, at A14 (reporting on an M.I.T. Media Lab study that found that Amazon facial-recognition software is less accurate with regard to darker-skinned women than lighter-skinned men, and Amazon's criticism of the study); Tracy Jan, Mortgage Algorithms Found To Have Racial Bias, Wash. Post, Nov. 15, 2018, at A21 (reporting on a University of California at Berkeley study that found that black and Latino home loan customers pay higher interest rates than white or Asian customers on loans processed online or in person); Tony Romm & Craig Timberg, Under Bipartisan Fire from Congress, CEO Insists Google Does Not Take Sides, Wash. Post, Dec. 12, 2018, at A16 (reporting on Congresspeople's concerns regarding Google algorithms which were voiced at a House Judiciary Committee hearing with Google's CEO).

[4] See, e.g., Danielle Keats Citron, Technological Due Process, 85 Wash. U. L. Rev. 1249, 1288–97 (2008); Natalie Ram, Innovating Criminal Justice, 112 Nw. U. L. Rev. 659 (2018); Rebecca Wexler, Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System, 70 Stan. L. Rev. 1343 (2018).

[5] See, e.g., Margot E. Kaminski, Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability, 92 S. Cal. L. Rev. 1529 (2019); Joshua A. Kroll et al., Accountable Algorithms, 165 U. Pa. L. Rev. 633 (2017); Anne L. Washington, How To Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate, 17 Colo. Tech. L.J. 131 (2018) (arguing for standards governing the information available about algorithms so that their accuracy and fairness can be properly assessed). But see Jon Kleinberg et al., Discrimination in the Age of Algorithms (Nat'l Bureau of Econ. Research, Working Paper No. 25548, 2019), http://www.nber.org/papers/w25548 [https://perma.cc/JU6H-HG3W] (analyzing the potential benefits of algorithms as tools to prove discrimination).

[6] See generally Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015) (discussing and critiquing internet and finance companies' non-transparent use of data tracking and algorithms to influence and manage people); Anupam Chander, The Racist Algorithm?, 115 Mich. L. Rev. 1023, 1024 (2017) (reviewing Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information (2015)) (arguing that instead of "transparency in the design of the algorithm" that Pasquale argues for, "[w]hat we need . . . is a transparency of inputs and results") (emphasis omitted).

[7] See, e.g., Aziz Z. Huq, Racial Equity in Algorithmic Criminal Justice, 68 Duke L.J. 1043 (2019) (arguing that current constitutional doctrine is ill-suited to the task of evaluating algorithmic fairness and that current standards offered in the technology literature miss important policy concerns); Sandra G. Mayson, Bias In, Bias Out, 128 Yale L.J. 2218 (2019) (discussing how past and existing racial inequalities in crime and arrests mean that methods to predict criminal risk based on existing information will result in racial inequality).

The computer science literature is filled with a proliferation of measures, each purporting to capture fairness along some dimension. This Article provides a pathway through that morass. It makes three contributions: one conceptual, one normative, and one legal. This Article argues that one of the dominant measures of fairness offered in the literature tells us what to *believe*, not what to do, and thus is ill-suited as a measure of fair treatment. This is the conceptual claim. Second, this Article argues that the ratio between false positives and false negatives offers an important indicator of whether members of two groups scored by an algorithm are treated fairly, vis-à-vis each other. This is the normative claim. Third, this Article challenges a common assumption that anti-discrimination law prohibits the use of racial and other protected classifications in all contexts. Because using race within algorithms can increase both their accuracy and fairness, this misunderstanding has important implications. This Article's third contribution is to show that the law poses less of a barrier than many assume.

We can use the controversy over a common risk assessment tool used by many states for bail, sentencing, and parole to illustrate the controversy about how best to measure fairness.[8] The tool, called COMPAS, assigns each person a score that indicates the likelihood that the person will commit a crime in the future.[9] In a high-profile exposé, the website ProPublica claimed that COMPAS treated blacks and whites differently because black arrestees and inmates were far more likely to be erroneously classified as risky than were white arrestees and inmates despite the fact that COMPAS did not explicitly use race in its algorithm.[10] The essence of ProPublica's claim was this:

> In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at

[8] See Julia Angwin et al., Machine Bias, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/BA53-JT7V].

[9] Equivant, Practitioner's Guide to COMPAS Core 7 (2019), http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf [https://perma.cc/LRY6-RXAH].

[10] See Angwin et al., supra note 8 ("Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions.").

> almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants.[11]

Northpointe[12] (the company that developed and owned COMPAS) responded to the criticism by arguing that ProPublica was focused on the wrong measure. In essence, Northpointe stressed the point ProPublica conceded—that COMPAS made mistakes with black and white defendants at roughly equal rates.[13] Although Northpointe and others challenged some of the accuracy of ProPublica's analysis,[14] the main thrust of Northpointe's defense was that COMPAS does treat blacks and whites the same. The controversy focused on the manner in which such similarity is assessed. Northpointe focused on the fact that if a black person and a white person were each given a particular score, the two people would be equally likely to recidivate.[15] ProPublica looked at the question from a different angle. Rather than asking whether a black person and a white person with the same score were equally likely to recidivate, it focused instead on whether a black and white person who *did not* go on to recidivate were equally likely to have received a low score from the algorithm.[16] In other words, one measure begins with the score and asks about its ability to predict reality. The other measure begins with reality and asks about its likelihood of being captured by the score.

The easiest way to fix the problem would be to treat the two groups equally in both respects. A high score and low score should mean the same thing for both blacks and whites (the measure Northpointe emphasized), and law-abiding blacks and whites should be equally likely to be mischaracterized by the tool (the measure ProPublica emphasized).

---

[11] Id.

[12] Northpointe, along with CourtView Justice Solutions Inc. and Constellation Justice Systems, rebranded to Equivant in January 2017. Equivant, Frequently Asked Questions 1, http://my.courtview.com/rs/322-KWH-233/images/Equivant%20Customer%20FAQ%20-%20FINAL.pdf [https://perma.cc/7HH8-LVQ6].

[13] See William Dieterich et al., COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity, Northpointe 9–10 (July 8, 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/N5RL-M9RN].

[14] For a critique of ProPublica's analysis, see Anthony W. Flores et al., False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country To Predict Future Criminals. And It's Biased Against Blacks.", 80 Fed. Prob. 38 (2016).

[15] See Dieterich et al., supra note 13, at 9–11.

[16] See Angwin et al., supra note 8 ("In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.").

Unfortunately, this solution has proven impossible to achieve. In a series of influential papers, computer scientists demonstrated that, in most circumstances, it is simply not possible to equalize both measures.[17] The reason it is impossible relates to the fact that the underlying rates of recidivism among blacks and whites differ.[18] When the two groups at issue (whatever they are) have different rates of the trait predicted by the algorithm, it is impossible to achieve parity between the groups in both dimensions.[19] The example discussed in Part I illustrates this phenomenon.[20] This fact gives rise to the question: in which dimension is such parity more important and why?

These different measures are often described as different conceptions of fairness.[21] This is a mistake. The measure favored by Northpointe is relevant to what we ought to *believe* about a particular scored individual. If a high-risk score means something different for blacks than for whites, then we do not know whether to believe (or how much confidence to have) in the claim that a particular scored individual is likely to commit a crime in the future. The measure favored by ProPublica relates instead to what we ought to *do*. If law-abiding blacks and law-abiding whites are

[17] See, e.g., Richard Berk et al., Fairness in Criminal Justice Risk Assessments: The State of the Art, Soc. Methods & Res. OnlineFirst 1, 23 (2018), https://journals.sagepub.com/doi/-10.1177/0049124118782533 [https://perma.cc/GG9L-9AEU] (discussing the required trade-off between predictive accuracy and various fairness measures); Alexandra Chouldechova, Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments, 5 Big Data 153, 157 (2017) (demonstrating that recidivism prediction instruments cannot simultaneously meet all fairness criteria where recidivism rates differ across groups because its error rates will be unbalanced across the groups when the instrument achieves predictive parity); Jon Kleinberg et al., Inherent Trade-Offs in the Fair Determination of Risk Scores, 67 LIPIcs 43:1, 43:5–8 (2017), https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf [https://perma.cc/S9DM-PER2] (demonstrating how difficult it is for algorithms to simultaneously achieve the fairness goals of calibration and balance in predictions involving different groups).

[18] See Bureau of Justice Statistics, U.S. Dep't of Justice, 2018 Update on Prisoner Recidivism: A 9-Year Follow-up Period (2005–2014) 6 tbl.3 (2018), https://www.bjs.gov/-content/pub/pdf/18upr9yfup0514.pdf [https://perma.cc/3UE3-AS5S] (analyzing rearrests of state prisoners released in 2005 in 30 states and finding that 86.9% of black prisoners and 80.9% of white prisoners were arrested in the nine years following their release); see also Dieterich et al., supra note 13, at 6 ("[I]n comparison with blacks, whites have much lower base rates of general recidivism . . . ."). Of course, the data on recidivism itself may be flawed. This consideration is discussed below. See infra text accompanying notes 33–37.

[19] This is true unless the tool makes no mistakes at all. Kleinberg et al., supra note 17, at 43:5–6.

[20] See infra Section I.A.

[21] For example, Berk et al. consider six different measures of algorithmic fairness. See Berk et al., supra note 17, at 12–15.

not equally likely to be mischaracterized by the score, we will not know whether or how to use the scores in making decisions. If we are comparing a measure that is relevant to what *we ought to believe* to one that is relevant to *what we ought to do*, we are truly comparing apples to oranges.

This conclusion does not straightforwardly suggest that we should instead focus on the measure touted by ProPublica, however. A sophisticated understanding of the significance of these measures is fast-moving and evolving. Some computer scientists now argue that the lack of parity in the ProPublica measure is less meaningful than one might think.[22] The better way to understand the measure highlighted by ProPublica would be to say that it *suggests* that something is likely amiss. Differences in the ratio of false positive rates to false negative rates indicate that the algorithmic tool may rely on data that are themselves infected with bias or that the algorithm may be compounding a prior injustice. Because these possibilities have normative implications for how the algorithm should be used, this measure relates to fairness.

The most promising way to enhance algorithmic fairness is to improve the accuracy of the algorithm overall.[23] And we can do that by permitting the use of protected traits (like race and sex) within the algorithm to determine what other traits will be used to predict the target variable (like recidivism). For example, housing instability might be more predictive of recidivism for whites than for blacks.[24] If the algorithm includes a racial classification, it can segment its analysis such that this trait is used to predict recidivism for whites but not for blacks. Although this approach would improve risk assessment and thereby lessen the inequity highlighted by ProPublica, many in the field believe this approach is off the table because it is prohibited by law.[25] This is not the case.

The use of racial classifications only sometimes constitutes disparate treatment on the basis of race and thus only sometimes gives rise to strict

[22] See Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (arXiv, Working Paper No. 1808.00023v2, 2018), http://arxiv.org/abs/1808.00023 [https://perma.cc/ML4Y-EY6S].

[23] See Sumegha Garg et al., Tracking and Improving Information in the Service of Fairness (arXiv, Working Paper No. 1904.09942v2, 2019), http://arxiv.org/abs/1904.09942 [https://perma.cc/D8ZN-CJ83].

[24] See Sam Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness, 2017 Proc. 23d ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining 797, 805.

[25] See id. ("[E]xplicitly including race as an input feature raises legal and policy complications, and as such it is common to simply exclude features with differential predictive power.").

scrutiny. The fact that some uses of racial classifications do not constitute disparate treatment reveals that the concept of disparate treatment is more elusive than is often recognized. This observation is important given the central role that the distinction between disparate treatment and disparate impact plays in equal protection doctrine and statutory anti-discrimination law. In addition, it is important because it opens the door to more creative ways to improve algorithmic fairness.

The Article proceeds as follows. Part I develops the conceptual claim. It shows that the two most prominent types of measures used to assess algorithmic fairness are geared to different tasks. One is relevant to belief and the other to decision and action. This Part begins with a detailed explanation of the two measures and then explores the factors that affect belief and action in individual cases. Turning to the comparative context, Part I argues that predictive parity (the measure favored by Northpointe) is relevant to belief but not directly to the fair treatment of different groups.

Part II makes a normative claim. It argues that differences in the ratio of false positives to false negatives between protected groups (a variation on the measure put forward by ProPublica) suggest unfairness, and it explains why this is so. This Part begins by clarifying three distinct ways in which the concept of fairness is used in the literature. It then explains both the normative appeal of focusing on the parity in the ratio of false positives to false negatives and, at the same time, why doing so can be misleading. Despite these drawbacks, Part II argues that the disparity in the ratio of false positive to false negative rates tells us something important about the fairness of the algorithm.

Part III explores what can be done to diminish this unfairness. It argues that using protected classifications like race and sex within algorithms can improve their accuracy and fairness. Because constitutional anti-discrimination law generally disfavors racial classifications, computer scientists and others who work with algorithms are reluctant to deploy this approach. Part III argues that this reluctance rests on an overly simplistic view of the law. Focusing on constitutional law and on racial classification in particular, this Part argues that the doctrine's resistance to the use of racial classifications is not categorical. Part III explores contexts in which the use of racial classifications does not constitute disparate treatment on the basis of race and extracts two principles from these examples. Using these principles, this Part argues that the use of protected classifications within algorithms may well be permissible. A conclusion follows.

I. PREDICTIVE PARITY AND BELIEF: THE CONCEPTUAL CLAIM

Scholars describe the dilemma as one that pits different conceptions of fairness against each other.[26] One could therefore go on to ask which measure better comports with what fairness requires. This question is answered, at least in part, by recognizing that the measures are geared to different tasks.

