MEASURES OF ALGORITHMIC FAIRNESS
Solving for Fairness [?]

Deborah Hellman

University of Virginia School of Law
Draft, please do not cite or circulate
February 8, 2019

Abstract.

MEASURES OF ALGORITHMIC FAIRNESS

Deborah Hellman

INTRODUCTION

At an event celebrating Marin Luther King Day this year, 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 assumptions. And if you don't fix the bias, then you are just 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

[1] Cite to coverage of event.

the Daily Wire.[2] But, the author of this comment assumes, math is just math and the idea that math can be unfair is crazy.

This recent controversy is just one of many to challenge the fairness of algorithmic decision-making.[3] The use of algorithms, and in particularly machine learning and artificial intelligence, has attracted significant attention in the legal literature as well. The issues raised are varied, including concerns about transparency,[4] accountability,[5] privacy[6] and fairness.[7] This Article focuses on fairness – the issue raised by Ocasio-Cortez. It focuses in particular on the question of *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.

One particular dispute, which dates back to 2016, attracted significant attention. At issue was, and still is, whether an algorithmic tool widely used to assess recidivism risk discriminates against blacks. That debate has proved particularly generative as critics of the algorithm relied on one conception of fairness while its defenders used another. At this point, computer scientists and statisticians entered the debate and attempted to develop ways to achieve fairness in both dimensions. Unfortunately, this proved impossible to achieve, for reasons explained below. As a result, the debate about the fairness of algorithms has entered a second phase. The questions now on the table include whether one should combine the different ways of measuring algorithmic fairness into one composite measure in a way that produces the best result overall. And, if not, which of the competing measures should we prefer and why?

This second stage debate about algorithmic fairness is marked by a proliferation of measures of algorithmic fairness. This multitude of measures suggests that it is especially important now to think more deeply about what these different measures actually capture and how, if at all, they relate to fairness. This project is likely also to be generative. While algorithms use math, to be sure, *fairness* in algorithmic decision-making is not the ubiquitous *x* of high school math class. Solving for fairness is likely to be conceptually complex and contested. In addition, the exploration of the ways in which we might do so reveals underappreciated ambiguity in

---

[2] Cite tweet. The coverage Ocacio-Cortez's comment is mixed. *See e.g.* https://slate.com/news-and-politics/2019/02/aoc-algorithms-racist-bias.html.

[3] Cite to headlines in the past year in which algorithms are charged with bias.

[4] Citron Tech Due Process

[5] include

[6] Pasquale, Black Box Society

[7] Huq, Mayson,

both antidiscrimination law and the moral values on which it rests.

This Article makes two contributions to this endeavor. First, it highlights an overlooked conceptual distinction between the two most influential measures used to assess whether algorithms are fair. One measure is most apt to questions of *belief* and the other to questions of *action*. Recognizing this difference between the measures provides a reason to favor one measure (or type of measure) over the other when the algorithmic tool is used in the context of decision-making. Making this choice has costs however. It would be sensible, therefore, to look for ways to mitigate these costs. However, a common assumption that antidiscrimination law prohibits the use of racial and other protected classifications in all contexts is inhibiting those who design these algorithms from mitigating the costs in the most obvious ways. This Article's second contribution is show that the law poses less of a barrier than many assume.

This description of the Article's contributions is abstract. To make it more concrete, consider the controversy from a few years ago that helped to spark the debate.

In an expose in May of 2016, ProPublica took to task the risk assessment tool used by many states to aid in decision making about whom to release on bail and on parole.[8] The tool, called COMPAS, assigns risk scores to each person. There are several factors that go into these scores but race is not among them.[9] The higher the number, the greater the risk of recidivism. In addition, the numbers themselves are designed to correspond to predicted probabilities such that, if the tool works as intended, eight out of ten of the people who score an 8 on the tool will, in fact, recidivate. ProPublica asserted that COMPAS treated blacks and whites differently. More specifically, ProPublica noted that black arrestees and inmates were far more likely to be erroneously classified as risky than were white arrestees and inmates. The essence of their 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 almost twice the rate as white defendants. White defendants were mislabeled as low risk more often than black defendants."[10]

---

[8] Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." By Julian Angwin, Jeff Larson, Surya Mattu and Kauren Kirchner, ProPublica, May 23, 2016.

[9] Are the public or proprietary. Check.

[10] *See supra* note 8.

Northpointe[11] (the company that developed and owns 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.[12] While Northpointe and others challenged some of the accuracy of ProPublica's analysis,[13] the main thrust of their 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 – eight say, to continue with that example – the two people would be equally likely to recidivate. In fact, eight of ten blacks and eight of ten whites with scores of eight did recidivate. 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, they 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. In other words, ProPublica and Northpointe were focused on different measures when assessing whether blacks and whites were treated similarly by the risk assessment device.

The easiest way to fix the problem would be to equalize with regard to both measures. 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). 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.[14] The reason it is impossible relates to the fact that the underlying rates of recidivism among blacks and whites differ.[15] Whenever the two groups at issue (whatever they are) have

---

[11] Note that company has changed its name – find citation.

[12] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

[13] For a critique of ProPublica's analysis, *see* Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, *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 Federal Probation, No. 2, 38-46 (2016).

[14] *See e.g.* Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arXiv:1609.05807v2 [cs.LG] 17 Nov 2016; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A study of bias in recidivism prediction instruments*, arXiv:1703.00056v1 [stat.AP] 28 Feb 2017. Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, Sociological Methods & Research, 1-42 (2018).

[15] Of course, the data on recidivism itself may be flawed. This consideration is

different rates of the trait predicted by the algorithm, it is impossible to achieve parity between the groups in both dimensions. This fact gives rise to the question: in which dimension is such parity more important and why?

These different measures are characterized as different conceptions of fairness.[16] Part I argues that this is a mistake. The measure favored by Northpointe is relevant to what we ought to *believe* about a particular scored individual. If a score of 8 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 an individual is likely to commit a crime in the future if we do not know his race. The measure favored by ProPublica relates instead to what we ought to *do*. If peaceful blacks and peaceful whites are not equally likely to be mischaracterized by the score, we will not know whether to use the scores produced by the tool in making decisions about bail or parole. 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. For this reason, both measures are not equally relevant to questions of fairness. Therefore, Part I argues, we should privilege balancing error rates[17] rather than predictive parity.

We should prefer a measure that equalizes error rates unless the law prohibits this preference. Fortunately, it does not. As both measures call attention to the ways in which the algorithmic tool produces a disparate impact on a protected group, antidiscrimination law is likely to be agnostic about which measure designers of such algorithms equalize. Both approaches are legally permissible, as Part I argues, and so the choice of which measure to privilege belongs to those who utilize these tools.

Equalizing in each dimension has costs, however. Part II explores what might be done to mitigate the costs of choosing each measure. It focuses particularly on how the algorithm's designers might mitigate the cost of favoring the measure that Part I suggests should be preferred. Suppose we equalize error rates in one of the several ways that we might do that. The upshot of insisting on equalizing error rates will be a loss of predictive parity. It will no longer be the case that a black person and a white person with the same score will be equally likely to recidivate. As a result, when

---

discussed *infra* at ___.

[16] Move footnote from later to here.

[17] There are different ways in which one might balance error rates. One could require the false positive rate to be the same, the false negative rate to be the same, both to be the same or the proportion of false positives to false negatives to be the same. This Article does not take a stand about *which* of these various ways of equalizing error rates is most important. It insists only that it is error rates that matter to decisions affecting action and so fairness between the two groups affected by algorithmic decision-making requires a focus on error rates.

decision-makers see a score, they won't know quite what it means. Part II explores ways that this loss of information can be lessened by taking note, within the algorithm, of the fact that some factors used by the algorithm are more predictive for one race than other.

Suppose, for example, that housing instability is more predictive of recidivism for whites than for blacks. 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. While this approach would improve risk assessment and reduce the loss of information brought about by a focus on equalizing error rates, many in the field believe this approach is off the table because prohibited by law.[18] Part II argues that this may not be the case. In particular, it argues that mere use of a racial classification does not always constitute disparate treatment.

Algorithmic tools lay bare the mechanisms of decision-making and thereby allow us to examine more precisely and in more detail which methods of thought and decision are legally permitted or prohibited by antidiscrimination law. What this analysis reveals are important ambiguities in a foundational concept of antidiscrimination law. Part II demonstrates that the concept of *disparate treatment* is blurry and elusive. This is important given the central place that the distinction between disparate treatment and disparate impact plays in equal protection doctrine and statutory antidiscrimination law.

The argument in the Article unfolds as follows. Part I presents the argument 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. To develop this argument, it begins with a detailed explanation of the two measures. It then explores the factors that affect belief and action in individual cases. Turning to the comparative context, Part I argues that parity in error rates is more central to fairness than predictive parity. This Part also considers how effects on third parties and other considerations affect fairness more broadly. It concludes by arguing that antidiscrimination law permits the equalization of error rates.

Part II explores the costs of this approach. It argues that these costs can be mitigated by using protected classifications like race and sex within algorithms. Part III concludes.