A. The Measures and What They Measure

To begin, it will be helpful to get a clear idea of what exactly the relevant "fairness" measures are and why it is impossible to equalize both. In order to explain this to a non-technical audience, I will present a contrived example that exhibits the relevant properties of the COMPAS controversy so that the reader can see and understand each of the measures. In the example I propose, I imagine that there are two social groups in the society: the Greens and the Blues.

> *The Case of the Disease Test*: Suppose there is a medical test used to determine who is sick with a given disease. The test does not perfectly report who is sick and who is not but is reasonably reliable for both the Blues and the Greens, as depicted below. Table 1-1 below represents the results for the Greens. The actual outcome (noted as sick or healthy) is represented in the columns and the predicted outcome (noted as positive/+ or negative/-) is represented in the rows.

TRUE OUTCOME				TRUE OUTCOME			
		Sick	Healthy			Sick	Healthy
TEST	+	60 [a]	20 [b]	TEST	+	16 [a]	5 [b]
RESULT	–	6 [c]	14 [d]	RESULT	–	22 [c]	57 [d]
Table 1-1 (Greens)				Table 1-2 (Blues)			

In the case of the Greens, 60 of the 100 who took the test had a positive test result and are in fact sick. These are the true positives. 20 of the 100 who took the test got a positive test result but are not sick. These are the false positives. 6 of the 100 who took the test got a negative test result

[26] See, e.g., Kleinberg et al., supra note 17, at 43:5.

despite the fact that they are in fact sick. These are the false negatives. And 14 of the 100 who took the test got a negative test result and are not sick. These are the true negatives.

Based on this data, the probability that a Green person is sick if she has tested positive for the disease is (a/(a+b), 60/(60+20)) or .75. Call this the *positive predictive value* or PPV. The probability that a Green is healthy if she tests negative for the disease is (1- (c/(c+d)), 1- (6/(6+14))) or .7. Call this the *negative predictive value* or NPV.

Compare these results to those of the other socially salient group in this society, the Blues. As Table 1-2 indicates, 16 of the 100 Blues who took the test got a positive result and are sick (true positives). 5 of the 100 Blues got a positive result and are not sick (false positives). 22 of the 100 Blues got a negative result even though they are sick (false negative), and 57 of the 100 Blues got a negative result and are healthy (true negative). The probability that a Blue person is sick if she has a positive test result is 16/(16+5) = .76, as the shaded boxes in Table 1-2 illustrate. Thus, the PPV for the Blues is very similar to that for the Greens. And the probability that a Blue person is healthy if she has a negative test result is 1- (22/(22+57)) = .72. The NPV for the Blues is roughly equivalent to that of the Greens. The test thus makes equally accurate predictions, approximately, for the Blues and the Greens.

Yet, if we ask a different question, these tables reveal something different. Rather than ask what the probability is that a Blue or Green person is sick, given her test result, we might ask instead what the probability is that a sick Blue or a sick Green will get an accurate (i.e. positive) test result. The shaded boxes in tables 2-1 and 2-2 below highlight this question.

TRUE OUTCOME				TRUE OUTCOME			
		Sick	Healthy			Sick	Healthy
TEST	+	60 [a]	20 [b]	TEST	+	16 [a]	5 [b]
RESULT	–	6 [c]	14 [d]	RESULT	–	22 [c]	57 [d]
Table 2-1 (Greens)				Table 2-2 (Blues)			

For a sick Green who takes the test, the probability that she will get an accurate, positive result is 60/(60+6) = .91 For a sick Blue who takes the test, the probability that she will get an accurate, positive result is quite different: 16/(16+22) = .42. We get dissimilar results as well when we compare what happens to healthy Greens and healthy Blues who take the test. For a healthy Green who takes the test, the test accurately provides a negative test result in 14 of the 34 cases, or 41% of the time. Whereas for a healthy Blue who takes the test, the test accurately reports a negative result in 57 out of 62 cases, or 91% of the time.

This simple example does not quite replicate the situation described in the ProPublica exposé but is close enough to illustrate the tension between the two measures.[27] The test is (approximately) equally accurate in predicting health for the Greens and Blues. If a Blue or a Green get a positive result, that result is accurate approximately 75% of the time. Yet the errors are of very different types. For the Greens, a sick person is highly likely to get a correct result, but a healthy person is not. Another way to put this point would be to say that the false positive rate is high for the Greens and higher than the false negative rate for Greens. Contrast that result with the situation for the Blues. For the Blues, a healthy person is highly likely to get an accurate test result whereas a sick Blue is not so fortunate. For the sick Blue, the test only gives the correct answer in 42% of cases. For the Blues, therefore, the false negative rate is high and is much higher than the false positive rate.

The basic point is this. The test is equally accurate for Blues and Greens. But, when errors occur, the types of errors that occur are different. For Greens the errors are more likely to be false positives and for Blues the errors are more likely to be false negatives.

In what follows, I will use these numbers and tables—which in the literature are called "confusion tables"[28]—to refer both to the medical example described above and to apply to a situation in which the same data is used to determine who should be released on parole. I use the same data for a hypothetical parole example to keep things as simple as I can, given the complexity of the underlying issue. To translate the confusion tables for that context, we would say that the test is a risk assessment algorithm which scores people as either high or low risk (high risk = positive, low risk = negative) and that rather than sick and healthy, the person actually recidivates (sick) or does not (healthy). To make the

[27] COMPAS did not use a binary scoring mechanism like the positive or negative result in the example in the text. Instead, people were given a risk score of 4 or 8, for example, which indicates how dangerous they are predicted to be relative to the scored group. Northpointe, Practitioner's Guide to COMPAS Core 8–11 (Mar. 19, 2015), http://www.northpointeinc.com/downloads/compas/Practitioners-Guide-COMPAS-Core-_031915.pdf [https://perma.cc/74D9-ET8T].

[28] See Berk et al., supra note 17, at 4 (explaining that "a cross-tabulation of the actual binary outcome Y by the predicted binary outcome $\hat{Y}$" is called, within the field of machine learning, a "confusion table" or a "confusion matrix").

Green/Blue example analogous to the dispute about COMPAS, the Greens would be blacks and the Blues would be whites. For blacks who will recidivate, the test accurately predicts that result 91% of the time. For blacks who will not, the test's accuracy falls to 41%. The results for whites (the Blues) are almost reciprocal. For those who will recidivate, the test is only accurate 42% of the time, but for those who will not, the test accurately yields that prediction in 91% of cases. Yet, as with the disease case, for both blacks and whites, a risk score of high risk is approximately 75% accurate for each group. Let me reiterate: my use of this data in an example dealing with parole decisions is entirely fabricated. I use it for purposes of exposition because it shares the same structure as the COMPAS example.

TRUE OUTCOME				TRUE OUTCOME			
SCORE		Will Recidivate	Will Not Recidivate	SCORE		Will Recidivate	Will Not Recidivate
	High Risk	60 [a]	20 [b]		High Risk	16 [a]	5 [b]
	Low Risk	6 [c]	14 [d]		Low Risk	22 [c]	57 [d]
Table 3-1 (Blacks)				Table 3-2 (Whites)			

Does this hypothetical risk assessment tool treat blacks fairly as compared to how it treats whites? The best response to the ProPublica exposé would be to adjust the algorithm so as to treat blacks and whites equally in both dimensions. However, this is impossible except under highly specific circumstances that are likely to be rare in practice.[29] As Kleinberg and co-authors explain: "Our main result . . . is that these conditions [equalizing both measures] are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases."[30]

As this hypothetical illustrates, when base rates[31] of some trait differ between two groups, it will be impossible to equalize both measures. In the disease hypothetical, the Greens are sicker than the Blues (66% of Greens are sick while only 38% of Blues are). Similarly, when that

[29] See supra note 17 and accompanying text.
[30] Kleinberg et al., supra note 17, at 43:3.
[31] The term "base rate" refers to the rate at which the condition occurs in the relevant population.

hypothetical case is used to illustrate the problem in the recidivism context, the base rate for recidivism is different for blacks as compared to whites, meaning that more blacks actually will recidivate than will whites (if these data are accurate). This is also the case in the data relied on by Northpointe.[32] In my hypothetical, I suppose these base rates differ quite substantially in order to use the same tables as in the disease example and to illustrate in fairly stark terms how the difference in base rates gives rise to an inability to equalize both measures.

One caveat is important to note before proceeding. The data that establish the base rate could themselves be unreliable and indeed could be inaccurate in predictable and biased ways. The base rate data about recidivism do not—and indeed cannot—report *actual* recidivism because researchers do not have access to this information. Instead they report arrests.[33] If policing practices make it the case that blacks who *actually* recidivate are more likely to be arrested than are whites who *actually* recidivate, then the reported base rates will not reflect the trait they purport to measure and thus should be viewed skeptically.[34] This is a point made frequently by critics of algorithms and of the data on which they are trained.[35] This problem, called "measurement error" in the computer

[32] See Dieterich et al., supra note 13, at 6 (noting that "in comparison with blacks, whites ha[d] much lower base rates of general recidivism (0.39 vs. 0.51) and violent recidivism (0.09 vs. 0.14)" in the study's main sample).

[33] See Bureau of Justice Statistics, supra note 18, at 3; see also Sandra G. Mayson, Dangerous Defendants, 127 Yale L.J. 490, 562 (2018) ("Pretrial risk assessment tools should instead measure crime risk in terms of the likelihood of rearrest for a serious violent crime in the pretrial phase.").

[34] Some scholars suggest that the algorithms should be trained on data on rearrests for violent crimes only because this data is less likely to be skewed by biased policing practices. See, e.g., Mayson, supra note 33, at 562 (discussing why pretrial risk assessment tools should assess whether a person will commit a serious violent crime, not just any crime).

[35] See, e.g., Pauline T. Kim, Auditing Algorithms for Discrimination, 166 U. Pa. L. Rev. Online 189, 191 (2017) (arguing that algorithms in general should be audited for bias "because the causes of bias often lie not in the code, but in broader social processes"); Abigail Z. Jacobs & Hanna Wallach, Measurement and Fairness, ACM Conference on Fairness, Accountability, and Transparency, FAT*, at 8 (2019) (emphasizing the gap that exists between a complex trait that is difficult to measure and the proxy trait that is used to capture it and the ways in which this disparity allows the replication of bias as, for example, "[u]sing previous salary as a measure of quality would replicate, and likely exacerbate, past patterns of inequality, including by race and gender").

science literature,[36] is an important issue and one whose significance will be addressed later.[37]

Measurement error is not a problem that is unique to the context in which automated algorithms or machine learning are used. In a canonical sex discrimination case from the 1970s, Justice Brennan made the same point. In *Craig v. Boren*, men challenged an Oklahoma law that allowed women to purchase low alcohol beer at age 18 but required men to be 21 to purchase the same product.[38] The state defended the law by arguing that young men have higher rates of drunk driving than do young women.[39] Justice Brennan, writing for the Court, found this argument unpersuasive. In his view, data showing that young men are more likely to be arrested for drunk driving than are young women may be unreliable because "'reckless' young men who drink and drive are transformed into arrest statistics, whereas their female counterparts are chivalrously escorted home."[40] Unavoidably, arrest statistics reflect both actual offending rates and policing practices.

The potential for bias in the data upon which both people and machines rely is certainly important[41] and provides a reason to be skeptical about some base rate data.[42] To start, I put this concern aside. In Part II, I return to it and consider how worries about measurement error should inform choices about how to use algorithmic data.

So far, drawing on the ProPublica controversy, I have focused on two measures that could be used to assess whether an algorithm is fair. We could focus on whether the scores produced by the algorithm are equally

[36] See, e.g., Sharad Goel et al., The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment 7 (Dec. 26, 2018) (unpublished manuscript), https://ssrn.com/abstract=3306723 [https://perma.cc/ZT5S-J4XK].

[37] See infra Section II.D.

[38] 429 U.S. 190, 191–92 (1976).

[39] Id. at 200–01.

[40] Id. at 202 n.14.

[41] See supra note 34 and accompanying text. For a detailed analysis of the many ways in which the Fourth Amendment to the U.S. Constitution, as understood currently, permits racial profiling by the police, see Devon W. Carbado, From Stopping Black People to Killing Black People: The Fourth Amendment Pathways to Police Violence, 105 Calif. L. Rev. 125 (2017). Because policing blacks more heavily than others contributes to bias in the base rate data, racial profiling is relevant to the reliability of data.

[42] For a discussion of whether nondiscrimination norms require a skepticism about base rate data involving protected groups beyond what good epistemic practice requires, see Deborah Hellman, The Epistemic Commitments of Nondiscrimination (Va. Pub. Law & Legal Theory Paper Series, Paper No. 60, 2018), https://ssrn.com/abstract=3273582 [https://perma.cc/G3JS-82BD].

predictive for each group or we could focus on whether the error rates produced by the algorithm are equal. These are not the only measures that are offered as tests of fairness in the technical literature.[43] But, for simplicity, and because the heart of the controversy appears to focus on these two measures, I begin my discussion with them.

Different scholars use different names to describe these two measures (or variants of them).[44] Alexandra Chouldechova uses the term "predictive parity" to describe the situation in which a black person and white person with the same score are equally likely to recidivate.[45] I find that term accessible and useful. However, it focuses only on positive predictive value (PPV). We could focus on whether the PPV *and* NPV are both equal for the two groups at issue. Where both are equal, we can call this *equal predictive value* or EPV. In this Article, I will use the terms equal predictive value (EPV) or predictive parity to capture the first of the potential measures. While there are differences between these two terms, for the most part I gloss over them. The first measure focuses on whether the score is equally accurate for the two groups, but EPV is more demanding than predictive parity. In my hypothetical, the disease test T exhibits predictive parity for Greens and Blues. Similarly, my hypothetical recidivism algorithm (using the same numbers) has predictive parity for blacks and whites. In my example, the NPV is also roughly equal for the two groups, so these examples also exhibit EPV.

---

[43] See Berk et al., supra note 17, at 12–15.

[44] For example, Jon Kleinberg and co-authors characterize the property of equal accuracy of the score across groups as "calibration within groups" and define it as the condition when risk assignment scores "mean what they claim to mean, even when considered separately in each group." Kleinberg et al., supra note 17, at 43:5. More formally, they define calibration within groups in this way:

> *Calibration within groups* requires that for each group *t*, and each bin *b* with associated score v_b, the expected number of people from group *t* in *b* who belong to the positive class should be a v_b fraction of the expected number of people from group *t* assigned to *b*.

Id. Richard Berk and co-authors call this feature "conditional use accuracy equality." Berk et al., supra note 17, at 14–15. They explain this concept by asking the following question: "Conditional on the prediction of success (or failure), is the projected probability of success (or failure) the same across protected group classes?" Id. at 15. Sharad Goel and co-authors call it simply "calibration." Goel et al., supra note 36, at 9 (defining "calibration" as the requirement that "outcomes are independent of protected attributes after controlling for estimated risk").

[45] See Chouldechova, supra note 17, at 155 (defining predictive parity as follows: "A *score* $S=S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership").

Alternatively, we could equalize the error rates. Scholars also have different terms for the situation in which these are equal. For example, Jon Kleinberg and his co-authors use the terms "balance for the positive class" and "balance for the negative class" to indicate when the false positive and false negative rates are the same for each group.[46] Chouldechova uses the term *error rate balance*,[47] a term which I find most accessible and so will adopt in this Article.[48]

To summarize, algorithms are used to predict some endpoint of interest—sickness, recidivism, or a multitude of other possible traits. These algorithms generally avoid the use of classifications that are protected by anti-discrimination law, like race or sex.[49] However, when the groups defined by protected traits have different base rates of the target trait, it will be impossible to have parity between the groups along all the possible dimensions of fairness. We have focused on two of those dimensions. The algorithm can exhibit equal predictive value such that scores will be equally predictive of the target trait for members of one group as for members of the other. Or, the algorithm can exhibit error rate

[46] See Kleinberg et al., supra note 17, at 43:3. They define "balance for the negative class," for example, as follows:

> [A] violation [of this condition] . . . would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.

Id. at 43:5. Berk and his co-authors call this "conditional procedure accuracy equality," Berk et al., supra note 17, at 14 (explaining that this measure is the "the same as considering whether the false negative rate and the false positive rate, respectively, are the same for African Americans and whites"), and Goel et al. call it "classification parity," Goel et. al., supra note 36, at 9 (defining "classification parity" to mean that "certain common measures of predictive performance (like false positive or negative rates) be equal across groups defined by the protected attributes").

[47] See Chouldechova, supra note 17, at 155 (defining "error rate balance" in the following way: "A score $S=S(x)$ satisfies error rate balance at a threshold s_{HR} if the false positive and false negative error rates are equal across groups.").

[48] Kleinberg's terminology focuses on cases that are non-binary, see Kleinberg et al., supra note 17, at 43:7, and Chouldechova's on binary terms, see Chouldechova, supra note 17, at 161.

[49] See generally Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 Calif. L. Rev. 671, 677–80 (2016) (emphasizing the fact that "data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes" despite not explicitly classifying on these bases); Kroll et al., supra note 5, at 685 (explaining that a "commonly understood way to demonstrate that a decision process is independent of sensitive attributes is to preclude the use of those sensitive attributes from consideration" but insisting that such an approach is "naive").

balance such that people of each group who have or lack the target variable are equally likely to be accurately scored by the test.

B. Predictive Accuracy and Belief

The fact that we cannot have both equal predictive value and error rate balance in most circumstances leads to the question: which should we prefer and why? That question focuses on whether equal predictive accuracy or equal rates of false positives or false negatives is more important. Before we tackle that question, it is helpful to step back and focus on the epistemic and practical significance of both accuracy and the type of error (false positive or false negative) in individual cases, where no comparative question is on the table. We need to know how we might fail to treat blacks and whites the same.

1. Individual Cases

If a test or algorithm has a high degree of predictive accuracy, it provides us with information. If a positive test result is correct 99% of the time, then it provides help in answering the following question: given this evidence (the test result), what should I believe? In the example just described, I should believe what the test predicts to be the case. A high degree of predictive accuracy does not, however, tell us how to act. To see why, consider the following example:

> *Leslie, the Baby, and the Bat*: One day, Leslie found a live bat in her house when her daughter was a baby. Although the bat eventually left her house, Leslie's pediatrician nonetheless recommended treating her young daughter with rabies shots. Why? While the doctor thought it unlikely that the baby had been bitten by the bat without the baby waking and crying out, and also thought it unlikely that the bat had rabies (as few do), still the doctor recommended treatment because rabies is fatal if not treated very soon after exposure. If the doctor were to put a percentage to the likelihood that the girl had rabies, it would have been extremely low. However, because the cost of a false negative judgment was so high (not treating someone who has contracted rabies leads to death), the doctor recommended treatment.

As this example illustrates, what we ought to believe (the baby does not have rabies) and what we ought to do (treat the baby for rabies) are affected by different considerations.[50] For Leslie and her baby, the cost of acting on a false negative assessment is so high that it makes practically no difference whether the doctor's belief that the baby does not have rabies is highly likely to be true. Decisions about what to do depend crucially on the costs of errors, as this example shows.

Different types of errors have different costs. What the costs are for each of the errors we might make (the false positive and the false negative) affect what we ought to do. Consider another example:

> *Different Legal Standards*: John is arrested and tried for punching Bill in the nose. The evidence presented at trial supports the proposition that John punched Bill. Sue is a member of the jury that hears the evidence. Sue believes that John punched Bill but isn't certain. Her level of confidence in the truth of the proposition that John punched Bill is 75%.

Is this level of confidence sufficient for Sue to vote to hold John responsible for this assault? It depends. If John is being tried for the crime of assault, Sue should vote to acquit. Sue's level of confidence in her belief that John punched Bill is insufficient to meet the legal standard required in a criminal case because, in order to support John's assault conviction, she must believe beyond a reasonable doubt that John punched Bill. By contrast, if Bill is suing John for the tort of assault (a civil claim), Sue should vote to find John liable. In a civil case, a juror must only believe that it is more likely than not that John punched Bill to find him liable for assault, and Sue has sufficient confidence in her belief that he did.

What explains the difference between the criminal and civil context is the cost of mistakes in each context.[51] The very high burden of proof in a criminal case reflects society's judgment that the cost of a false positive (convicting an innocent) is extremely high and much higher than the cost of a false negative (letting a guilty person go free). By contrast, in the civil case, the cost of a false positive (holding an innocent person liable) is approximately the same as the cost of a false negative (failing to hold a

[50] Some philosophers argue that moral and pragmatic considerations also affect the beliefs a person should hold. See, e.g., Rima Basu, The Wrongs of Racist Beliefs, 176 Phil. Stud. 2497, 2513–14 (2018) (arguing that rational racist beliefs can wrong a person); Michael Pace, The Epistemic Value of Moral Considerations: Justification, Moral Encroachment, and James' 'Will To Believe,' 45 Noûs 239, 241, 251 (2011) (arguing that pragmatic reasons properly affect the choice whether to care more about avoiding false beliefs or acquiring true beliefs).

[51] I use the term "cost" here broadly so that it includes not only monetary costs but also personal costs and moral costs.

guilty person liable). As a result, the burden of proof is much lower in the civil context. The point to emphasize about these two contexts is this: a person on a jury could have the same degree of confidence in the accuracy of the claim that John punched Bill in both the criminal and civil trial yet still do different things (vote to acquit, vote to hold liable) because of the stakes. What we believe is a function of the evidence; what we do is a function of what we believe and the stakes of acting on our beliefs if they turn out to be mistaken.[52]

If we lose some degree of predictive accuracy, what else will we lose? Faced with the score produced by a test or algorithm, we will not know precisely what to believe, as the significance of the test or score will be lessened. Loss of predictive accuracy compromises knowledge or, to be more precise, we lose confidence in the information provided by the algorithm.[53]

The type of error we might make matters when assessing how we ought to act. As the famous Blackstone ratio expresses,[54] the two types of errors we might make are often not of equivalent significance. It is better that ten guilty men are freed than that one innocent is wrongly convicted. In other words, a false positive matters ten times as much as a false negative in the criminal context. In the civil context, the errors are of roughly the same weight. It is for this reason that the burden of proof is so much higher in the criminal than the civil context.

The two examples—*Leslie, the Baby, and the Bat* and *Different Legal Standards*—illustrate two points about the relationship between predictive accuracy and action. Accurate belief is sometimes not necessary in order to decide how to act, as the bat example demonstrates. Even if the doctor is uncertain about how likely it is that the baby has rabies, she nonetheless knows how to act. In addition, accurate belief is sometimes not sufficient to know how to act either, as the example of *Different Legal Standards* makes clear. Even when we know precisely how likely it is that John punched Bill, we do not know what to do without

[52] Again, some philosophers believe that the cost of error is relevant to belief as well. See, e.g., Pace, supra note 50, at 257. If they are correct, that only strengthens the claim that I argue for here, i.e., that error rate balance should be prioritized over predictive parity.

[53] Another way to express this idea is to say that our "credence" is lowered. See infra note 55 and accompanying text.

[54] See 4 William Blackstone, Commentaries *358 ("[F]or the law holds, that it is better that ten guilty persons escape, than that one innocent suffer.").

making a normative judgment about how to weigh each type of error against each other, a weighing implicit within each legal standard.

I do not want to overstate the point. Clearly accurate beliefs are often important for decision and action. In fact, Part III will argue for an approach that increases accuracy. Rather, my goal in this Part to is get a better handle on how and why predictive accuracy matters in order to better understand the significance of a lack of parity in this dimension. It is to that question that we now turn.

2. Comparative Cases

With a clearer sense of the significance of accuracy in the individual case, we can now ask about the comparative context. When we lack equal predictive value, do we thereby compromise fairness between blacks and whites scored by the algorithm?

Return to the disease example to explore this question. The screening test in this hypothetical is approximately 75% accurate for both the Greens and the Blues. If a physician tests a patient and gets a positive result, she has reason to be fairly confident that the patient has the disease (and this is so even if she is unable to know whether the person is a Green or a Blue). More precisely, and to borrow a philosophical term, the doctor has a credence of .75 in the proposition that the patient has the disease.[55] Since the test exhibits predictive parity, it is equally accurate for Greens as for Blues. Why is this important? Most obviously, it allows the doctor to know how confident to be in the test even if she is unaware or unable to know the "color" (Green or Blue) of the person involved. In other words, parity in the predictive accuracy of the result provides information value when we can't (for practical or legal reasons) distinguish between or among groups.

This is unsurprising. As predictive accuracy relates directly to belief, so too does parity of predictive accuracy. But what of fairness? Does a lack of predictive parity compromise fairness? Without predictive parity, the scores that members of each group receive are not equally meaningful. Does the fact that the test is more accurate for one group than for another mean that it is unfair? To explore this question, consider the following example:

[55] The term "credence" is one used by epistemologists. For example, Sarah Moss defines credences as "subjective probabilities measured on a scale from 0 to 1." Sarah Moss, Probabilistic Knowledge 2 (2018).

> *Pedagogical Choice*: A professor must decide what type of exam to give to her students. Suppose that she can choose all essay questions or all multiple-choice questions or some combination thereof. Suppose further that with an exam of all essay questions, the exam will do a better job reflecting the actual knowledge of men than it will do of women and that for an exam of all multiple-choice questions, the reverse is true. The professor chooses to have 75% multiple choice questions and 25% essay questions.[56]

In *Pedagogical Choice*, the grade on the test means something different for female test-takers than it does for male test-takers. In particular, the exam is a more reliable indicator of actual knowledge for women than for men. We can now pose the question we are interested in: has either group been treated unfairly? It is hard to answer that question without knowing more. The test is less accurate for men, but in what way is it less accurate? Does it give men better scores than they deserve, worse scores than they deserve, or does it skew equally in both directions? Surely this information matters in assessing whether the test is fair to men. In other words, it is not the fact that the test isn't equally accurate for men and women that matters to fairness, it is how the inaccuracy operates.

But isn't there some unfairness in being judged by a less accurate measure than is applied to another group? I hear the voices of studious law students in my head asking this question. Suppose that for male students, the test is a less accurate indicator of knowledge than it is for female students but that the manner in which it is less accurate is that it produces more false positives—i.e. more men who don't know the material well get good grades. In one sense, men are benefited by this loss of predictive accuracy. But in another sense, they are harmed. For the

[56] If the professor makes this choice in order to disadvantage one group or another, then this will likely be legally problematic because intentions are relevant under current anti-discrimination law. See infra note 95; see also Richard H. Fallon, Jr., Constitutionally Forbidden Legislative Intent, 130 Harv. L. Rev. 523 (2016) (mapping the various ways in which intention matters in constitutional law, across several doctrines, and arguing that given the confusion in the doctrine, intention ought not to matter to constitutional law). Whether intentions matter to permissibility from a moral perspective is controversial. Micah Schwartzman believes intention should matter. See Micah Schwartzman, Official Intentions and Political Legitimacy: The Case of the Travel Ban 2 (Va. Pub. Law & Legal Theory Paper Series, Paper No. 2018-22), https://papers.ssrn.com/so113/abstract=3159393 [https://perma.cc/UT3E-NDAW] (arguing that intentions should be relevant to the permissibility of governmental action). In my view, intentions should not matter to permissibility in the context of assessing discrimination. See Deborah Hellman, When Is Discrimination Wrong? 138–68 (2008).

well-prepared male student who would have done well on either sort of exam, he loses the ability to distinguish himself from other male test-takers who do as well, even though they know less. This individual man is surely harmed by the fact that the test is less accurate for men than for women. But, assuming that we are unable to know whether a particular test-taker is a man or a woman (which is the assumption that gives rise to the dilemma we are exploring), then prepared female test-takers, who are also inappropriately grouped together with less prepared male test-takers, are also unable to separate themselves from these less well-prepared male test-takers. If this is correct, male test-takers haven't been treated unfairly as compared to female test-takers. Rather, we might say that very prepared test-takers are treated unfairly in being subject to a test that does not separate them from some less prepared test-takers (who happen to be men).