I. WHICH MEASURE SHOULD WE PRIORITIZE?

Scholars describe the dilemma as one that pits different conceptions of fairness against each other. One could therefore go on to ask, which

[18] Cite – bring footnote from later forward.

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 of 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 relevant measures. In the example I propose, I imagine that there are two hypothetical 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 with the disease 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 is represented in the columns and the predicted outcome is represented in the rows.

TRUE OUTCOME		
	Sick	Healthy
TEST RESULT	60 [a]	20 [b]
	5 [c]	15 [d]

Table 1-1 (Greens)

TRUE OUTCOME			
		Sick	Healthy
TEST RESULT	+	16 [a]	5 [b]
	-	19 [c]	60 [d]

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. Twenty of the 100 who took the test got a positive test result but are not sick. These are the false positives. 5 of the 100 who took the test got a negative test result despite the fact that they are in fact sick. These are the false negatives. And, 15 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. The probability that a Green is sick if she tests negative for the disease is (c/c+d, 5/5+15) or .25. In other words, the test is 75% accurate – as illustrated by the shaded boxes in Table 1-1 above.

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). 20 of the 100 Blues got a negative result even though they are sick (false negative) and 59 of the 100 Blues got a negative result and are healthy (true negative). As these numbers indicate, the test is 76% accurate for the Blues. The probability that 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. And the probability that Blue person is sick if she has a negative test result is 20/(20+59) = .25. 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 the tables below highlight this question. Compare.

TRUE OUTCOME		
	Sick	Healthy
TEST RESULT	60 [a]	20 [b]
	5 [c]	15 [d]

Table 1-1 (Greens)

TRUE OUTCOME			
		Sick	Healthy
TEST RESULT	+	16 [a]	5 [b]
	-	19 [c]	60 [d]

Table 1-2 (Blues)

For a sick Green who takes the test, the probability that she will get a positive result is 60/(60+5) = .92 For a sick Blue who takes the test, the probability that she will get a positive result is quite different: 16/(16+19) = .46. We get dissimilar results as well when we compare what happens to healthy Green and healthy Blues who take the test. For a healthy Green who takes the test, the test accurately provides a negative test result in 15 of the 35 cases or .43. Whereas for a healthy Blue who takes the case, the test accurately reports a negative result in 60 out of 66 times or 92% of the time.

This simple example does not quite replicate the situation described in the ProPublica exposé but is, I hope, close enough to illustrate the tension between the two measures.[19] 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 in 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

[19] COMPAS did not using 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 that 4 or 8, respectively, of 10 people given that score will recidivate if released.

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 (92%) whereas a sick Blue is not so fortunate. For the sick Blue, the test only gives the correct answer in 44% of cases. For the Blues, therefore, the false negative rate is high and is much higher than the false positive rate.

Greens:		Blues:	
False positive rate	.57	False positive rate	.07
False negative rate	.07	False negative rate	.54

If we are concerned to treat the Blues and the Greens fairly (as compared to each other, not as compared to another approach), is the test a good one?[20] On the one hand, when the test is given, it is equally accurate for the Blues as for the Greens. On the other hand, the frequency of each types of error varies. Where the cost of the two types of errors are different, the burden of the test's errors will be different for the different groups. And what are the costs of each of these errors? For a test that predicts disease, a false positive result may lead to unnecessary treatment and a false negative may lead to the failure to treat an ill person. For a test that predicts recidivism used in the context of bail or parole, a false positive may lead to unnecessary incarceration; a false negative may lead to the release of a dangerous person. So, one way to capture the issue would be to ask: which is more important: equal accuracy or equal burden?

In what follows, I will use these numbers and tables – which in the literature are called "confusion tables"[21] – 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 simple. 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 actual recidivates (sick) or does not (healthy). To make the Green/Blue example analogous to the dispute about COMPAS, the Greens

[20] How one ought to think about what fairness between groups of people requires is itself contested. One possible answer would be to say that we must treat the two groups the same. While this answer is problematic for many reasons, it is clearly unsatisfactory here. The two groups are not treated the same in some dimension. The relevant question is which dimension matters and why.

[21] *See* Berk et al. *supra* note 14 at 4 (explaining that "a cross-tabulation of the actual binary outcome Y by the predicted binary outcome $\hat{Y}$" are called, within the field of machine learning, a 'confusion table' (also 'confusion matrix')."

would be African-Americans (blacks). If a black person will recidivate, the test accurately predicts that result 92% of the time. If they will not, the test's accuracy falls to 43%. The results for whites (the Blues) is almost reciprocal. If they will recidivate, the test is only accurate 44% of the time but if they will not, the test accurate yields that prediction in 92% of the cases. Yet, as with the disease case, for both blacks and whites, a risk score of high risk is 75 or 76% accurate. Let me reiterate: my application of this data to an example dealing with parole decisions is entirely fabricated. I use it to make a point and because it is shares features (though exaggerated) with the COMPAS example.

TRUE OUTCOME		
SCORE hig	Dangerous	Peaceful
	60 [a]	20 [b]
	5 [c]	15 [d]
Table 1-1 (Blacks)		

TRUE OUTCOME		
SCORE	Dangerous	Peaceful
High-risk	16 [a]	5 [b]
Low-risk	19 [c]	60 [d]
Table 1-2 (Whites)		

Does this hypothetical risk assessment tool treat blacks fairly as compared to how it treats whites? The first response to the ProPublica exposé was that the algorithm should be adjusted so as to treat blacks and whites equally in both dimensions. However in a series a papers, scholars demonstrated that this is impossible except under highly specific circumstances that are likely to be rare in practice.[22] As Kleinberg and co-authors explain: "Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases."[23] It is impossible to equalize both measures because of the difference in base rates.[24] In the disease hypothetical, the Greens are sicker than the Blues (65% of Greens are sick while only 35% of Blues are). Similarly, when that 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 really will commit crimes if released than will whites (if this data were accurate). This is also the case in the data relied on by Northpointe. 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 make the point

[22] *See supra* note 14.
[23] Kleinberg, et. al, *supra* note 14 at 3.
[24] The term "base rate" refers to [insert definition].

clear and accessible. Note however that the COMPAS itself does not take race into account. Rather, it is the fact of differences in recidivism rates between blacks and whites that leads to the fact that the test cannot both equalize predictive accuracy for both groups and equalize the error rates for each group.

One important 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 recidivism rates are not – and indeed cannot – report *actual* recidivism. Instead they report arrests. 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 do not reflect the trait they purport to measure and thus should be viewed skeptically.[25] This is a point made frequently by critics of the use of algorithms and of the data on which they are trained.[26] This problem, called "measurement error" in the computer science literature[27] is an important issue and one that those who rely on arrest statistics, for example, must take into consideration. It is not, however, a criticism 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 makes the same point. In *Craig v. Boren*,[28] 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. The state defended the law by arguing that young men have higher rates of drunk driving than do young women. 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 as "'reckless' young men who drink and drive are transformed into arrest statistics, whereas their female counterparts are chivalrously escorted [home]."[29] Unavoidably, arrest statistics reflect both actual

---

[25] 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.* Sandra Mayson – others?

[26] *See* e.g. Pauline Kim, *Auditing Algorithms for Discrimination*, 166 Penn L. Rev. Online 189, __ (2017), Abigail Z. Jacobs and Hanna Wallach, *Measurement and* Fairness, ACM Conference on Fairness, Accountability, and Transparency, FAT*, 2019 (emphasizing the gap that exits 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").

[27] *See* e.g. Sharad Goel, Ravi Shroff, Jennifer Skeem and Christopher Slobogin, *The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment*, draft paper (on file with author) at 7.

[28] 190 U.S. 429 (1976).

[29] *Id*. at 203.

offending rates and policing practices.

The potential for bias in the data that both people and machines rely on is certainly important[30] and provides a reason to be skeptical about some base rate data.[31] To start, I put this concern aside. I begin by assuming the differences in the base rates reflected in the data are not distorted by biased law enforcement. I make this assumption not because I think it is true. Rather because it allows us to see and examine a different and equally important controversy. Later, I revisit the possibility of biased data and examine how that affects the analysis.

So far, and drawing on the ProPublica controversy, I have focused on two different measures that could be equalized. The scores produced by the algorithm could be equally predictive or the error rates produced by the algorithm could be equal.[32] But for simplicity, and because the heart of the controversy appears to focus on those two types of measures, I will limit our focus to a comparison between these measures. Different scholars use different names to describe the two measures (or a close enough variant for our purposes).[33] Alexandra Chouldechova uses the term "predictive parity," to describe the situation in a black person and white person with the same score are equally likely to recidivate.[34] Because I find her term the most

---

[30] 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 Cal. L. Rev. 125 (2017).

[31] Whether nondiscrimination norms require a skepticism about base rate data about protected groups beyond what good epistemic practice requires is a subject I explore in another paper. Deborah Hellman, *The Epistemic Commitments of Nondiscrimination*, draft on file with author.

[32] For example, Berk, et al. consider six different measures which could plausibly be measures of algorithmic fairness in their view. Berk *supra* note 14 at 13-15.