This claim of unfairness has a different character altogether. It isn't a claim about unfairness on the basis of sex. Instead, it is a claim that everyone is entitled to be treated by the most accurate test available (or feasible, or imaginable). It is a claim that another test could have done a better job of identifying and stratifying the best from the very good, the good, etc. This is not a claim about whether one group of test-takers is being treated fairly vis-à-vis another group of test-takers. In fact, it isn't a comparative claim at all.[57] Rather it is a claim to a right to the best available decision-making tool. That this is a good claim—legally or morally—I find doubtful.[58] But what it is not is a claim of unfairness between groups.[59]

Let me summarize the argument of this Part. Predictive accuracy provides information that informs belief. As the first two hypothetical examples (the bat and the legal standards) demonstrate, this information is neither necessary nor sufficient to tell one how to act. Given the

[57] For a discussion of the difference between comparative and non-comparative conceptions of justice and how they relate to claims of wrongful discrimination, see Deborah Hellman, Two Concepts of Discrimination, 102 Va. L. Rev. 895 (2016).

[58] See Hellman, supra note 56, at chs. 4–5.

[59] There may be growing interest in a measure of individual fairness among computer scientists. See, e.g., Cynthia Dwork et al., Fairness Through Awareness, 2012 Proc. 3d Innovations Theoretical Computer Sci. Conf. 214, 214–15 (describing a framework to achieve fairness in classification that includes measuring "individual-based fairness") (emphasis omitted). I believe this approach can capture the non-comparative idea that people should be treated rationally but not the fairness-based concern with treatment as an equal. See Hellman, supra note 57, at 902, 933.

relationship between predictive accuracy and belief and predictive accuracy and action, how should we think about the significance of equal predictive accuracy? Equal predictive accuracy is important because it tells us how much confidence to have in a test or score in those contexts in which we do not or cannot know to which group a person belongs. In other words, equal predictive accuracy also relates primarily to questions of belief and not to questions of action. A lack of predictive parity might nevertheless be unfair. While I concede that tests that are more accurate for one group than for another could constitute a form of unfairness, it matters how that inaccuracy operates. Giving members of a group better scores than they deserve may well be less morally troubling than the reverse. *Pedagogical Choice* demonstrates the fact that fairness is more closely tied to this sort of question than to accuracy pure and simple.

II. ERROR RATES AND FAIRNESS: THE NORMATIVE CLAIM

In the last Section, we saw that a lack of predictive parity primarily affects belief. If an algorithm lacks predictive parity, then we cannot know precisely what to believe about a scored individual. At the same time, lack of predictive parity only indirectly affects action. While there can be unfairness in the fact that information about one group is less accurate or meaningful than about another, the unfairness that most commentators seem interested in is less theoretical and more practical. In this Part, I argue that a difference in the ratio of false positives to false negatives between legally protected groups is suggestive of this practical sense of unfairness. Before presenting this argument, it will be helpful to clarify several different ways that the term "fairness," in its more practical sense, might be used.

A. Fairness Three Ways

The concept of fairness can be used in several ways and can refer to many different normative ideas. This presents potential problems, as scholars and commentators discussing algorithmic "fairness" may use that term to refer to different ideas. It will thus be helpful to clarify some of the broad conceptual distinctions that divide this moral landscape. The first conceptual distinction is between a comparative and non-comparative conception of fairness.[60] The comparative conception of

[60] See Hellman, supra note 57.

fairness examines whether *X* was treated fairly, as compared to how *Y* was treated, where *X* and *Y* can be either individuals or groups. Was John treated fairly, given how Jane was treated? Were men treated fairly, given how women were treated? Were blacks treated fairly vis-à-vis whites? By contrast, a non-comparative conception of fairness asks whether *X* is treated as she ought to be treated, without regard to how any other person or group is treated. In the non-comparative conception of fairness, we compare *X*'s treatment to some standard but not to the treatment of any other actual or hypothetical people.

If our focus is algorithmic fairness, the comparative conception of fairness can be further divided into two sub-types. One can ask whether individuals or groups scored by the algorithm are treated fairly as compared to others who are also scored by the algorithm. This is the kind of fairness identified by ProPublica.[61] They asked whether blacks scored by the algorithm were treated fairly as compared to whites who were scored by the algorithm. Alternatively, one could focus on both people scored by the algorithm and people affected by this scoring practice. For example, if the algorithm is used in the criminal justice context, one might ask whether it treats potential crime victims fairly as compared to how it treats scored individuals. Or one might ask whether the algorithm treats blacks fairly as compared to whites but include both blacks and whites scored by the algorithm and those not scored but affected by the scoring practice.[62]

In what follows, I focus on the comparative conception of fairness and, in particular, on the comparison between how two protected groups scored by the algorithm are treated vis-à-vis each other.

B. Error Ratio Parity

In this Section I argue that we should focus on whether the ratio between the false positive rate and the false negative rate is the same for relevant groups scored by the algorithm. I call this measure *Error Ratio Parity* or ERP. In what follows, I acknowledge that this measure alone does not determine whether an algorithm is fair or unfair. Still, a lack of

[61] For a complete discussion of ProPublica's perspective, see supra notes 8–16 and accompanying text.

[62] See, e.g., Huq, supra note 7, at 1111 (considering the normatively relevant inquiry to be "whether the costs that an algorithmically driven policy imposes upon a minority group outweigh the benefits accruing to that group").

Error Ratio Parity is importantly suggestive of unfairness when the group at issue is one that has been mistreated in the past, or so I contend.

An algorithm, like any test or procedure, is likely to be imperfect. It does not perfectly predict or report the trait, quality, or state it is designed to identify. A recidivism predictor sometimes predicts a person will recidivate when she will not and predicts a person will not recidivate when she will. Similarly, an exam may yield a high grade for an unprepared student who lacks the relevant knowledge or yield a low grade for a prepared and knowledgeable student. Designers of the algorithm must determine how to weigh the costs of each of these types of errors the test or algorithm could make. This assessment affects how they draw the lines between the categories at issue (high versus low risk; A versus C grades, etc.). The designer of the algorithm must balance the harm of mistakenly giving a knowledgeable student a low grade against the harm of erroneously giving a slacker an A, for example.[63]

There is no one-size-fits-all answer to this question. Sometimes false positives are more costly than false negatives, and sometimes the reverse is true. For example, if the task is to identify potential terrorists at airports, the algorithm's designers are likely to judge the cost of a false positive to be low and the cost of a false negative to be high. If the algorithm picks out someone as a potential terrorist who is not, there is inconvenience and possibly loss of privacy and stigma but no loss of life. If the algorithm fails to identify a terrorist, the costs can be deadly. For that reason, the tool adopted will be likely to have a high false positive rate. It might identify as a potential terrorist anyone with a non-negligible chance of being a terrorist. In order to be certain not to miss any potential terrorist, the algorithm might even select everyone (literally). If this were the upshot, we would hardly need an algorithm, but you see the point. How sensitive the tool should be, and thus how close to this limit, depends in part on the cost of the false negative.

In other contexts, it is the cost of the false positive rather than the false negative that is most concerning. Our procedure for determining who is convicted of a crime provides a good example. Consider, again, the "Blackstone ratio": "[I]t is better that ten guilty persons escape, than that

[63] When the score represents a prediction of future events in the form of a likelihood that the event will occur, it isn't correct to say that the score mistakenly characterizes a person as low risk who does not go on to recidivate. When the score predicts the future rather than representing the present, it is less clear how we should characterize the concepts of Type I and Type II errors. This is an important topic that needs to be addressed.

one innocent suffer."[64] This ratio is arrived at by determining the cost to the community of the risk involved in releasing a guilty and potentially dangerous person as compared to the cost to the individual (as well as to his family and community) of erroneously convicting an innocent. While the costs of releasing a guilty person may be high, it is because the community weighs so heavily the harm of erroneously incarcerating an innocent that this ratio is arrived at.

The different costs of false positives and false negatives explain why we treat airline travelers differently than criminal defendants. We adopt a different rule governing how confident we must be that the person in question has the relevant trait before we take action. We need only a small suspicion that a traveler is a terrorist before we search him; we need to be extremely confident that the defendant is a criminal before we convict. We can express the rule we apply in either of two ways. Rule A might say the following: at a certain level of confidence in the truth of the relevant fact (*T* is a terrorist; *D* is a criminal), take a particular action. Rule B might say: at a particular ratio between the two types of errors, take a particular action. In other words, the ratio between the false positives and false negatives can be understood as another way of articulating the rule applied in each context.

If blacks and whites scored by an algorithm were subject to different rules of the form of Rule A, we would have no doubt that this would constitute disparate treatment on the basis of race. So too, I contend, if blacks and whites scored by the algorithm are subject to different rules of the form of Rule B. This too constitutes disparate treatment on the basis of race.

Consider *Different Legal Standards* again. Suppose that if John is white, the jury can only vote to convict him if they find him guilty beyond a reasonable doubt. If John is black, however, they may convict him if they believe it is more likely than not that John is guilty. If this were the case, we would have no doubt that we have disparate treatment on the basis of race. My claim is that this disparate treatment can be understood in multiple ways. It might stem from the fact that Rule A-type rules of different forms have been applied to different races: the confidence level required for convicting blacks is lower than the level required for convicting whites. But the disparate treatment could also stem from the fact that Rule B-type rules of different forms are used: the defendant-

[64] Blackstone, supra note 54, at *358.

friendly 10:1 Blackstone error ratio applies to whites but not to blacks. In a very real sense, each treatment can be expressed in either form.

To summarize, sometimes we will want to make sure we have very few false negatives (in an algorithm that identifies terrorists, for example). Other times, we will want to make sure that we have very few false positives (as in the Blackstone ratio). These determinations depend on the costs of each type of error, which is in part a function of how we intend to respond to each determination. Keeping someone in jail is a more serious cost to both the individual and to society than is an intrusive search at an airport, for example. One way to think about algorithmic fairness, then, would be to ask whether the algorithm strikes the same balance between (the costs of) false positives versus false negatives for each of the groups scored by the algorithm. The more difficult question is how to assess whether the algorithm does so.

We want to ensure that the algorithm strikes the same balance in the way it weighs false positives as compared to false negatives for the two relevant groups scored by the algorithm.[65] At first blush, lack of error ratio parity seems to indicate that an algorithm does not. With COMPAS, false positives outweigh false negatives for blacks and false negatives outweigh false positives for whites.[66] This is particularly worrisome where, as here, otherwise the contexts are the same. In both contexts, there is a risk in releasing a dangerous person and a harm in failing to release someone who is peaceful.

Fairness between protected groups scored by the algorithm requires that we balance false positives versus false negatives in the same way for each group. What we should not do—to put the point colorfully—is treat blacks like terrorists and whites like Englishmen by weighing false negatives as especially costly for blacks and false positives as especially

[65] In a recent article, Marcello Di Bello and Collin O'Neil argue that the moral concept of equal protection requires that all people face an equal risk of mistaken conviction. See Marcello Di Bello & Collin O'Neil, Profile Evidence, Fairness, and the Risks of Mistaken Convictions, 130 Ethics 147 (2020). My view differs from theirs in that I am only concerned about disparities between protected groups—not ensuring equal risk for all.

[66] See Angwin et al., supra note 8 ("The 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. White defendants were mislabeled as low risk more often than black defendants."); cf. Jeff Larson et al., 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/7F6S-FLCK] (describing how ProPublica analyzed false positives and false negatives in the COMPAS scores obtained from Broward County, Florida).

costly for whites. The reason the ProPublica story about COMPAS was so incendiary is because the algorithm appears to do just this.

C. The Limitations of Error Ratio Parity

This appearance is misleading, however. To set policy, we must first determine how to balance the two types of errors we might make. In the criminal conviction context, for example, Blackstone proposed a balance of ten false negatives (guilty who go free) to one false positive (innocent who is convicted). Once the 10:1 balance is set, error ratio parity demands that it apply equally across different groups. The problem is that this ratio is not specific to any particular subgroup of arrested individuals. To see this, imagine that the police simply arrest the first 100 people they encounter on the street. In this context, one would hope that there will be a very different ratio, one with many more people who go free than 10:1. This is because there is no reason to think that these random people have committed a crime. Similarly, if the police only arrest people they have strong reasons to believe have committed crimes, the ratio of people released to incarcerated will also be different than 10:1. The ratio, as an expression of the balance between the two types of errors, tells us the following: when we have a particular amount of evidence of guilt, set the ratio at 10:1. Thus, for blacks and whites about whom we have the same reason to believe are dangerous, the ratio of false positives to false negatives should be the same. But the error rates depicted in a confusion table show not only what happens to individuals for whom we have the same amount of evidence, but for all individuals (those for whom we have both more and less). As a result, if the information we have indicates that, collectively, one group is more likely to recidivate than the other, more people in that group will be scored as high risk (both correctly and incorrectly). The false positive rate is thus likely to be higher. Therefore, information about error rates alone does not directly tell us whether we are balancing the cost of false positives as compared to false negatives in the same way for the two relevant groups.[67] Only when we know the underlying characteristics of the groups—random arrestees or suspected criminals, for example—can we make any claims about fairness or unfairness.