[33] For example, Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan characterize the property of equal accuracy of the score across groups as "calibration within groups" and define it as follows: "conditioned on the bin to which an individual is assigned, the likelihood that the individual is a member of the positive class is independent of the group to which the individual belongs." *See supra* note 14 at 4. More formally, they define calibration with 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 14 explaining 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?" Sharad Goel and coauthors call it simply "calibration." Goel, *et al. supra* note 27 at 9 (defining "calibration" as the requirement that "outcomes are independent of protected attributes after controlling for estimate risk").

[34] Chouldechova, *supra* note 14 at 4 (defining predicative parity as follows: "A score $S= S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among

accessible and illustrative, this Article will use that term. 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. Alternatively, we could equalize the error rates. Scholars also have different terms for the situation in which these are equal. For example, John Kleinberg uses 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.[35] Chouldechova uses the term "error rate balance,"[36] a term which I again find most accessible and so will adopt in this Article.

To summarize, algorithms are used to predict some endpoint of interest – sickness, recidivism, etc. These algorithms avoid the use of classifications that are protected by anti-discrimination law, like race or sex. However, when the groups defined by protected traits have different rates of the target trait, it will be impossible to have parity between the groups along all the possible dimensions of interest. We have focused on two of those dimensions. The algorithm can exhibit *predictive parity* such that a score 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 balance* such that people of each group who have or lack the target variable are equally likely to be accurately scored by the test.

Which measure should be preferred? The usual answer is *predictive parity*.[37] For example, Kleinberg and co-authors claim that "[a] first basic goal in this literature is that the probability estimates provided by the algorithm should be *well-calibrated*" both as a whole and "this condition should hold when applied separately in each group as well."[38] The

---

high-risk offenders is the same regardless of group membership").

[35] Kleinberg, et. al. *supra* note 14 at 2. He defines "balance for 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 5. Berk calls this "conditional procedure accuracy equality," Berk, et. al. *supra* note at 14 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 call is "classification parity," Goel, et. al. *supra* note 27 at 9 (defining "classification parity" as that "certain common measures of predictive performance (like false positive or negative rates) be equal across groups defined by the protected attributes").

[36] Chouldechova, *supra* note 14 at 4 (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 rate and false negative error rates are equal across groups").

[37] *See e.g.* Kleinberg, Mayson,

[38] Kleinberg, *supra* note 14 at 2.

terminology is different, but Kleinberg preference for calibration for each group is equivalent to predictive parity.[39] In what follows, I develop the argument for the alternative approach. While there are different ways that we might focus on error rates, the Article argues being fair to each of the groups at issue requires attention to error rates.

B. Belief versus Action and Why it Matters

The fact that we cannot have both *predictive parity* and *error rate balance* is most circumstances leads to the question: which should we prefer and why? That question focuses on whether *equal* predictive accuracy or *equal* burden is more important. Before we tackle that question, it is helpful to step back and focus on the epistemic and practical significance of both predictive accuracy and the burden of errors in *individual* cases, where no comparative question is on the table. After all, not all lack of parity between groups, even protected groups, is important. If you learned that the data about blacks was recorded in blue ink and the data about whites was recorded in black ink, this would hardly matter. "A difference, to be a difference, must make a difference," after all.[40] A better understanding of the significance of loss of predictive accuracy and changes in error rates in individual cases will help us understand the significance of the unequal predictive accuracy and unequal error rates for fair treatment of the groups affected.

1. Accuracy and burden in 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 an answer to the following question: Given this evidence (the test result), what should I believe? In that example, 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 examples.

Leslie, the baby and the bat: One day, Leslie found a live bat in her

[39] Aziz Huq also argues for the position defended here but for different reasons. Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L. J. __ (2019) (forthcoming). Huq's defense of privileging error rate balance (he uses different terminology) over predictive parity is specific to the criminal justice context and the effect on blacks.

[40] This quote is attributed to Gertrude Stein, cite.

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 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 putting 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.[41] 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. And, that remains trues in cases in which those costs are less dramatic and less extreme than in this example.

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 conviction, she must believe *beyond a reasonable doubt* that John punched Bill in order to vote to convict John of assault. By contrast, if Bill is suing John for the tort of assault (a civil claim), Sue should 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 more confidence in her belief that he did than that.

What explains the difference between the criminal and civil context is the cost of mistakes in each context.[42] In the criminal case, the cost of a

[41] Some philosophers argue that pragmatic and moral considerations also affect belief. *See e.g.* [input citations].

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

false positive (convicting an innocent) is extremely high and much higher than the cost of a false negative (letting a guilty person go free) in the judgment of our society, as evidenced by the fact that we set a very high burden of proof for the criminal context. 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 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.[43]

When we lose predictive accuracy, what will be our answer to the question *Given the evidence provided by the test or algorithm's score, what should I believe*? The answer may well be: I don't know. Loss of predictive accuracy compromises knowledge or, to be more precise, we lose confidence in the information provided by the algorithm.[44] But, as *Leslie, the baby and bat* demonstrates, we may still know how to act.

Now compare that situation to one in which the rate of false positives or false negatives rises or falls. Suppose we are confident that a given medical test is highly accurate. But are uncertain about the rate of false positives or of false negatives. This may not matter to decision and action in some cases as well like those in which the cost of one type of error is so high as to dwarf all else. Because rabies is fatal if untreated, even if a rabies test administered on the baby had an uncertain false positive rate, still we would treat the baby for rabies. This is because we would treat unless we were nearly certain that the baby does not have rabies. In other cases, the error rates make a difference. *Different legal standards* illustrates that fact. In a criminal case, the famous aphorism, usually attributed to Blackstone,[45] cautions that it is better that 10 guilty men are freed than that one innocent is wrongly convicted. The level of confidence a jury must have in the judgment that the defendant committed the offense derives from this view about the importance of a low false positive rate. If

[43] Again, some philosophers believe that the cost of error is relevant to belief as well. *See e.g.* Michael Pace, "The Epistemic Value of Moral Considerations: Justification, Moral Encroachment, and James' 'Will to Believe,'" *Noûs* 45 (2011): 239–268. 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.

[44] Another way to express this idea is to say that our "credence" is lowered.

[45] cite

the false positive rate were to change, our trial procedure would no longer be justified.

2. What is lost by forgoing predictive parity?

With a clearer sense of the significance of accuracy and error rates in the individual case, we can now ask about the comparative context. We are focused on fairness and, in particular, on treating the two groups at issue fairly vis-à-vis each other. What we want to know is this. When we lack predictive parity, do we thereby compromise fairness between blacks and whites scored by the algorithm? When we have an imbalance in the error rates, do we thereby compromise fairness between blacks and whites scored by the algorithm? And if fairness is implicated in giving up each, which is the more serious fairness problem?

To answer these questions, we begin by focusing on what is lost if we forgo predictive parity. 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. More precisely, and to borrow a philosophical term, the doctor has a credence of .75 in the proposition that the patient has the disease.[46] Since the test exhibits predictive parity, it is equally accurate for Greens as for Blues. Why is this important? Two possibilities come to mind. First, perhaps treating blacks and whites fairly requires that an equally accurate assessment tool be used on each group. Second, perhaps predictive parity is important because forgoing it leads to a loss of information which impedes rational decision-making. I consider each possibility below.

a. Fairness and equal accuracy

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, *by itself*, mean that it is unfair? To assess this question, consider the following example.

The 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

[46] The term "credence" is one used by epistemologists. For example, Sarah Moss defines it in this way: [insert]

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% essay questions and 25% multiple choice.[47] In such a case, the grade on the test means something different for women test takers than it does for men test takers. In particular, the exam is a more reliable indicator of actual knowledge for men than for women. In such a case, has either group been treated unfairly? I think it is hard to answer that question without knowing more. The test is less *accurate* for women, but in what way is it less accurate? Does it give them better scores than they deserve, less good scores than they deserve, or does it skew equally in both directions? Surely this information matters to assessing whether the test is fair to women. But when we go on to answer these questions, we are no longer just focused on predictive parity but instead have turned to error rates. If knowledgeable women and knowledgeable men are equally likely to have the test result fail to reflect their knowledge and unprepared women and men are equally likely to have the test record them as knowledgeable, then the test treats each group fairly. In other words, it isn't the inaccuracy itself that matters for fairness, it is how the inaccuracy operates. But if we focus on how it operates, we have shifted our attention to error rates.

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 women students the test is a less accurate indicator of knowledge than it is for men but that the manner in which it is less accurate is that it produces more false positives – i.e. more women who don't know the material well get good grades. In one sense women are benefited by this loss of predictive accuracy. But in another sense, they are harmed. For the well-prepared female student who would have done well on either sort of exam, she loses the ability to distinguish herself from other female test takers who do as well, even though they know less. This individual woman is surely harmed by the fact that the test is less accurate for women than for men. 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 male test-takers, who are also inappropriately grouped together with less prepared female test takers, are also unable to separate themselves from these less well-prepared female test takers. If this

[47] If the professor makes this choice in order to disadvantage one group or another, this is likely to be legally problematic as intentions are relevant under current antidiscrimination law. *See* cases, Schwartzman, but see Fallon. Whether intentions matter to permissibility from a moral perspective is more controversial. *See* Scanlon, Hellman. To make this concern irrelevant, suppose the professor makes this choice for other reasons – either good pedagogical reasons or bad reasons, like laziness.

is correct, women as a group haven't been treated unfairly as compared to men as a group. 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 women).