[67] See Corbett-Davies & Goel, supra note 22, at 11–15.

This conclusion does not mean the lack of error ratio parity is meaningless, however.[68] Yes, it results from the base rate distribution of the target trait. There are more false positives for blacks in the COMPAS data because the data shows that blacks commit more crime and so the algorithm will predict more black crime and will do so imperfectly. Nevertheless, the disparity in error ratios is meaningful because by highlighting the consequences of base rate differences, we see the real-world effects on people that this base rate distribution gives rise to. As a result, this measure provides an additional normative reason to explore the ways in which the data may be biased and the ways in which the data may be the product of prior unfairness that we should avoid entrenching. In other words, the lack of error ratio parity raises the stakes and as such requires us to look more deeply and more carefully at what is going on, as the next Section explores.

D. Why Error Ratio Parity Is Relevant to Fair Treatment

The lack of error ratio parity is important because it highlights the real-world way in which the differences in base rates manifest and, in so doing, creates an obligation to interrogate them—both factually and morally.

First, the fact that base rate variances yield such stark differences in the error rate ratios gives us a moral reason to make extra efforts to ensure that the data on which the algorithm relies is accurate. To be sure, we always want reliable data. But when the stakes of inaccuracies are high, we should make special effort to confirm the accuracy of the underlying inputs. And, as others have noted, base rate data about groups who have suffered discrimination in the past is especially at risk of inaccuracy. If arrest statistics are a function of policing practices as well as actual crime rates, then reliance on arrests to predict recidivism has problems. This "measurement error"[69] is most likely what Representative Ocasio-Cortez had in mind when she claimed that algorithms just "automat[e] the bias."[70]

The data on which algorithms rely will not perfectly reflect the trait it purports to reflect in most instances. Test scores are not perfect reflections of knowledge or ability. Arrests are not perfect reflections of actual crime. The neutral sounding term "measurement error" conveys the ubiquity of

[68] Id. at 11–16 (stressing the limitations of these measures).
[69] Goel et al., supra note 36, at 7.
[70] See Li, supra note 1.

the problem. Some traits simply cannot be measured directly, and proxies will be the best we can do. However, sometimes these proxies are skewed in predictable and problematic ways. When they are, we should do what we can to combat these biases. This is an issue that has attracted significant attention in both the popular press and the academic literature. For example, Sandra Mayson argues that in the criminal justice context, predictive algorithms should only use arrest data for serious, violent crime (instead of using all arrests as an input) because this data is likely to be more reliable.[71]

The fact that the ratio between the false positive rates and false negative rates is so different for the groups involved provides an additional reason to investigate whether the data on which we rely is inaccurate in a way that is biased against protected groups. Sometimes, after investigation, we may be satisfied that the data is as accurate as practicable. Other times, we will not. The fact that there is an additional reason is important because there will always be trade-offs involved in improving data. Getting better data could be costly. Whether it is worth the cost will depend on the reasons that weigh on the other side. What I am suggesting is that the disparity in error ratios should count as a reason to expend resources improving the data.

Second, the lack of error ratio parity might also indicate that the algorithm is compounding a prior injustice. Accurate data on base rate differences may result from prior injustice. For example, suppose that low educational attainment is predictive of recidivism. And suppose that blacks are more likely to have left school early because the schools they attended were inferior. If an algorithm uses educational attainment to predict recidivism, it may use the fact that blacks were unfairly treated in the past to justify treating them worse today. This is the problem I term "*compounding injustice*."[72]

Consider another example. Suppose that inmates who have themselves been victims of child abuse are more likely to recidivate than those who have not been victims. A parole board might take that factor into consideration when making parole decisions. If so, there is no inaccuracy

[71] See Mayson, supra note 33, at 562 (concluding that, despite some difficulties, "arrest for a serious violent crime is still . . . the best measure available").

[72] Statutory prohibitions on disparate impact can be justified by the duty to avoid compounding injustice. See Deborah Hellman, Indirect Discrimination and the Duty To Avoid Compounding Injustice, *in* Foundations of Indirect Discrimination Law 105, 107–09 (Hugh Collins & Tarunabh Khaitan eds., 2018). This example is drawn from that chapter.

if victims of child abuse are more likely to recidivate. But something seems troubling about this practice, nonetheless. The fact that this person is more likely to recidivate is due to the fact that he has been the victim of injustice himself. If the parole board takes this factor into account in determining whether to release him on parole, it compounds the prior injustice by carrying it forward into another domain.

Differential base rates for blacks and whites may well be the result of prior injustice. This is especially true when what is measured by the base rate is health, employment, education, or interaction with the criminal justice system.[73] When the algorithm uses the data that is the product of the prior injustice, the error ratio disparity instantiates the way this injustice is carried forward into another domain.

Worries about automating bias and compounding injustice arise particularly when lack of error rate parity exists between legally protected groups. Given what we know of United States history, base rate differences in crime between blacks and whites are importantly different than base rate differences in disease between two random groups, like the Greens and the Blues. In the case of racial differences, we have good reason to suspect the factual problem of measurement error (automating the basis) and the moral problem of compounding injustice are the cause of the differential base rates. The lack of error rate parity therefore provides a moral reason (1) to investigate the accuracy of data more than one otherwise might and (2) to hesitate in using the data so as to avoid possibly compounding prior injustice.

E. Rebuttal and Reply

Some scholars suggest that the harm to racial minorities scored by the algorithm can be made up for, to some degree, by benefits to other members of the same protected group who are affected by the scoring practices. For example, Aziz Huq argues that we ought to assess the permissibility of algorithmic tools used in the criminal justice context by assessing whether they provide more benefit than harm to racial

[73] See, e.g., Bureau of Labor Statistics, Rep. 1082, Labor Force Characteristics by Race and Ethnicity, 2018 (2019), https://www.bls.gov/opub/reports/race-and-ethnicity/2018/pdf/-home.pdf [https://perma.cc/LM9S-GAEL] (showing U.S. 2018 unemployment rate at 3.9% overall and at 3.5% for whites, while jobless rate for blacks was 6.5%); David R. Williams et al., Race, Socioeconomic Status, and Health: Complexities, Ongoing Challenges, and Research Opportunities, Annals N.Y. Acad. Sci. 69 (2010) (overviewing research indicating health disparities among racial groups and socioeconomic statuses).

minorities as a group—both those scored by the algorithm and those affected by the use of the tool. In his view, "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."[74] Because much violent crime is intra-racial,[75] the victims of those racial minorities who would recidivate are likely to be other members of the same racial groups. In his view, the proper way to evaluate the fairness of the algorithm is to focus on how minorities as a whole are affected, and in particular, on whether the deployment of the algorithm lessens or worsens the racial stratification of society.[76]

This argument surely has substantial appeal. However, it depends on an unstated assumption about what is the relevant fairness question to ask. My focus has been on the first of the comparative questions: are blacks scored by the algorithm treated fairly as compared to whites scored by the algorithm? Huq's focus is on how blacks (both those scored and those not scored) are affected as compared to how whites (both scored and not scored) are affected. In my view, the narrower comparison is the morally relevant one to ask, as the argument below demonstrates.

To see why, return to the example I call *Pedagogical Choice*. This time, I will put some numbers to the scenario I described and, to keep things simple, will use the same confusion tables I used in the Green/Blue

[74] Huq, supra note 7, at 1128. The group who would otherwise not go on to inflict any social harm is those to whom the algorithm assigns false positives. Id. at 1125–26.

[75] See Rachel E. Morgan, U.S. Dep't of Justice, Bureau of Justice Statistics, NCJ 250747, Race and Hispanic Origin of Victims and Offenders, 2012–15, at 1 (2017), https://www.bjs.gov/content/pub/pdf/rhovo1215.pdf [https://perma.cc/H4G4-XXCP] ("During the 4-year aggregated period from 2012 to 2015, half (51%) of violent victimizations [including rape or sexual assault, robbery, aggravated assault, and simple assault] were intraracial . . . ."); Rachel E. Morgan & Barbara A. Oudekerk, U.S. Dep't of Justice, Bureau of Justice Statistics, NCJ 253043, Criminal Victimization, 2018, at 13 tbl.14 (2019), https://www.bjs.gov/content/pub/pdf/cv18.pdf [https://perma.cc/36ZY-URRJ] (showing that "[b]ased on victims' perceptions, the largest percentage of violent incidents committed against white, black, and Hispanic victims were committed by someone of the same race or ethnicity," where violent incidents do not include homicide since statistics are based on victim surveys); Fed. Bureau of Investigation, Uniform Crime Reports: Crime in the United States 2018, Expanded Homicide Data Table 6, https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/tables/expanded-homicide-data-table-6.xls [https://perma.cc/T37D-VCSQ] (showing that race of the offender and victim are the same in most murders of white and black victims).

[76] See Huq, supra note 7, at 1127–28.

disease case and the Black/White recidivism case (modeled on COMPAS).

TRUE OUTCOME				TRUE OUTCOME			
GRADE		Prepared	Unprepared	GRADE		Prepared	Unprepared
	A	60 [a]	20 [b]		A	16 [a]	5 [b]
	C	6 [c]	14 [d]		C	22 [c]	57 [d]
Table 4-1 (Women)				Table 4-2 (Men)			

In this scenario, the test exhibits equal predictive value, as a grade of A or C is approximately equally predictive of actual knowledge for both women and men. For women, an A grade accurately reports knowledge in 75% of the cases; for men, an A grade accurately reports knowledge in 76% of the cases. Similarly, C grades are equally predictive of lack of knowledge (approximately) for both men and women. A grade of C for women accurately reports lack of knowledge in 70% of cases and for men it accurately reports lack of knowledge in 72% of cases. Yet fewer prepared men get As than get Cs. And, to add insult to injury, more unprepared women get As than get Cs. In other words, this test lacks parity in the ratio of false positives to false negatives (error ratio parity). For women, there are far more false positives than false negatives. For men, the reverse is true; there are far more false negatives than false positives.

This lack of ERP does not show that the test is unfair, as I explained above.[77] But it does raise questions. This time, however, it is men who are arguably treated unfairly, not women.[78] What I want to explore with this example is the argument that if the test treats men unfairly, this unfairness could be made up for by some benefit to other men not scored by the algorithm. In order to explore that argument, consider the following hypothetical example:

[77] See supra Section II.C.

[78] While the law treats policies that disadvantage men the same as those that disadvantage women, see, e.g., *Craig v. Boren*, 429 U.S. 190, 204 (1976) (striking down a law that provided a higher drinking age for young men than for young women), the lack of ERP is more suggestive of unfairness when there are other reasons to worry about automating bias and compounding injustice. When the group affected is not a previously disadvantaged group, these reasons are less likely to apply. Whether we should adopt the same symmetry as the law does is a question I leave for another day.

> *The Slacker Bump*: Suppose that men are less well-prepared than women for jobs in the current labor market that require more skills. If prepared men scored by the exam/algorithm are mischaracterized by the test at high rates as unprepared (i.e. given grades of C), they will present less competition to unprepared men who have not taken the test. And, if many jobs are still fairly gender segregated, the harm to the skilled men erroneously scored by the algorithm will benefit unskilled men with whom they are likely to compete. Can we conclude that the test treats men fairly?

My answer—and yours as well, I hope—is that this argument is unsuccessful. It initially seems plausible in the criminal justice context because of the implicit shift to a different fairness question. However, in my view, we cannot make up for unfairness to men scored by the algorithm with a benefit to other men not scored by the test but still affected by this scoring. Similarly, if COMPAS is used to predict recidivism and we worry that it treats blacks scored by the algorithm unfairly as compared to whites scored by the algorithm, we cannot make up for this unfairness with a benefit (if one exists) to other blacks affected by the scoring practice. The *Slacker Bump* example shows that while benefit to others has moral relevance, it doesn't ameliorate the unfairness between the two groups scored by the algorithm.

Let me summarize what has been covered thus far. Part I presented a dilemma. When underlying base rates for some trait are different between two groups, it is mathematically impossible to achieve equal predictive value and error rate balance. This generated the question: what does fairness require? The conceptual intervention of Part I emphasized that equal predictive value is a measure that is best suited to a belief-related question: given this evidence, what should I believe? As such, it is not particularly well-suited as a measure of fairness. Part II presented the argument that it is the ratio between false positive and false negative rates—a measure I term error ratio parity—that we should focus on. This Part presents the argument that fairness between groups scored by the algorithm requires that an algorithm set the balance between false positives and false negatives in the same way for each group. *Error ratio parity* is not a direct measure of this conception of fairness, as this Part explains. Nonetheless, a lack of error ratio parity between a previously disadvantaged group and its counterpart (blacks and whites, for example) is suggestive of unfairness and provides a normative reason to engage in further investigation and for caution. In the next Part, I consider how this unfairness can be mitigated.

III. RACIAL CLASSIFICATION WITHOUT DISPARATE TREATMENT: THE LEGAL CLAIM

Lack of error ratio parity is suggestive of unfairness. How might this unfairness be lessened? Two possibilities come to mind. First, one can mitigate the burden of errors. Second, one can improve accuracy and thereby limit the frequency of errors. Unfortunately, there are barriers to the adoption of each strategy. If the effect of a high-risk score in the context of an algorithm used to predict recidivism were the provision of helpful services rather than incarceration, the unfairness of more false positives for blacks than for whites would clearly be of less moral concern. Changing how states act in response to the scores is thus one strategy to limit unfairness. The barriers to adopting this approach are practical and political. Alternatively, one could improve the accuracy of algorithms overall and thereby limit errors by including race, sex, and other protected traits within the algorithms themselves. Computer scientists and others who design algorithms recognize the ways in which permitting algorithms to use racial classifications within algorithms will improve accuracy.[79] However, they largely refrain from doing so because they believe the law forbids this practice.[80] The barrier to the adoption of this strategy is a perception of illegality. But, as Section III.B argues, this perception may well be overstated. The legal claim that constitutes the third contribution of this Article is this: use of racial classifications within algorithms does not (or not clearly) constitute disparate treatment on the basis of race. As a result, the law provides less of a barrier to mitigating the unfairness of algorithms than many believe.