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, from the good, etc. This is not a claim about whether one group (women) is being treated fairly vis-à-vis the other. In fact, it isn't a comparative claim at all.[48] Rather it is a claim to a right the best available decision-making tool. Whether this is a good claim – legally or morally – I find doubtful. [49] But what it is not is a claim of unfairness between groups.

b. Fairness and loss of information

The second reason that forgoing predictive parity may fail to treat one group fairly vis-à-vis the other relates to the loss of information that lack of predictive parity entails. Consider. If we were to favor error rate balance and thereby lose predictive parity because we cannot have both, then a positive test result would mean something different for a Green person than it would for a Blue (to return to the disease test hypothetical from the start of the Article). And if we do not or cannot know whether a person is Green or Blue, we will not know how confident to be that a positive test result means that the person has the disease. In other words, predictive parity affects what we should believe about a person based on the evidence that the test provides, if we must form that belief without knowing whether the person is Blue or Green.[50] Loss of predictive parity, when we do not know what group a person belongs to, leads to inferior information. This loss of information may compromise decision-making about treatment.

Kleinberg and coauthors argue in favor of privileging predictive parity for precisely this reason. They argue that predictive parity is important because if it exists it means that "we are justified in treating people with the same score comparably with respect to the outcome, rather than treating

[48] I describe the difference between comparative and non-compartive conceptions of justice and how they relate to claims of wrongful discrimination in Deborah Hellman, *Two Concepts of Discrimination*, 102 VA. L. REV. 895 (2016).

[49] Cite my book, chapters 4 and 5.

[50] As Kleinberg, et al. explain "calibration within groups" (or predictive parity) "asks that the scores mean what they claim to mean, even when considered separately in each group." Kleinberg, et al. *supra* note 14 at 4.

people with same score differently based on the group they belong to."[51] Is this correct? Consider the disease case again. Let us further suppose some facts about the disease in question and the burdens of treatment. For simplicity, I will assume that the treatment always works. If the disease is life-threatening if left untreated, and the treatment itself is not too burdensome, then we are always going to treat and it will not matter whether we know exactly how accurate the test is. In such a case, we need not be bothered about the fact that the score means something different for a Green than for a Blue. This is the lesson of the bat case with which I began (though rabies shots are painful). Even if we do not have predictive parity and so are unsure exactly how likely it is that the person has the disease, we still want to treat the person for the disease. Similarly, where the treatment is costless or even pleasant (frequent massages, for example), then treatment will likely be recommended even if we don't know how likely it is that the person has the disease and so the loss of predictive parity will not threaten our ability to know how to treat a person in such a case either.

The point to stress here is that without predictive parity, we do not know what to believe about whether the person has the disease without knowing whether she is a Green or a Blue. Or, more precisely, we may not know how much credence to have in the proposition that the person has the disease, without knowing whether she is a Green or a Blue. But we may well know how to treat her nonetheless. And this is so because the decision regarding how to act rests not only on what we believe to be the case but also on the costs of each type of error we might make, as well as the likelihood of those errors.

The examples I use to make this point involve very high costs or very low costs for each types of error. Clearly many cases will fall in between where, to continue with the disease example, the burdens of treatment and of the disease if left untreated are more moderate. In such cases, it will be helpful to know both what the test result means more precisely for the person involved and the costs of each type of error we might make.

But not always. Sometimes in these cases too, we can get by fairly well without predictive parity. If we give up predictive parity and we don't know whether the person is a Green or a Blue, then there will be two possible values, X and Y, for the credence we should have in a positive test result. In the cases where X and Y are fairly close together, it is possible that the costs associated with treatment and with the disease make it the case that we *still* know what to do, even if we don't know whether the person is a Green or Blue.

---

[51] *Id*. at 4-5.

In other cases, the disparity between what the test result indicates for a Blue versus for a Green may be great enough and/or the differences in the costs of each type of error close enough that we will not know how to treat a person without knowing with precision how accurate the test is. The first point I want to emphasize is that this is only sometimes the case. But if the question is which parity we can afford to give up – predictive parity or error rate balance – what this discussion illustrates is that if we are focused on how to treat people, predictive parity is sometimes expendable.

To summarize, loss of predictive parity leads to loss of information in those contexts in which we do not, or cannot, know what group a person belongs to. This loss of information affects our ability to form beliefs, or the confidence we have in those beliefs. This may be a practical problem where information is useful. But the first point to stress is that the effect that loss of predictive parity has on our beliefs is not (or not obviously) a problem of fairness, as I argued in Part I.B.2.a above.[52] Second, the loss of information is not even a practical problem in many situations. There will be a range of cases in which decision-makers know how to act even though they are less than certain about what an algorithmic score indicates, as this part has shown.

3. What is lost by forgoing error rate balance?

Predictive parity and error rate balance are both characterized in the literature as competing conceptions of fairness.[53] But, as we have seen above, a lack of predictive parity does not compromise fairness between groups. What of error rate imbalance?

Return to the case of *pedagogical choice*. When discussing whether lack of predictive parity was unfair to either women or men, I argued that it would be difficult to answer that question without knowing more about whether the test produces more false positives for women (high grades, despite lack of knowledge) or more false negatives (low grades, despite

[52] If loss of information has effects on third parties, this could raise issues of fairness to those third parties. This concern is addressed in Part I.D.1 *infra*. However, fairness to third parties is a different kind of concern than fairness to the two groups scored by the algorithm. Alternatively, perhaps there is unfairness in the situation in which an algorithm is more accurate for one group than for another. This claim is addressed in Part I.D.2 *infra*.

[53] *See e.g.* Berk, et. al, *supra* note 14 at 3 (claiming that "when attempts are made to clarify what fairness can mean, there are several different kinds that can conflict with one another…"); Kleinberg, et. al. *supra* note 14 at 4 (describing calibration within groups (which is similar to predictive parity), balance for the negative class and balance for the positive class (which together make up error rate balance) as "each reflecting a potentially different notion of what it means for a risk assignment to be 'fair'"); Chouldechova*, supra* note 14 at 3 (describing the project of her paper as one of "[a]ssessing fairness").

significant knowledge). This intuition suggests that equalizing error rates is what matters to questions of fairness. If prepared women and prepared men are equally likely to be mischaracterized by the test as unprepared (i.e. given low grades), this parity bears on whether the test is fair. How inaccuracy manifests – as false negatives or false positives – matters greatly. This is because they are different *kinds* of inaccuracy and as such are likely to have different consequences in the world. A good grade for an unprepared test taker is so much less important *to that person* than is a bad grade for a prepared student.

In that example, I singled out *false negative* rates and asserted that balancing these is important for fairness in the test design. In the context of a risk assessment tool used to predict recidivism, equalizing the *false positive* rate may seem more pressing to fairness between legally protected groups. We want to know whether peaceful blacks and peaceful whites are equally likely be mischaracterized by a risk assessment algorithm. Again, this is because getting out of prison is desirable, staying in is burdensome. What matters for fairness is equalizing the burden (when there is one) of mistakes. Because the costs of false positives and false negatives are unlikely to be the same in any given context, equal predictive accuracy does not produce fairness if it is achieved by more false positives for one group and false negatives to the other. What matters to the individuals assessed by algorithmic tool *how* it mischaracterizes them, when it does, not just that it mischaracterizes them sometimes.

The main claim of this part is that it is parity in the dimension of error rates that matters most to questions of fairness between groups. Above, I also offered some tentative thoughts about which error rates might matter most in particular contexts. Those thoughts are speculative. The next step in this analysis would focus on *when* balancing false positive rates, false negative rates, both or the ratio of one to the other is most significant in a given context and why. I leave this question for another day. Suffice it to say that the costs associated with each type of error in the given context is likely to make one of those measures more apt to issues of fairness.

How people are treated is what matters to questions of fairness. If these tools are used in the contexts of decisions and action, these decisions will have consequences. Fairness requires attention to these consequences. A tool that leads to more peaceful blacks remaining in jail than peaceful whites is not made more fair by the fact that more dangerous whites are released than dangerous blacks. The compensating false negative error (dangerous whites who are scored as low risk) helps to achieve predictive parity but does nothing toward making the tool more fair.

At the risk of being glib, it may well make sense to say that predictive parity treats scores equally and error rate balance treats socially salient

groups of people equally. The first metric is a measure of the meaning of information. The parity it achieves is conceptual. The second metric is a measure of the treatment of groups. The parity it achieves is practical.

C. Optimization and its impact

Decision-making methods, whether algorithmic or non-algorithmic, must make choices about how to balance the different types of errors they might make. Sometimes false positives are most problematic. Sometimes false negatives. Designers can adjust the tool they are using to be care more about avoiding one form of error than the other. 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 positive to be high. If the algorithm picks out someone as a potential terrorist who is not, very little is lost. 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 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 positive. If the result of identifying everyone as a potential terrorist is that everyone will be searched, then this may well be the best policy to adopt. In fact, the search everyone approach has benefits that derive precisely from its uniformity. There is no stigma in being identified as a "potential terrorist" if everyone is identified in the same way, though there will, of course, be costs in terms of expense (to hire searchers) and inconvenience for travelers.