A. Reduce the Burden of Errors

Error ratio imbalance exists when the ratio between false positives and false negatives differs between two groups. If one type of error is more burdensome than the other, this imbalance may be cause for moral concern. One strategy for mitigating the moral significance of this

[79] See, e.g., Corbett-Davies & Goel, supra note 22, at 2 ("In contrast to the principle of anti-classification, it is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics.").

[80] See Corbett-Davies et al., supra note 24, at 804–05.

differential burden, therefore, would be to alter the consequences of these errors.

In an insightful recent article, Sandra Mayson argues for exactly this approach.[81] If the effect of classification as high risk were "greater access to support and opportunities" rather than incarceration, "a higher false-positive rate among black defendants would be less of a concern."[82] In other words, if the burden were more of a benefit, the disparate impact of the error rate imbalance would create less unfairness.

I agree with Mayson that lessening the consequences of errors helps to ameliorate the unfairness of error ratio imbalance. The goal of this approach would be to equalize the costs of errors between the two relevant groups. If we cannot equalize the error rates themselves, this approach strives to equalize the overall burden such differential error rates produce by adjusting the consequences of errors.

The drawbacks of this approach are likely to be practical—in two ways. First, Mayson's recommendations are fairly demanding and likely to be difficult to achieve, as she acknowledges.[83] Second, it will be necessary to figure out how to adjust such costs to each context. Mayson is focused on the criminal justice context and so her policy recommendations are geared to that context.[84] When algorithms are used to make employment decisions or decisions about whether to issue loans, for example, different strategies will be needed. In the abstract, it is hard to assess whether there will in fact be ways to lower the burdens of each form of error in all the myriad situations in which the need to do so will arise.

B. Improve Accuracy Overall by Using Protected Traits

The fact that one cannot equalize both predictive accuracy and error rates depends on two conditions. First, it occurs because the base rate for the target trait is different for the two groups at issue. Second, it occurs because the test is not perfectly accurate. Part II explored the moral significance of differential base rates and argued that different error rate ratios provide relevant information that indicates that the algorithm may be unfair. In this Part, I highlight the oft-neglected fact that improving the accuracy of algorithms will also diminish both errors in an absolute sense

---

[81] See Mayson, supra note 7, at 2286–93.

[82] Id. at 2287.

[83] See id. at 2290 (recognizing that "[c]hanging the default response to risk would require overcoming . . . institutional and cultural barriers").

[84] See id. at 2286–93.

as well as the divergence in error rate ratios between groups.[85] One obvious way to improve accuracy would be for algorithms to include protected traits like race and sex.[86] Algorithms are designed to be "race blind" because their designers, as well as many legal scholars, assume that use of racial classifications within algorithms is legally prohibited.[87] In what follows, I argue that this conclusion is overstated.[88] While the use of protected traits within algorithms is likely legally impermissible in some instances, it is likely permissible in others. To preview the conclusion: I conclude that an algorithm may not deploy different "cut points" for blacks than for whites, meaning that it cannot set different risk scores for what it determines to be high risk for one race than for another. But an algorithm can take race into account to determine what other traits should be brought to bear to determine actual risk.

The use of different cut scores would constitute disparate treatment on the basis of race, but the use of race to determine what other factors to include within an algorithm does not. This conclusion highlights the fact that the legal category of "disparate treatment" is more elusive than is often recognized, a conclusion that has both practical and conceptual significance. It matters practically because if the designers of algorithms are persuaded that they may use protected traits in the manner I describe, both fairness and accuracy will be improved. It matters conceptually because it demonstrates the way in which the categories of *disparate*

---

[85] See Garg et al., supra note 23 (demonstrating that improving accuracy improves fairness, using several different conceptions of fairness).

[86] Skeem, Monahan, and Lowenkamp argue risk assessment devices used in the criminal justice context should explicitly take account of sex or risk "overestimating women's likelihood of recidivism." Jennifer Skeem et al., Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men, 40 Law & Hum. Behav. 580, 591 (2016). Whether current constitutional and statutory law permits such explicit gender-based classification is unclear. How the analysis presented in this Article would change if the protected trait were sex rather than race would require a related but somewhat different analysis.

[87] See Corbett-Davies et al., supra note 24, at 804–05.

[88] The view that algorithms may consider race and other protected traits in some fashion is gaining some currency in the legal literature. See, e.g., Jason R. Bent, Is Algorithmic Affirmative Action Legal?, 108 Geo. L.J. (forthcoming 2020) (arguing that using race in algorithms may survive strict scrutiny in some contexts). In my view, the term "algorithmic affirmative action," which Bent borrows from Anupam Chander, see id. at 4–5, misleadingly conveys that the explicit use of race within algorithms provides minorities with a benefit when compared with non-minorities. The use of race within algorithms that I endorse is a way to ensure that predictions for each group are as accurate as they can be. Cf. Chander, supra note 6, at 1027 (arguing that algorithms should be designed "in race- and gender-conscious ways to account for existing discrimination lurking in the data").

treatment and *disparate impact* are less distinct and more porous than current legal doctrine acknowledges.

1. Different Thresholds Versus Different Tracks

In the context of traits that are not legally protected, algorithm developers are free to improve accuracy in several ways: they can segment the data to create two different predictive algorithms, set different thresholds at which a particular action is warranted for the two groups, or use the trait within the algorithm to determine how other traits are brought to bear to predict the relevant target variable. Where race and other protected traits are involved, however, computer scientists feel they are constrained by law. In this Subsection, I highlight two different ways that the protected trait "race" could be used by an algorithm and argue that one of these ways is legally problematic and the other is not. There are surely many variants other than these two ways an algorithm could use racial categories. I do not mean to suggest these are the only possibilities. Rather, I select them because one seems clearly legally problematic and the other is, at least plausibly, legally permissible.

a. Legal Background

A brief primer on U.S. anti-discrimination law may be helpful first. Most laws and policies classify and thus draw distinctions between people on the basis of some trait. For example, commonplace and fairly uncontroversial laws require that a person be sixteen to drive or require that a person pass the bar exam to practice law. The first law distinguishes on the basis of age and the second on the basis of bar-passage. While most distinction-drawing is clearly legally permissible (as these two examples demonstrate), some distinction-drawing raises potential legal problems. Only when the law or policy classifies on the basis of particular traits or affects groups defined by those traits does anti-discrimination law become engaged. These traits, referred to as "protected traits," include both race and sex, as well as a limited list of other traits, which are either recognized by courts (in the context of constitutional law) or specified within the relevant statutes (in the context of statutory anti-discrimination law). As a matter of U.S. constitutional law, this list of traits is more limited than under U.S. statutory law. For example, in the United States, disability is not a protected characteristic as a matter of constitutional law[89] but is as a matter of statutory law.[90] In addition, different bodies of law apply to different actors. Constitutional law only applies to governmental actors, while statutory law applies to specified private actors as well. But the class of particular private actors the statutory law applies to is itself determined by the relevant statutes at issue. In what follows, I focus on constitutional law because the central example I have focused on—the use of risk assessment tools by states and localities to determine whom to release on bail or whom to release early from prison—would be governed by constitutional law.[91]

"Disparate treatment" on the basis of either race or sex gives rise to heightened judicial review and is disfavored by U.S. constitutional law.[92] For simplicity, I will focus here on race.[93] Both explicit racial classification and the intention to classify on the basis of race constitute disparate treatment on the basis of race. Whether it is invidious intent or racial classification that is the "touchstone"[94] of an equal protection violation is controversial.[95] Sometimes the Supreme Court emphasizes

[89] See City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 440–43 (1985).

[90] See Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 327 (codified as amended in scattered sections of 42 and 47 U.S.C.).

[91] An extension of the analysis presented in this Article might focus instead on statutory anti-discrimination law. The conclusion that both lack of predictive parity and error rate imbalance constitute forms of disparate impact would remain the same. A statutory analysis would go on to consider whether this disparate impact violates the relevant statutes at issue.

[92] See Adarand Constructors, Inc. v. Pena, 515 U.S. 200, 227 (1995) ("[W]e hold today that all racial classifications, imposed by whatever federal, state, or local governmental actor, must be analyzed by a reviewing court under strict scrutiny."); Craig v. Boren, 429 U.S. 190, 197 (1976) ("To withstand constitutional challenge, previous cases establish that classifications by gender must serve important governmental objectives and must be substantially related to achievement of those objectives.").

[93] An extension of this analysis could consider whether sex-based classifications would be treated differently. This is an important project and one I hope to take up in a second article.

[94] See Washington v. Davis, 426 U.S. 229, 242 (1976) (insisting that "[d]isproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution").

[95] Compare *Adarand*, 515 U.S. at 227 ("[A]ll governmental action based on race—a *group* classification long recognized as 'in most circumstances irrelevant and therefore prohibited,'—should be subjected to detailed judicial inquiry to ensure that the *personal* right to equal protection of the laws has not been infringed." (citation omitted)), with *Washington*, 426 U.S. at 240 (describing "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").

classification[96] and sometimes the Court emphasizes intent.[97] However, when a law or policy contains an explicit racial classification, it often does not matter what the reason or purpose for the classification is. Strict scrutiny is applied. The Supreme Court's affirmative action cases support this view. For example, if a public university considers the race of an applicant in its admissions process, the explicit use of race is subject to "strict scrutiny" and only permitted to the extent that it is justified by a compelling governmental interest.[98] This is true even when the affirmative action policy is the result of a benign motive, such as a desire to remedy past societal discrimination.[99] On the other hand, intention matters when there is no explicit racial classification. If a facially neutral classification (i.e. not race, sex, or some other protected trait) is used deliberately as a proxy for a protected characteristic, the use of the so-called "facially neutral" (or non-protected) classification also gives rise to heightened judicial review.[100]

Both an invidious intention and use of explicit racial classification can constitute disparate treatment on the basis of race and thus give rise to strict scrutiny. With this background in mind, we can now see why neither lack of predictive parity nor lack of error ratio parity constitutes disparate treatment on the basis of race. First, an algorithm designed to achieve

[96] See Grutter v. Bollinger, 539 U.S. 306, 326 (2003) ("[A]ll racial classifications imposed by government 'must be analyzed by a reviewing court under strict scrutiny.'" (quoting *Adarand*, 515 U.S. at 227)); Gratz v. Bollinger, 539 U.S. 244, 270 (2003) ("It is by now very well established that 'all racial classifications reviewable under the Equal Protection Clause must be strictly scrutinized.'" (quoting *Adarand*, 515 U.S. at 224)).

[97] See supra note 95.

[98] See *Grutter*, 539 U.S. at 327–28; *Gratz*, 539 U.S. at 270.

[99] See City of Richmond v. J.A. Croson Co., 488 U.S. 469, 500, 505 (1989) (striking down the City's requirement that municipal contractors subcontract 30% of their contract amount to minority-owned businesses because "[r]acial classifications are suspect, and that means that simple legislative assurances of good intention cannot suffice" and "the city has failed to demonstrate a compelling interest in apportioning public contracting opportunities on the basis of race").

[100] See Hunt v. Cromartie, 526 U.S. 541, 547, 554 (1999) (reversing the district court's grant of summary judgment to defendant where plaintiffs showed circumstantial evidence that race motivated the state's congressional redistricting plan); *Washington*, 426 U.S. at 241. Where a non-protected trait is used to target people with a protected trait in order to promote integration rather than in order to harm the protected group, this practice is likely permissible. See Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 789 (2007) (Kennedy, J., concurring) ("School boards may pursue the goal of bringing together students of diverse backgrounds and races through other means, including strategic site selection of new schools . . . ."). However, as Justice Kennedy is no longer on the Supreme Court, his own views about these issues are less important going forward.

equal predictive value is not adopted with an invidious intent. While the designers may well recognize that this choice will result in error rate or ratio imbalance, this fact alone will be insufficient to turn this disparate impact into an instance of disparate treatment. The Supreme Court has insisted that a screening tool must have been adopted "because of" the disparate impact and not merely "in spite of" these foreseeable consequences in order to give rise to strict scrutiny.[101] Therefore, the fact that the algorithm is designed to achieve predictive parity and foreseeably produces error rate or ratio imbalance does not lead to the conclusion that this algorithm constitutes disparate treatment on the basis of race.

Similarly, if an algorithm's designers were to make the choice to equalize error rate ratios and thereby forgo equal predictive value, this too would constitute disparate impact, not disparate treatment, and so be legally permissible. There is no reason to think that this choice is adopted in order to produce the disparate impact of unequal predictive value. Thus, a choice to privilege error ratio parity, which predictably produces a lack of predictive parity, would also be legally permissible. Alternatively, the designers of an algorithm may choose to adopt some amalgam between predictive parity and error ratio parity. This too would be legally permitted. Algorithmic designers have many legally permitted options.[102]

How far does this legal permissibility extend? If the algorithm's designers attempt to reduce the error ratio disparity and improve accuracy overall by using race within the algorithm, is this permissible?

b. Different Thresholds

One way an algorithm could use racial classifications to limit error ratio disparities would be to set different thresholds for the target trait for each

[101] See Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 (1979); *Washington*, 426 U.S. at 239–41. Indeed, where a facially neutral screening tool is adopted to benefit rather than harm a protected group, such a policy will likely not give rise to strict scrutiny. In the Supreme Court's affirmative action cases, the Court repeatedly encourages universities to adopt facially neutral means of increasing minority enrollment and suggests that such endeavors are to be celebrated, not scrutinized. See, e.g., Fisher v. Univ. of Tex. at Austin, 570 U.S. 297, 312 (2013) ("[S]trict scrutiny imposes on the university the ultimate burden of demonstrating, before turning to racial classifications, that available, workable race-neutral alternatives do not suffice."); *Grutter*, 539 U.S. at 343 ("We take the Law School at its word that it would like nothing better than to find a race-neutral admissions formula and will terminate its use of racial preferences as soon as practicable.").