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": "Better ten guilty people go free than that one innocent person is convicted." This ratio is arrived at by determining the cost to the community of the risk involved in releasing a guilty and potentially dangerous person into the community 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 values the harm of erroneously incarcerating an innocent so highly that this ratio is arrived at.

In assessing whether an algorithm treats blacks and whites equally, one way we might assess this is to focus on whether the tool strikes the same balance between the costs of false positives and false negatives for blacks

and whites in the given context. It would be unfair if we treat blacks like terrorists and whites like Englishman, to use a colorful analogy. Yet COMPAS seems to go a fair way in that direction. Because false positives outweigh false negatives for blacks and false negatives outweigh false positives for whites, the algorithm expresses a value of the relative costs of each. This is particularly worrisome where, as here, otherwise the contexts are likely to be quite similar. In both contexts, there is a risk in releasing a dangerous person and a harm in failing to release someone who is peaceful. What distinguishes the cases in the race of the individual held in custody. Equalizing the ratios of false positives to false negatives for blacks and whites would express that society holds constant the value it assigns to the cost of incarceration for the white and black individual.

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 typs 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. When we adopt a decision-rule that incorporates such a balance, we assume that the costs are born by scored individuals *as a group* or society *as a whole*. But, as we have seen, that is not always the case. Where the costs of the more burdensome type of errors are born more by one subgroup of the population than another, we do not treat members of the group fairly. In a very real sense, it is *as if* we applied a different rule.

D. Fairness All-Things-Considered

There is an important ambiguity in the literature regarding these different measures of algorithmic fairness. Sometimes the notion of fairness is focused on the following question. Does the tool treat blacks fairly as compared to whites? Or whites fairly as compared to blacks? That is the question that this Article has so far focused on. But the notion of fairness is a capacious one and loosely deployed. It might also encompass other notions of fairness. These other notions of fairness include fairness to third parties as well as broader notions still. I consider some of these below.

1. Fairness to third parties

In addition to asking whether an algorithmic tool treats one subgroup of those whom it scores fairly as comparted to another, we might ask instead

whether it treats those whom it scores fairly as compared to affected others. Borrowing from contract law, we might term these others "third parties" as they are not party to the scoring mechanism but are nonetheless affected by it.

In the context in which an algorithmic tool is used to predict recidivism risk in order to determine whom to release on bail or whom to parole, the relevant third parties include people (and their relatives and friends) who might be harmed by released accused and inmates who go on to commit crimes. In addition, the relatives and friends of people released are also affected third parties. These effects could be positive or negative. In the disease hypothetical, for example, affected third parties include people who might contract an untreated disease if it is contagious and family members and friends of affected persons who may suffer materially and psychologically from the illness of loved ones. This catalogue of possible affected third parties is illustrative but clearly not exhaustive.

The last section focused on optimization, the balance set between false positives and false negatives given the interests and costs involved in a particular context. That discussion envisioned a simplified example; it compared the harm to the individual of a false positive with the harm to the community of a false negative. When we focus on how third parties may be affected by how the balance between false positives and false negatives is struck, we should also recognize that the community can also be harmed by a false positive and the individual could be harmed by a false negative. The harms to the community of a false positive include the costs – both human and monetary – of unnecessary treatment (in the disease context) and unnecessary incarceration (in the criminal justice context). The harm to the individual of a false negative are obvious in disease case (under-treatment) but possible in the criminal justice context as well. When we speak about algorithmic *fairness*, rather than asking about whether a protected group and its counterpart (blacks and whites, women and men) are treated fairly vis-à-vis each other, we might instead be asking about whether the way the algorithm is designed and used sets the right balance between the interests of the scored individual and those of society more generally, including the affected third parties.

The interests of third parties will be affected by any loss of information produced by forgoing predictive parity. For example, when a score is less meaningful for one group than it is for another or when the score is less meaningful than it could be, decision-makers using the algorithm may make decisions about whom to release that are different than they would make with more accurate information. In the context of releasing people from prison, the effect of this loss of information might be releasing more people or different people than would be released if decision-makers had more

accurate information for all groups. The people who are released may recidivate and in so doing may harm third parties whose interests matter also to any all things considered calculation of what one ought to do. If abandoning predictive parity leads to a loss in predictive accuracy, then predictive parity may matter to fairness to third parties.[54]

This is a genuine concern and does provide a reason that counts against abandoning predictive parity. But there are two caveats worth noting. First, we should not confuse all things considered conceptions of fairness with fairness between two groups. The argument of this Article is focused on the latter issue. The goal, in Part I, is to show that fairness between groups requires attention to error rate balance not to predictive parity. However, what matters at the end of the day is an all things considered judgment about which measure to prefer. So, we might say that fairness between groups weighs in favor of error rate balance and harm to third parties or the interests of society will sometimes count on the other side of the ledger. When it does, one will need to weigh up these concerns, and others, and determine what to do. Second, when we turn our attention to an all-things-considered perspective, fairness between groups and fairness to their parties are not the only fairness-related concerns. The next section provides another example of a fairness related concern that would enter such an all-things-considered judgment.

2. Automating Bias and Compounding Injustice

Once we open the door to other sorts of fairness, fairness to third parties is not the only relevant concern we let in. We might also worry about the fairness of taking the facts as they are as a starting place, especially when those facts themselves are the product of injustice. This concern includes two related ideas. First, the data on which the algorithm relies might be an inaccurate reflection of the underlying facts. 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 is called "measurement error"[55] and is most likely what Representative Ocasio-Cortez had in mind when she claimed that algorithms just "automate the bias."[56] Second, the data may themselves be accurate but the disparities they reflect may themselves be *caused by* 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

[54] **How does the point is this paragraph relate to the point in the prior paragraph?**
[55] cite
[56] cite

recidivism, it may use the fact that blacks were unfairly treated in the past to justify treating them worse today. That seems problematic, as explained below.

Consider the problem of *automating the bias* first. Sometimes the data on which algorithms rely does not accurately reflect the trait it purports to reflect. 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 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 ways. When they are, we should do what we can to combat these biases. This is an issue that has attracted significant attention in the both the popular press and the academic literature. For example, Sandra Mayson argues that in the criminal justice context, predictive algorithms should use arrest for violent crime rather than all arrests as an input but this data is likely to be more reliable.[57]

Alternatively, we might adopt prophylactic measures that aim to compensate for bias in our data, even when we cannot be sure how much of it exists exactly. Just as a driver may steer slightly to the right when he knows that a deflated tire is pulling him left in order to achieve the aim of driving straight down the road, we could adopt measures that pull us away from bias. The Rooney rule, a National Football League rule, requires League teams to interview at least one racial or ethnic minority candidate for all head coach positions, provides a good example. If one worries that an unconstrained algorithm (either the kind that operates informally in a person's head or the kind that is automated) relies on biased inputs, the Rooney rule counteracts that bias. Interestingly, in many circumstances, this constraint improves decision-making, as judged only by reference to the hiring of a coach with a particular set of skills and capacities.[58] The reason that the rule helps decision-makers to achieve their aims is due to the fact that it compensates for bias in the inputs that they themselves may not recognize.

Second, consider the problem of compounding injustice.[59] Suppose that inmates who have themselves been victims of child abuse are more likely to recidivate than those who have not been victims. The parole board might take that factor into consideration when making parole decisions. If so,

---

[57] cite

[58] Cite paper on Rooney rule from FAT.

[59] I argue that 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, (T. Khaitan, editor, Hart Publishing, 2018). This example is drawn from that chapter.

there is no inaccuracy or error if victims of child abuse *are* more likely to recidivate. But something seems wrong about this practice. The fact that this person is more likely to recidivate is due to the fact that he has himself been the victim of injustice. 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.

In some race and sex discrimination cases, the Supreme Court appears to adopt this rationale for invalidating a law. Consider, for example, *Califano v. Goldfarb*.[60] There, the Supreme Court struck down a statute providing that the spouses of men would automatically qualify for social security benefits but requiring spouses of women to show they were dependent in order to qualify for the same benefits. In 1977, when the case was decided, gender was likely a very good proxy for dependency. In that sense, there was no measurement error. Nevertheless, the Court invalidated the law.[61] Had the gender distinction in the law been permitted to stand, the statute would have compounded the societal injustice that led to the fact that women were more likely to be dependent on their spouses than men by providing female wage earners less generous social security benefits than male wage earners.[62] When differential base rates are themselves the result of prior injustice, the practices that perpetuate these disparities risk compounding that prior injustice. In order to avoid replicating and reinforcing prior injustice, we may have a special obligation to ensure error rate balance.

Unfairness to third parties, automating bias and the concern about compounding injustice, as well as other possible fairness considerations, would all need to be included in an *all things considered* evaluation of what fairness requires. This is no easy task. What is clear from the discussion is that several additional factors would be relevant and that these factors could weigh on different sides. For example, fairness to third parties would plausibly count in favor of predictive parity and the unfairness of compounding injustice would plausibly count in favor of error rate balance. As it is difficult to say how such concerns would be weighed in every context, an all things considered evaluation is unlikely to favor one measure over another in general. Instead, we should note that other fairness considerations could be in play and be attentive to them as they arise.

---

[60] 430 U.S. 199 (1977).

[61] *Id*. This case is similar in rationale to *Frontiero v. Richardson*, 411 U.S. 677 (year), (invalidating a presumption that the spouses of male service members were dependent but requiring the spouses of female service members to prove dependency to qualify for benefits) and Weinberger v. Wiesenfeld, 420 U.S. 636 (year) (invalidating a law that restricted so-called "mother's benefits" to widows).

[62] Cite race case discussed in my prior article.

If we limit our focus to the more discrete inquiry regarding fairness between socially salient groups of people who are scored by the algorithmic tool, there is something useful to say. If we are wondering which measure – predictive parity or error rate balance – does a better job of ensuring that each of the relevant groups is treated fairly vis-à-vis the other, we should favor error rate balance for two, related, reasons. First, because error rates affect what we ought to do and not what we ought to believe, they more clearly relate to issues of fairness than do differences is the meaning of scores. Second, the costs of each type of error and the balance between them indicates how the algorithm values the individuals affected. We treat groups fairly when we balance those costs in the same way for each group.

E. Does the Law Constrain the Choice?

An analysis of which measure one should prefer and why would be irrelevant if the law forbids or requires either measure. This section considers that question. U.S. antidiscrimination law is organized around a distinction between two forms of discrimination: disparate treatment and disparate impact.[63] In cases of "disparate treatment," a law, policy or practice draws a distinction among people on the basis of a protected trait.[64] In cases of "disparate impact," a law, policy or practice treats everyone the same or, alternatively, it draws a distinction among people on the basis of a non-protected trait. In so doing, the law or policy affects members of a protected group in a different or worse way than it affects others.[65] In U.S. law, disparate treatment is the more serious offense. As a matter of constitutional law, only disparate treatment gives rise to strict scrutiny. As a matter of statutory law, both disparate treatment and disparate impact discrimination are potentially prohibited but, in fact, much disparate impact is found to be justified under the reasons permitted by law.[66] Thus, central to determining the legal status of algorithmic tools that lack either predictive parity or error rate balance will be the determination of how to categorize these effects. Do they constitute disparate treatment on the basis of race (or some other protected trait) or do they merely give rise to a disparate racial impact? In this part, I argue that both situations should be understood as forms of disparate impact. The upshot of this result is that neither is legally prohibited and thus policy makers are free to choose which

---

[63] In other countries, especially in Europe and Canada, the preferred terms are "direct" and "indirect" discrimination. *See e.g.* THE ROUTLEDGE HANDBOOK OF THE ETHICS OF DISCRIMINATION, Kasper Lippert-Rasmussen, editor, (2018).

[64] *See supra* note **Error! Bookmark not defined.**.

[65] *Id.*

[66] Cite Michael Selmi.

measure to prefer.

A brief primer on U.S. antidiscrimination law may be helpful first. Most laws 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 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 classifies on the basis of particular traits or affects groups defined by those traits, does antidiscrimination 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 antidiscrimination law). As a matter of constitutional law, this list of traits is more limited than under statutory law. For example, in the United States disability is not a protected characteristic as a matter of constitutional law[67] but is as a matter of statutory law.[68] 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 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 use of risk assessment tools by states and localities to determine whom to release on bail or whom to release early from prison is governed by Constitutional law.[69]

Disparate treatment on the basis of both race and sex give rise to heightened judicial review and are thus both disfavored by U.S. Constitutional law. For simplicity, I will focus here on race.[70] 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 *intention* or racial *classification* that is the "touchstone"[71] of an equal

---

[67] Cleborne v. Cleborne Living Center.

[68] Cite ADA

[69] An extension of the analysis presented in this Article would focus instead on statutory antidiscrimination 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.

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

[71] Washington v. Davis, 426 U.S. at ___ (insisting that "[d]isproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination

protection violation is controversial.[72] Sometimes the Supreme Court emphasizes classification[73] and sometimes the Court emphasizes intention.[74] 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 applied. The Supreme Court's affirmation 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.[75] This is true despite a remedial or other benign purpose for adopting policy. Yet, 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.[76] Is an invidious intention the condition that offends the Constitution or is it racial classification?[77] The normative foundation of equal protection jurisprudence is uncertain.

What is certain is that without racial classification or invidious intention, a law or policy does not constitute disparate treatment on the basis of race. With this background in mind, we can now see why lack of predictive parity and error rate imbalance are each forms of disparate impact. Neither involves explicit racial classification. Rather, both measures call attention to the disparate racial impact of utilizing facially neutral measures. If a tool lacks predictive parity, then the same score will mean something different for blacks than it will for whites. Like in the hypothetical law school exam example discussed earlier, the particular test was more meaningful for men than for women. The exam format thus produced a disparate impact on women. The screening tool itself was facially neutral however, as it consisted of a combination of multiple choice

---

forbidden by the Constitution").

[72] Cite one case for each view.

[73] *See infra* note 75 and accompanying text for a discussion of the way in which the Supreme Court's affirmative action jurisprudence supports this conclusion.

[74] *See supra* note __ and accompanying text for a discussion of the Supreme Court's rejection of disparate impact as alone sufficient to give rise to strict scrutiny and emphasizing the importance of intention.

[75] Cite *Grutter* and *Gratz*.

[76] Interestingly, despite the fact that the Supreme Court says it treat actions that are motivated by good and bad intentions the same, when the state employs a facially neutral classification in order to benefit a disadvantaged group, the Court has allowed – even lauded – such efforts. The policy of admitting the top 10% of high school graduates around the state of Texas to the state's universities is a good example. *See* cite Grutter/Gratz endorsing this policy.

[77] Cite my Two Concepts of Discrimination

or essay questions that produced this disparate impact. Such a facially neutral screening method only gives rise to strict scrutiny if it is adopted in order to produce the disparate impact, "because of" the disparate impact and not merely "in spite of" these foreseeable consequences.[78]

The fact that a tool produces error rate imbalance is also a form of disparate impact. The algorithmic tool employed does not use a racial classification. Nor is the tool adopted in order to yield the disparate false positive burden on blacks that is produced, as *Washington v. Davis* and *Personnel Administrators v. Feeney* require.[79] At the most, one could argue – as John Hart Ely famously did[80] – that a willingness to tolerate this disparate burden on blacks bespeaks a differential sympathy that is as problematic from the perspective of the Equal Protection Clause as is a specific intent to harm. While Ely offers a powerful normative argument for this view, thus far the Supreme Court has been unwilling to find that such differential sympathy also gives rise to strict scrutiny.[81]

Part I sets up a dilemma. In most cases, it will not be possible to equalize both of two important measures with which we might assess algorithms. The first, predictive parity, insists that the algorithm be equally predictive of the trait it aims to predict for each of two legally protected groups – blacks and whites, for example. The second, error rate balance, insists that the error rates (false positives, false negatives or both) be equal for each of the protected groups. When the relevant groups differ, as they often do, in the underlying rates of the predicted trait, equalizing both measures is not a mathematical possibility. The fact gives rise to a need to choose. Which measure should those who design and implement algorithms prefer and why?

Part I argues that the answer to this question depends on being more precise about the question. If we are concerned about fairness between the protected groups, then we should prefer a measure that focuses on error rates. This is because though predictive parity is relevant to the meaningfulness of the scores, unequally predictive scores, alone, are not unfair. Error rate imbalance, by contrast, is a fairness concern because it indicates whether the inaccuracy in the measure produces the same type of

---

[78] Personel Admin v. Feeney and Washington v. Davis. 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.* Grutter and Gratz and Fisher.

[79] *Id*.

[80] Cite Democracy and Distrust

[81] Cite – Wash v. Davis again or something else.

error for each group. Because false positives and false negatives are unlikely to be equally burdensome and fairness requires equalizing the burdens of errors, it is equality in the dimension of error rates that matters to fairness between groups.

If those who design and study algorithms are focused on treating protected groups fairly vis-à-vis each other, they should focus on error rates. But perhaps worries about unfairness are not, or not only, focused on this question. Perhaps they are also concerned about a more holistic, all-things-considered conception of fairness. If so, there are several issues that will be relevant. Forgoing predictive parity will produce a loss of information which may well have costs for affected third parties. But this concern is not the only broader fairness issue. Also relevant are concerns about reproducing the errors imbedded in the data. In addition, we might worry that even accurate data may itself reflect the effects of prior injustice that we should avoid reinforcing. All this is to say that a holistic concern with fairness is likely to be complex and so it will be difficult to sort out which measure should be preferred from an all-things-considered perspective.

That being the case, perhaps we should focus on avoiding unfairness in the narrower sense and then work to mitigate the costs, both moral and practical, that this choice entails. Part II takes up this mitigation project.

II. MITIGATING THE COSTS OF OUR CHOICE

Each choice has costs. If we privilege predictive parity or privilege error rate balance, there will be effects that are undesirable. While the law does not require that we mitigate such costs, a concern with fairness more globally, suggests that we should mitigate these costs if possible. As Part I argued that we should focus on balancing error rates and thus be willing to forgo predictive parity, the bulk of this Part will be devoted to how we might mitigate the costs of that choice. Part II.B. addresses that question. But because others might make a different choice, I begin in Part II.A. with a briefer account of how one might mitigate the effects of instead adopting predictive parity.