[102] Cf. Huq, supra note 7, at 1083 (asserting that "the dominant intent- and classification-focused calibration [of Equal Protection doctrine] is ill suited to the forms and dynamics of algorithmic criminal justice tools").

racial group. If whites with a score of six or higher are labeled "high risk" and blacks with a score of eight or higher are labeled "high risk," then the algorithm employs different thresholds or "cut scores" for each racial group. This approach is widely viewed as legally prohibited.[103] As a descriptive matter, I agree that race-specific thresholds would trigger strict scrutiny under constitutional law, and that differential thresholds would be unlikely to survive such demanding judicial review. A different threshold for each racial group would employ an explicit racial classification, and different treatment for members of each racial group would follow. As this conclusion is uncontroversial, I will not discuss it further.

c. Different Tracks Within Algorithms

But there is another way that racial classifications could be used. Instead of setting different thresholds for each racial group, an algorithm might use race within the algorithm to determine what other traits should be used to predict the target variable. This approach could improve both accuracy and fairness. Suppose that some of the traits that predict recidivism are more predictive for one race than for another. For example, Sam Corbett-Davies and co-authors consider the possibility that "housing stability might be less predictive of recidivism for minorities than for whites."[104] If so, perhaps we might utilize two tracks within the algorithm. For whites, housing stability would be included in the predictive algorithm. For blacks, it would not. However, Corbett-Davies and his co-authors worry that using housing stability for whites but not for blacks would require using race explicitly in the algorithm and that doing so would raise legal problems.[105] As a result, they report, "it is common to simply exclude features with differential predictive power."[106] The result of doing so, in their view, is to exacerbate disparate racial impact.[107]

[103] See, e.g., Corbett-Davies et al., supra note 24, at 804 (arguing that "race-specific thresholds . . . would likely trigger strict scrutiny" (emphasis omitted)).

[104] Id. at 805.

[105] See id. at 804.

[106] Id. at 805.

[107] See id. at 804–05 (noting that "discarding information may inadvertently lead to . . . redlining effects," which is discrimination accomplished "by ignoring information about the disfavored groups"); see also Huq, supra note 7, at 1101 ("The procedural purity demanded by an anticlassification rule, in sum, would come at a high price in terms of accuracy in algorithmic application.").

Sharad Goel and co-authors also point out that using separate algorithms for each racial group could help to ameliorate measurement error.[108] They offer the following example. Suppose that the existence and number of past drug sales is predictive of future criminal activity, but that accurate information about actual past drug sales is difficult to obtain. Past arrest or drug distribution convictions might then be used as a proxy for past sales. If we worry that arrest and conviction data is biased by policing practices in which minority communities are more heavily policed than white communities, it might be the case that past arrests for drug selling are more predictive of future criminal activity for whites than they are for blacks.[109] If so, we will increase the accuracy of the algorithm, in their view, by using "two separate statistical models, one for black defendants and another for white defendants."[110]

Joshua Kroll and co-authors, building on the work of Cynthia Dwork and co-authors,[111] provide another example in which a trait is more predictive for one group than for another.

> Consider, for example a system that classifies profiles in a social network as representing either real or fake people based on the uniqueness of their names. In European cultures, from which a majority of the profiles come, names are built by making choices from a relatively small set of possible first and last names, so a name which is unique across this population might be suspected to be fake. However, other cultures (especially Native American cultures) value unique names, so it is common for people in these cultures to have names that are not shared with anyone else. Since a majority of accounts will come from the majority of the population, for which unique names are rare, any classification based on the uniqueness of names will inherently classify real minority profiles as fake at a higher rate than majority profiles, and may also misidentify fake profiles using names drawn from the minority population as real. This unfairness could be remedied if the system were "aware" of the minority status of a name under consideration, since then the algorithm could know whether the

[108] See Goel et. al., supra note 36, at 7.

[109] See id. (noting that since "minorities who engage in drug-related crime are more likely to be arrested than whites who engage in the same behavior . . . , using recorded drug arrests as a proxy for actual drug sales may (incorrectly) rate black defendants as higher risk than white defendants who have engaged in similar criminal behavior" (citation omitted)).

[110] Id.

[111] Kroll et al., supra note 5, at 686 (citing Dwork et al., supra note 59).

> implication of a unique name is that a profile is very likely to be fake or very likely to be real.[112]

In each of these examples, the fact that the algorithm must be blind to differences among the populations creates a problem. If the algorithm could take account of the ways that housing stability is more relevant to recidivism risk for whites than for blacks, that drug sale arrests are less predictive of recidivism for blacks than for whites, and that unique names are more predictive of fraud for non-Native people than for Native Americans, prediction would be improved. In each of the examples, were the algorithm to take race into account in the way it processes other information, the algorithm would do a better job at its task. Both accuracy and fairness would be improved.

Does the law prohibit using racial categories in this way? The answer depends on whether using race within algorithms would constitute disparate treatment on the basis of race. Interestingly, it is not clear that it does.

In one sense, dividing the algorithm into two racial tracks and using different information to evaluate each track constitutes disparate treatment. On the white track, housing stability or instability would be factored into the analysis of whether the individual is at high or low risk of recidivism. On the black track, it would not. In another sense, dividing the algorithm into two racial tracks and using different information to evaluate each track treats each group the same and therefore does not constitute disparate treatment. For both blacks and whites, only relevant information is utilized, where relevance is defined as having a specified level of predictive power. So, while different factors are used to predict recidivism for blacks and for whites, only relevant factors are applied to each. The algorithm includes a racial classification, which suggests that strict scrutiny should be applied. But for each racial group, the algorithm brings to bear only relevant factors, which suggests that strict scrutiny should not be applied. This example, and others like it, puts pressure on what the law means, precisely, by the concept of disparate treatment.

2. Racial Classification Without Disparate Treatment

The law's treatment of explicit racial classifications is more complex and nuanced than scholars writing about algorithms have recognized thus

[112] Kroll et al., supra note 5, at 686–87 (footnote omitted).

far. In fact, not all racial classifications are subject to strict scrutiny. For example, racial classification is subject to lesser judicial oversight when used for information-gathering purposes only.[113] In addition, racial classifications are sometimes permitted when they do not rely on a racial generalization.[114] In each of these instances, courts find that the deployment of a racial classification does not constitute disparate treatment on the basis of race. The fact that racial classifications are sometimes legally permitted without passing heightened review opens the door to the possibility of using race within algorithms. To the extent that these strategies can improve accuracy overall, they can also improve fairness.

The arc of the argument presented below is as follows. I begin by considering two circumstances in which racial classification does not constitute disparate treatment. From these examples, I extract two principles. Using these principles, I examine the deployment of racial classifications within algorithms and conclude that this practice may not constitute disparate treatment on the basis of race and so may not give rise to heightened judicial review.

a. Information Not Use

If any use of a racial classification, in any context, constitutes disparate treatment on the basis of race, then the use of racial tracks within algorithms would do so as well. But this is not the case. Despite common assumptions to the contrary, the fact that a law or policy deploys a racial classification does not always constitute disparate treatment. For example, the commonplace practice of collecting information using racial categories does not appear to constitute disparate treatment on the basis of race. As Kim Forde-Mazrui notes,

> [I]t is no exaggeration to observe that millions of hours are spent every year by researchers and policymakers at all levels of government, including public universities—and in a wide variety of private organizations, often with government funding—investigating racial disparities in contexts such as health, family, education, employment,

[113] See infra Subsection III.B.2.a.
[114] See infra Subsection III.B.2.b.

> criminal justice, and virtually all areas of the civic, economic, and social life of the nation.[115]

The fact that the racial classifications used in these practices are ubiquitous suggests that they are permissible.

For the most part, the use of racial classification in data collection has been unchallenged. However, one district court has considered whether the United States Census may use racial categories.[116] As discussed below, the result of that challenge reinforces the conclusion that racial data collection does not constitute disparate treatment on the basis of race.[117]

The Census collects information about the number of people living in the United States, as required by the Constitution.[118] It also collects additional information about characteristics of the U.S. population, including information about race.[119] Racial information has been collected on the Census since 1790, though not with the same level of specificity as is solicited in the Census's current form.[120] The collection of this information was challenged in the district court case *Morales v. Daley*, decided in 2000. The plaintiffs argued that the deployment of racial categories on the Census should be subject to strict scrutiny,[121] and the government defended the use of the race-based classification on the ground that the information was "needed to assess racial disparities in health and environmental risks" and to meet redistricting requirements.[122]

---

[115] Kim Forde-Mazrui, The Canary-Blind Constitution: Must Government Ignore Racial Inequality?, 79 Law & Contemp. Probs. 53, 72 (2016).

[116] Morales v. Daley, 116 F. Supp. 2d 801 (S.D. Tex. 2000).

[117] The Supreme Court recently considered whether the Secretary of Commerce's decision to add a citizenship question to the 2020 Census was constitutionally permissible in *Dep't of Commerce v. New York*, 139 S. Ct. 2551 (2019). In holding that the Secretary abused his discretion because the stated reason for the addition was pretextual, Chief Justice Roberts noted that the Census asks a question about race but did not consider its constitutionality. See id. at 2561–62.

[118] Article I, Section 2, Clause 3 of the Constitution of the United States requires that an "actual Enumeration shall be made within three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years." U.S. Const. art. I, § 2, cl. 3.

[119] Collecting this information, however, is not constitutionally mandated.

[120] See *Morales*, 116 F. Supp. 2d at 809 (noting that the Census "has always included additional data points, such as race, sex, and age of the persons counted"); compare An Act Providing for the Enumeration of the Inhabitants of the United States, ch. 2, 2 Stat. 101 (1790) with U.S. Dep't of Commerce, United States Census 2010.

[121] *Morales*, 116 F. Supp. 2d at 810.

[122] Id. at 813.

In addition, the government argued that the collection of information, on its own, does not constitute disparate treatment and thus that strict scrutiny did not apply.[123]

The District Court for the Southern District of Texas upheld the use of racial classification—including the requirement that census respondents self-report their race under penalty of substantial fines—and declined to apply strict scrutiny.[124] The court reasoned that "Plaintiffs' position is based upon a misunderstanding of the distinction between collecting demographic data so that the government may have the information it believes at a given time it needs in order to govern, and governmental use of suspect classifications without a compelling interest."[125] Because collection of information is different from use, the former does not constitute disparate treatment and thus does not give rise to strict scrutiny.[126]

This example illustrates that what distinguishes a racial classification that constitutes disparate treatment from a racial classification that does not constitute disparate treatment is the relationship the classification has to real-world effects, i.e., collection versus use. In addition, the census example suggests that the effect of the racial classification must be direct and not merely the downstream consequence of such classification.[127] The collection of racial data on the Census is highly consequential, after all, with substantial impact in the real world, including for redistricting and for the allocation of governmental resources. And yet, these effects are insufficient to make racial classifications in the Census subject to strict scrutiny. The reason, one suspects, is that these effects are too remote.

b. No Racial Generalization

Even when racial classifications have direct effects on the people subject to them, these classifications are not always subject to strict scrutiny. The manner in which they are used also matters. Strict scrutiny

[123] Id. at 813–14.

[124] Id. at 809, 814–15.

[125] Id. at 814.

[126] While the court in *Morales* does not make crystal clear that it upholds the classification without applying strict scrutiny, that is the clear implication of its analysis. The case contains no discussion of whether the asserted governmental interests are "compelling," which would be required if strict scrutiny had been applied.

[127] See *Morales*, 116 F. Supp. 2d at 814–15 (holding that "requiring a person to self-classify racially or ethnically" does not violate due process, no matter the ultimate "use[s] [to which] such classifications have been put in the past").

applies where the use of a racial classification relies on a generalization about the racial group. And when it does not, the use of racial classifications does not always give rise to strict scrutiny.

Consider the following example. When an eyewitness or crime victim describes the perpetrator as a person of a particular race, police focus their investigations on people of that race. Notwithstanding the fact that a racial classification is used to determine whom to investigate, stop, or search, such conduct has not been considered to constitute disparate treatment on the basis of race and thus does not give rise to strict scrutiny.[128] For the person on whom police investigative efforts focus, it may well feel like disparate treatment on the basis of race.[129] Yet, as the Second Circuit in *Brown v. City of Oneonta* explains, it is not. Why not?

Reliance on a racial suspect description does not constitute disparate treatment on the basis of race because the police department in such a case does not rely on a racial generalization.[130] To be sure, the police department does rely on a generalization, and that generalization includes a racial classification. But the police are not relying on a generalization about people of a particular race, and thus the department is not employing a racial generalization.[131]

The police department in *Brown* presumably operated according to the following policy: *follow the suspect description*. Because the victim of an attack described her assailant as a black man, this policy led the police department to search black men. Such a policy is meaningfully different from a police department policy of policing black men more heavily than

[128] See Brown v. City of Oneonta, 221 F.3d 329, 333–34 (2d Cir. 2000), cert. denied, 534 U.S. 816 (2001) (holding that the search of all the black residents of Oneonta, New York, in response to a report from a crime victim that the perpetrator was black did not violate the Equal Protection Clause).