A. Reduce the Costs of Errors

If we insist on predictive parity, then we produce error rate imbalance. One group will have more false positives; another will have more false negatives. This matters, I argued, because the burden of false positives and false negatives is unlikely to be the same for the people affected. One

strategy for mitigating the cost of this differential burden, therefore, would be to alter the consequences of these errors.

In an insightful recent article, Sandra Mayson argues for exactly this approach.[82] If the effect of classification as high risk were "greater access to social services and employment" rather than incarceration, "a higher false-positive rate among black defendants would be less of a concern."[83] 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 rate imbalance. Thus, if those who design and implement algorithmic decisions continue to insist on predictive parity, they should strive to ameliorate the costs of this choice by adjusting the consequences that flow from each type of error. 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 produces 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 politically. 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. When algorithms are used to make employment decisions or load decisions, 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. Reduce the Loss of Information

If we insist on error rate balance, we give up predictive parity. As a result, the scores that members of each group receive will not mean the same thing. This is a problem because decision-makers will lose information which could be valuable. However, this loss of information could be avoided or minimized if the algorithm could explicitly take account of the protected trait at issue.

It is the fact that current law disfavors explicit race and sex-based classifications that creates the dilemma we have been exploring. To see this, consider first an example in which protected traits are not involved. Suppose there is a diagnostic test used to determine whether a woman has

[82] Sandra G. Mayson, *Bias In, Bias Out*, forthcoming Yale L. J. (2019).
[83] *Id*. at 43.

breast cancer. If the base rates for breast cancer are different for pre- and post-menopausal women, a doctor treating such patients would ask the woman which group she belongs to and take that fact into account in interpreting test results. It would be silly not to. Even if the groups correspond to racial categories, a physician might take that into account in the health care context and would certainly consider the sex of her patient.[84] Where categories or traits about people (both protected traits like race and sex and unprotected traits like being pre- or post-menopause) are associated with different rates of the target trait, ignoring this information has informational costs.

Forgoing predictive parity makes it the case that a score means something different depending on the group to which a person belongs. One way to handle that problem would be to take group membership into account when interpreting the score. Interestingly, then, the loss of information due to lack of predictive parity only arises because current interpretations of legal requirements appear to prohibit differentiation by racial group (and perhaps other socially salient groups).[85] We can thus mitigate the loss of information that results from forgoing predictive parity by taking group membership into account within algorithms.

Most scholars assume this approach is legally prohibited.[86] Were the groups at issue defined by an unprotected trait (pre- versus post-menopause, for example), algorithm developers would segment the data into two tracks and employ different approaches for each. Where race, and arguably other protected traits like sex, is involved, most scholars presume that states and localities cannot simply have different threshold scores for blacks and whites that determine whom to release from custody.[87] Nor can they, people presume, employ different predictive traits within the algorithm for blacks and whites. As a descriptive matter, I agree that race-specific

---

[84] Monahan and Skeem emphasize this point, stressing that "not to use gender as a risk factor for various health conditions would be unimaginable." John Monahan and Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, Annu. Rev. Clin. Psychol. 2016. 12:489-513 at 502.

[85] 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." *See* Jennifer Skeem, John Monahan and Christopher Lowenkamp, "Gender, Risk Assessment, and Sanctioning: The Cost of Treating Women Like Men," *Law and Human Behavior*, vol. 40, No. 5, 580-593 (2016) at 591. Whether current Constitutional and statutory law permits such explicit gender-based classification is unclear. Monahan believes it does, cite, Sonja Starr believes it does not. [possibly address in part IV].

[86] cite

[87] Corbett-Davies, Pierson, Feller, Goel and Hug, "Algorithmic decision making and the cost of fairness," ArXiv: 1701.08230v4 [cs.CY] 10 Jun 2017 at 8 (noting that race specific thresholds would trigger strict scrutiny).

thresholds would trigger strict scrutiny as a matter of constitutional law, and that such differential thresholds would be unlikely to survive such demanding judicial review. In this section, I take up a more modest way that racial classification might be utilized in algorithms and suggest that it's legal permissibility is, at worst, ambiguous.

1. Separate Tracks Within Algorithms

The law's treatment of explicit racial classifications is more complex and nuanced than scholars writing about algorithms have thus far recognized. Racial classification is legally permitted when used for information-gathering purposes only. The fact that racial classification is legally permitted so long as the races are still *treated the same* opens the door to using race in algorithms to ensure that different racial groups are treated equally. This section develops that argument.

When we insist on predictive parity, error rates become imbalanced, as described earlier. The problem is that a peaceful black person is less likely to be correctly identified by the algorithm than is a peaceful white person. It is possible that we might lessen these errors with more fine-grained information of the following sort. Suppose that some of the traits that predict recidivism, for example, 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."[88] If so, perhaps we might have two tracks of analysis. For whites, housing stability is included in the predictive algorithm. For blacks, it is not. However, Corbett-Davies and his coauthors worry that using housing stability for whites but not for blacks would require using race explicitly in the algorithm and that doing so will raise legal problems. As a result, they report, "it is common to simply exclude features with differential predictive power."[89] The result of doing so, in their view, is to exacerbate disparate racial impact.[90]

Sharad Goel and coauthors also point out that using separate algorithms for each racial group could help to ameliorate measurement error.[91] They offer the following example. Suppose that the existence and number of past drug sales is predictive of future criminal activity. However, it is hard to have accurate information about actual past drug sales. Rather what we have is a proxy – past arrests or convictions for drug selling. If we worry that arrest and conviction data is biased by policing practices in which

---

[88] *Id.* at 9.
[89] *Id*.
[90] *Id*. (noting that discarding information may inadvertently lead to redlining effects").
[91] Goel *et. al. supra* note 27 at 7.

minority communities are more heavily policed than white communities, it might be the case that past arrests for drug selling is more predictive of future criminal activity for white than it is for blacks. 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."[92]

Joshua Kroll and coauthors, [93] building on the work of Cynthia Dwork and coauthors[94] provide another similar example.

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

In each of these examples, the fact that the algorithm must be blind to real 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

---

[92] *Id*.

[93] Joshua A. Kroll, Juanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PENN. L. REV. 685-688 (2017)

[94] Cynthia Dowrk, Moritz Hardt, Toniann Pitassi, Omer Reingold and Richard Zemel, *Fairness Though Awareness*, 2012 PROC. 3RD INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214.

the algorithm to take race into account *in the way it processes other information*, the algorithm would be a better job at its task.

Does the law in fact 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. And examining why reveals that the concept of *disparate treatment* is fuzzy and hard to define.

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 in to 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 racial two 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 by 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 suggest 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, put pressure on what the law means, precisely, by the concept of *disparate treatment*.

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, racial classification does not always constitute disparate treatment. For example, he commonplace practice of collecting information using racial categories appears not to constitute disparate treatment. As Kim Forde-Mazrui notes "it 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, criminal justice, and virtually all areas of the civic, economic, and social life of the nation."[95] 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 case did consider whether

[95] Kim Forde-Mazrui, *The Canary-Blind Constitution: Must Government Ignore Racial Inequality*, 79 LAW AND CONTEMP. PROBLEMS 53, 72 (2016).

the Census may use racial categories.[96] The United States Census collects information about the number of people living in the United States, as required by the Constitution.[97] And, in addition, the Census also collects additional information about characteristics of the U.S. population including information about race (this information is not constitutionally mandated, however). 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.[98] The collection of such information, including racial information, was challenged in *Morales v. Daley*. The Plaintiffs in that case argued that the deployment of racial categories on the Census should be subject to strict scrutiny.[99] 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. [100] 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.[101] The District Court for the Southern District of Texas upheld the use racial classification in the Census – including the requirement that people self-report their race under penalty of substantial fines. The court in *Morales v. Daley* declines to apply strict scrutiny, despite the use of a racial classification on the grounds 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 classification without a compelling interest." [102] *Collection* of information is different from *use*, in the court's view, and the former does not constitute disparate treatment and thus does not give rise to strict scrutiny.

What distinguishes a racial classification within a law or policy that constitutes disparate treatment and a racial classification that does not constitute disparate treatment relates, according to this example, to whether the racial classification leads to effects in the world that might matter to

---

[96] Morales v. Daley, 116 F.Supp.2d 801 (2000) (upholding the collection of various pieces of information by the Census, including information about race, under the Equal Protection Clause and other constitutional clauses).

[97] Article I, Section 2, Clause 3 of the Constitution of the United States requires that an "actual Enumeration shall be made with three Years after the first Meeting of the Congress of the United States, and within every subsequent Term of ten Years…"

[98] 116 F.Supp.2d. at __ (noting that the Census "has always included additional data points, such as race, sex, and age of the persons counted").

[99] *Id*. at ___

[100] *Id.* at ___

[101] *Id*. at ___

[102] *Id*. at __.

those affected. If it does not, as the data collection example demonstrates, then strict scrutiny does not apply. In addition, the Census example suggests that the effect must be the direct effect of the classification itself and not merely downstream consequences of such classification. [103] 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 classification in the Census subject to strict scrutiny.