[129] Some scholars argue that it is and should therefore be subject to strict scrutiny. See R. Richard Banks, Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse, 48 UCLA L. Rev. 1075, 1080 (2001) (arguing that race-based suspect descriptions employ racial classifications and thus should warrant strict scrutiny if Equal Protection doctrine adheres to a norm of color-blindness, and going on to demonstrate that current doctrine only sometimes adheres to this norm).

[130] The Second Circuit concluded that the plaintiffs had not "identified any law or policy that contains an express racial classification" because the policy of the police department was, instead, to respond to the suspect description of the witness or victim, no matter the race. *Brown*, 221 F.3d at 337.

[131] See id. (emphasizing that plaintiffs "were not questioned solely on the basis of their race" but instead "on the altogether legitimate basis of a physical description given by the victim of a crime").

white men, for example (racial profiling).[132] Racial profiling is based on a generalization about blacks and their likelihood of committing crime. As the *Brown* court explained, "Plaintiffs do not allege that upon hearing that a violent crime had been committed, the police used an established profile of violent criminals to determine that the suspect must have been black."[133] If they did, the police would be generalizing about blacks, i.e. from the trait black, they would be inferring that such a person is likely to be a criminal (or more likely than the average person to be a criminal). The police in *Brown* relied on a different kind of generalization, one about the reliability of eyewitness descriptions. Their policy—*follow the suspect description*—implicitly relies on the generalization that eyewitness reports are more likely to be helpful than not (or are sufficiently likely to be accurate to justify the burdens imposed) or something of that nature.[134] Race is used within the policy in this particular case but only because the policy generalizes about eyewitnesses, not because it generalizes about blacks.[135]

[132] For a thoughtful description of the distinction between racial profiling and reliance on racial suspect descriptions, see Arthur Isak Applbaum, Response: Racial Generalization, Police Discretion, and Bayesian Contractualism, *in* Handled with Discretion: Ethical Issues in Police Decision Making 145, 145–58 (John Kleinig ed., 1996).

[133] *Brown*, 221 F.3d at 337.

[134] Whether this generalization is correct is up for debate. Fred Schauer, for example, emphasizes the way in which seemingly direct evidence, like eyewitness reports, is probabilistic in just the same way as profiles and other probabilistic evidence. See Frederick Schauer, Profiles, Probabilities, and Stereotypes 101–03 (2003).

[135] The Fourth Circuit adopted the same rationale in *Monroe v. City of Charlottesville*, 579 F.3d 380, 382 (4th Cir. 2009), cert. denied, 559 U.S. 992 (2010) (upholding the dismissal of an Equal Protection challenge to police seeking out and asking for DNA samples from young, African-American men in Charlottesville in response to victim's descriptions of a rapist as a young African-American man). In *Monroe*, the Fourth Circuit explained its reasoning as follows:

> This is not a case in which police created a criminal profile of their own volition and decided which characteristics, such as race, that the criminal possessed. Nor is this a situation where police were faced with conflicting or uncertain evidence as to the assailant's race and made the decision to pursue only African-Americans. Rather, as earlier indicated, the police decided to approach Monroe based on the similarity between him and the several elements of the victims' descriptions, not because of a plan to investigate African-Americans.

Id. at 388.

c. Principles and Application

These examples demonstrate that not all uses of racial classifications constitute disparate treatment or give rise to strict scrutiny.[136] Only some do. Thus, the mere fact that an algorithm uses race in predicting recidivism should not by itself give rise to strict scrutiny. How the algorithm employs the racial classification also matters. Drawing from these two examples—the collection of information using racial categories and the reliance on racial suspect descriptions—we can extract principles that help to guide us regarding what disparate treatment requires and how that doctrine bears on the use of racial classifications within algorithms. However, a note of caution is warranted. First, as the Supreme Court has not weighed in on either of these examples, they may turn out to be less significant than this Article assumes. The Court denied certiorari in both *Brown v. City of Oneonta*[137] and in *Monroe v. City of Charlottesville*,[138] a 2010 case from the Fourth Circuit that permitted police use of racial suspect descriptions for the same reasons as did *Brown*. Second, the analysis presented here works to make coherent and find an underlying rationale for a body of doctrine which may not be amenable to either.

With those caveats in mind, we can use these examples to provide guideposts for determining when the use of racial classifications does not constitute disparate treatment. Two principles emerge. First, the Census example suggests that the use of racial classifications must produce a proximate effect in order to constitute disparate treatment. Second, the permissibility of racial suspect descriptions suggests that when race is used within a generalization, only generalizations about racial groups constitute disparate treatment on the basis of race.

When race is used within an algorithm to determine what weight to give to other factors like housing stability, it lacks both of the features just mentioned. First, the effect produced by this use of a racial classification is not proximate. Rather, the use of race determines what other factors to employ in making a prediction about recidivism risk. The racial category

[136] Aziz Huq agrees that the apparent constitutionality of relying on racial suspect descriptions suggests that some uses of race in algorithms are constitutional, but he has a different explanation for why. See Huq, supra note 7, at 1096 (surmising that "[r]ace-based feature selections would then trigger no more constitutional concern than race-based suspect descriptions" because "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").

[137] 221 F.3d 329.

[138] 579 F.3d 380.

provides information that in turn can be used to determine what other traits should be incorporated into the algorithm. Like the racial information in the Census, this racial information is likely to have downstream consequences, but these effects are too remote from the use of the classification itself to constitute disparate impact on the basis of race.

Second, the generalization embodied in the algorithm is a generalization about the relationship between housing stability and recidivism, given a person of a particular race. This is analogous to the generalization about the reliability of eyewitness testimony, given a report about a perpetrator's race. While the algorithm relies on a generalization about what housing stability or instability indicates for people of each race, the generalization itself is not a racial generalization. It refers to the racial classification but not by relying on a racial generalization. And it does this in the same way as do suspect descriptions. Housing instability is predictive (or not), depending on race. Eyewitness reports are predictive (or not), given a report about the race of the assailant. Given this structural similarity, there is good reason to think that the use of race within algorithms is, and should be, permissible.

3. Ricci*'s Irrelevance*

Some scholars[139] appear to think that modifying an algorithm to avoid a racially disparate impact is specifically prohibited by the Supreme Court's decision in *Ricci v. DeStefano*.[140] If that were correct, the suggestion that a state could actually employ racial categories within an algorithm would be clearly impermissible as it would take racial awareness one step further. In my view,[141] these scholars overread *Ricci*. To see why, consider the facts of the case.

The New Haven Fire Department had developed a test to use in determining who would be promoted. Firefighters studied for this test, purchased review materials, and otherwise invested considerable time, energy, and money in preparing for the test.[142] When the results were

[139] See, e.g., Barocas & Selbst, supra note 49, at 724–26 (reading the holding in *Ricci* as prohibiting making changes to an algorithm "[a]fter an employer begins to use the model to make hiring decisions"); Kroll, supra note 5, at 694 (equating the racial awareness advocated here with disparate treatment).

[140] 557 U.S. 557 (2009).

[141] Other scholars agree, most notably Pauline Kim. See, e.g., Kim, supra note 35, at 191.

[142] *Ricci*, 557 U.S. at 562, 583–84.

revealed, the number of minority candidates eligible for promotion was extremely small.[143] As a result, the city decided not to certify the results, and so the firefighters who had passed the test were not eligible for promotion.[144] The city defended its decision on the ground that the disparate impact prong of Title VII of the Civil Rights Act prohibited it from using a screening mechanism that produced a disparate impact without sufficient reason.[145] The Supreme Court struck down the city's decision not to certify the results. In the Court's view, the city's decision itself constituted disparate treatment on the basis of race as applied to the firefighters who had passed the test.[146] In addition, the Court found that without "a strong basis in evidence" that the city would be liable under a disparate impact theory, it was not justified in taking such action.[147]

Kroll and co-authors,[148] as well as Barocas and Selbst,[149] read *Ricci* as prohibiting the intent to avoid a racially disparate impact and the very awareness of race that differential tracking within algorithms would accomplish.[150] These scholars misread *Ricci*, as Pauline Kim persuasively argues.[151] They ignore the fact that specific, identifiable people who had relied on the prior test were affected in *Ricci*—plaintiffs whose stories were relayed to the Court. By contrast, where an algorithm designer is aware that an approach will have a racially disparate impact in the abstract and so makes changes to avoid that impact, we have no specific, known

[143] Id. at 562; Respondents' Brief on the Merits at 5–6, *Ricci*, 557 U.S. 557 (Nos. 07-1428 & 08-328).

[144] *Ricci*, 557 U.S. at 574.

[145] Id. at 575.

[146] Id. at 592–93.

[147] Id. at 592.

[148] See Kroll, supra note 5, at 694–95 ("If an agency runs an algorithm that has a disparate impact, correcting those results after the fact will trigger the same kind of analysis as New Haven's rejection of its firefighter test results.").

[149] See Barocas & Selbst, supra note 49, at 725–26 (arguing that *Ricci* prohibits an employer from making changes to an algorithm after seeing that it will have a disparate impact on racial minorities).

[150] This interpretation overreads *Ricci* in my view. If the employer does not revoke offers from actual individuals, there is no reliance by actual people involved. If the employer uses the model, sees the impact, and then makes changes going forward that affect other potential hiring, *Ricci*'s rationale would not apply.

[151] See Kim, supra note 35, at 191 (arguing that Kroll misreads *Ricci* and that that case "narrowly addressed a situation in which an employer took an adverse action against identifiable individuals based on race, while still permitting the revision of algorithms prospectively to remove bias"); Pauline T. Kim, Data-Driven Discrimination at Work, 58 Wm. & Mary L. Rev. 857, 869 (2017).

people who are harmed, nor any reliance. *Ricci* does not speak to this sort of case and so has only limited value in assessing it.

The debate between Kroll, Barocas, and Selbst on the one hand and Kim on the other is focused on whether it is permissible to modify an algorithm prospectively in response to its projected disparate impact. That debate centers on whether mere awareness of racial impact is sufficient to give rise to strict scrutiny. Kim is clearly correct, in my view, that mere awareness of the racial impact of a proposed course of action does not give rise to strict scrutiny. If it did, the decision to adopt facially neutral policies because of their salutary effect in diminishing racial disparities in all sorts of areas would be constitutionally in jeopardy. Given that Justice Kennedy, who authored the opinion for the Court in *Ricci*, specifically endorsed approaches like choosing to site schools where they will enroll a racially diverse cohort of students,[152] we can safely conclude that we should not read *Ricci* to suggest that an awareness of the racial impact of actions by itself would give rise to strict scrutiny.

The awareness of race that undergirds the use of race within algorithms is not prohibited by *Ricci*. Instead, if that case bears on the question of whether algorithms can employ racial classifications at all, it supports the importance of a proximate effect to a finding of disparate treatment. In *Ricci*, it was the fact that the decision at issue had a direct effect on identifiable people that made a significant difference.

To summarize, Part III has explored how one might mitigate the unfairness that error ratio imbalance suggests and manifests. It first considered how one might do so by minimizing the costs of errors. Part III then turned to addressing how both fairness and accuracy might be improved by the deployment of racial classifications within algorithms. This Part then argued against the majority view that consideration of race within algorithms is always impermissible. Instead, it presented a picture of constitutional equal protection jurisprudence that would render this an open question.[153]

[152] See Parents Involved in Cmty. Sch. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 789 (2007) (Kennedy, J., concurring) ("School boards may pursue the goal of bringing together students of diverse backgrounds and races through other means, including strategic site selection of new schools.").

[153] Interestingly, the Wisconsin Supreme Court recently held that the use of gender within the COMPAS risk assessment tool does not violate due process because using gender improves accuracy. State v. Loomis, 881 N.W.2d 749, 766 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017) (explaining that "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose").

CONCLUSION

This Article makes three contributions to the debate about how best to measure algorithmic fairness. The first contribution is conceptual, the second is normative, and the third is legal. The two most prominent types of measures focus on whether the scores algorithms produce are equally predictive or instead on whether the error rates produced are equal. The conceptual contribution of the Article is to highlight that these different measures are best suited to answering different questions. The accuracy of scores relates to belief and is relevant to a person asking: given this data, what should I believe*?* Because the fairness that is usually at issue relates to how people are treated, a measure geared to questions of belief is ill-suited to this task, as Part I contends.

The second contribution is normative. It argues that fairness between groups scored by the algorithm requires that the way the algorithm balances the two types of errors it might make should be the same for each of the groups at issue. Different ratios between false positives and false negatives constitute different rules, in a very real sense. Yet, as Part II acknowledges, parity in the ratios of false positive rates to false negative rates does not determine that different ratios are employed for the two groups. Nevertheless, lack of parity in the ratios between false positive rates and false negative rates is suggestive of unfairness when the groups at issue have suffered disadvantage in the past. Lack of error ratio parity highlights the costs of differential base rates for racial groups and so provides a special reason to investigate bias in the data and to probe ways that the algorithm may be compounding prior injustice. For these reasons, this measure is important and worthy of our attention.

The third contribution is legal. We can mitigate the unfairness that lack of error ratio parity signals by improving the accuracy of algorithms. Unfortunately, an overstatement of current legal doctrine's resistance to racial classification has led computer scientists to forgo promising ways to improve the accuracy and fairness of algorithms by using racial classifications to determine what other traits should determine the algorithm's output. If algorithms use protected traits in a limited way to determine which other traits to consider within the algorithm, overall accuracy can be improved. Part III argues that constitutional law does not rule this strategy out. The concept of disparate treatment, which is central to equal protection doctrine, is not well defined. While the use of racial classifications by governmental actors usually constitutes disparate treatment on the basis of race, it does not always do so. The examples of racial classifications that do not give rise to heightened review can help us determine how courts should evaluate the use of race within algorithms when racial classifications are deployed to improve accuracy overall. Given the stakes, courts should be open to this approach.