The use of racial classifications is also not subject to strict scrutiny when the manner in which it uses race does not involve making generalizations about the way blacks, or whites, or people of some other race are. It is racial generalizations, in particular ones that are denigrating, that are legally prohibited. In order to explain this distinction, it is necessary to first lay a bit of groundwork. All classifications, including racial classifications, can be used in either of two ways: as a proxy for some other trait or not as a proxy for another trait.[104] For example, in the sex-discrimination case of *Craig v. Boren*, discussed in Part I, sex is used as a proxy for a person's likelihood to drink and drive.[105] But sex-based classifications are not always used as a proxy for another trait. For example, when a school or university adopts a single-sex policy, sex is not used as a proxy for another trait. Rather, a women's college, for example, seeks to have a single-sex study body and in fact seeks to highlight, by doing so, how varied are the women enrolled in terms of the other traits they bring. Similarly, race is sometimes used as a proxy and sometimes not. Racial profiling in policing is an example of using race as a proxy, in that case a proxy for likelihood of committing particular sorts of crime. Classic Jim Crow segregation, by contrast, does not use race as a proxy for other traits. Rather, those who adopted racial segregation in public schools and elsewhere intended to keep African-Americans from attending white schools, period, whatever other traits these students also had.[106]

---

[103] *Id*. at 814-815, explaining that "[t]he issue whether requiring a person to self-classify racially or ethnically, knowing to what use such classifications have been put in the past, can violate the due process implications of the Fifth Amendment. This court holds that such self-classifications do not"). While the court speaks of the due process clause, because we are dealing here with federal action, the Court is evaluating the implied equal protection requirements found in the Fifth Amendment's due process clause. *Cite Bolling v. Sharpe.*

[104] Cite my Two Types of Discrimination piece.

[105] Cite Craig, discussion in Part I. Sex is used as a proxy for other traits in the laws at issue in many sex-discrimination contexts. *See e.g.* Frontiero v. Richardson, cite (sex a proxy for having a dependent spouse), *Reed v. Reed* (sex as proxy for financial acumen).

[106] Describe how in affirmation action there is a debate about whether to understand

Where race is not used as a proxy, it is treated as especially legally problematic. As Justice Powell explained in *Regents of Univ. of California v. Bakke*,[107] "[if] petitioner's purpose is to assure within its student body some specified percentage of a particular group merely because of its race or ethnic origin, such a preferential purpose must be rejected not as insubstantial but as facially invalid."[108] By contrast, when race is used as a proxy, its status is more ambiguous. Sometimes it is subject to strict scrutiny but passes such scrutiny, as when race, together with other factors, contributes to the diversity of a university's study body. Just as a farm boy from Iowa may contribute a unique perspective, so too may a member of a racial minority.[109]

Most importantly for our purposes, when a racial classification is part of a generalization but not used in a way that involves generalizing about race or a racial group, the racial classification does not give rise to strict scrutiny at all. Consider, for example, the context when police rely on a suspect description that includes race. When eye witnesses or victims describe a 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 be considered to be disparate treatment on the basis of race.[110] For the person on whom police investigative efforts focus, it may well feel like disparate treatment on the basis of race.[111] Yet, as the Second Circuit in *Brown v. City of Oneonta* explains, it is not.[112]

The reason that reliance on a racial suspect description does not constitute disparate treatment on the basis of race, in the court's view, is that the police department of the City of Oneonta is not relying on a racial generalization. The police department *is* relying on a generalization, and

---

race as proxy or non-proxy – use various arguments from *Bakke*.

[107] 438 U.S. 265 (1978).

[108] *Id*. at ___. The rejection of such "discrimination for its own sake" has been reaffirmed by the Supreme Court in _____. Parents Involved good cite here.

[109] Find citation : Bakke or Grutter, I believe.

[110] *See* Brown v. City of Oneonta, 221 F.3d 329 (2d Cir. 1999) (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 does not violate Equal Protection but could violate the Fourth Amendment as race alone is insufficient to constitute reasonable grounds to arrest and search a person.).

[111] 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 (2001) (arguing….[fill in].)

[112] The Second Circuit concludes that the plaintiffs have not "identified any law or policy that contains an express racial classification" because the policy of the Police Department is, instead, to respond to the suspect description of the witness or victim, whatever it is. 221 F.3d at 337.

that generalization turns out to include race, but it is not relying on a generalization about people of a particular race and thus not employing a *racial generalization*. The police department operates according to the following policy: *follow the suspect description*, or something along these lines (or so we assume). In the particular instance, this policy led the police department to search black men because the victim of an attack described her assailant as black. Such a policy is meaningfully different from a police department policy of policing black men more heavily than white men, for example. The second policy would be based on a generalization about black men and their likelihood of committing crime. As the court in *Brown v. City of Oneonta* explained, "Plaintiffs does 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."[113] If they did, the police would be generalizing about blacks, i.e. from the trait black, they would be concluding 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 v. City of Oneonta* do rely on a generalization also but one of a very different character. They rely on a generalization about the reliability of eye witness descriptions. Their policy – *follow the suspect description* – implicitly relies on the generalization that eye witness reports are more likely to be helpful than not (or are sufficiently likely to be accurate to warrant the burdens imposed) or something of that nature.[114] Race is used within that in this particular case but this policy does not rely on a view about blacks, only a view about eye-witnesses.

These examples demonstrate that not all uses of racial classifications constitute disparate treatment or give rise to strict scrutiny. Only some do. This is important by itself. This fact shows that the concept of disparate treatment is less clearly delineated than one might initially suspect. Thus, the mere fact that an algorithm uses race in predicting recidivism should not by itself give rise to strict scrutiny. How the algorithm does so 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

---

[113] *Id*.

[114] Fred Schauer emphasizes the way in which seemingly direct evidence like eye witness reports is probabilistic in just the same way as profiles and other probabilistic evidence. FREDERICK SCHAUER, PROFILES, PROBABILITIES AND STEREOTYPES, 101-103 (2003). Interestingly, it was this generalization about the reliability of eye-witness reports about race that proved problematic constitutionally on Fourth Amendment grounds. 221 F.3d at 340-341.

Supreme Court has not weighed in on either of these examples, they may turn out to be less significant than this presentation assumes. 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.

If these examples provide guideposts for determining when the use of racial classifications constitutes disparate treatment, two principles emerge. First, the Census example suggests that the use of racial classifications must produce an effect that is proximate in order to constitute disparate treatment. Second, the suspect description example suggests that when race is used within a generalization, only racial generalizations 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 provides information that in turn can be used to determine what other traits to bring to bear. Like the racial information in the Census, this racial information is likely to have down steam 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. While the algorithm relies on a generalization about what housing stability or instability indicates for people of each race, the generalization itself is not, or not primarily, about race. As a result, there is good reason to think that the use of race within algorithms is and should be permissible.

Of course, a court may find it impermissible nonetheless – as these are fine distinctions and may strike some as splitting hairs. In addition, the current Supreme Court may be especially reluctant to give its imprimatur to the use of race by governmental officials. That said, unless the same Supreme Court is willing to repudiate the use of race in the Census or when relying on suspect descriptions, the inconsistency between those uses of racial classifications and a blanket prohibition will require explanation.

2. *Ricci*'s Irrelevance

Some scholars[115] appear to think that modifying an algorithm to avoid a

[115] Kroll, *supra* note ___ at 694 (equating the racial awareness advocated here with disparate treatment), Solon Barocas, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 724-726 (2016) (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"). This

racially disparate impact is specifically prohibited by the Supreme Court's decision in *Ricci v. DeStefano*.[116] If that were the case, 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, and that of other scholars,[117] this is overreading of *Ricci*. To see why, consider the facts of *Ricci*.

The Fire Department of the City of New Haven had developed a test to use in determining who would be promoted. Fire fighters studied for this test, purchased review material, and otherwise invested considerable time, energy and money in preparing for the test.[118] When the results were revealed, the numbers of minority candidates eligible for promotion was extremely small. As a result, the city decided not to certify the results, and so the firefighters who had passed the test were not be eligible for promotion.[119] The city defended its decision on the ground that the disparate impact prong of Title VII of the Civil Rights Act of 1964, as amended, prohibited it from using a screening mechanism that produced a disparate impact without sufficient reason.[120] 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 of the firefighters who had passed the test.[121] 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.[122]

Kroll and others, include Solon Borocas, read *Ricci* as prohibiting the intent to avoid a racially disparate impact[123] and the very awareness of race that differential tracking within algorithms would commend. As Pauline

---

interpretation over-reads *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.

[116] 557 U.S. 55 (2009).

[117] Other scholars agree, most notably Pauline Kim. *See e.g.* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. (2017) (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. 867 (DATE).

[118] Cite to *Ricci*.

[119] Cite to *Ricci*.

[120] Cite to *Ricci*.

[121] Cite to *Ricci*.

[122] Cite to *Ricci*.

[123] Kroll, *supra* note 93 at 694 (arguing that "[i]f 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").

Kim persuasively argues,[124] these scholars misread *Ricci*. 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. 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 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 and Barocas on the one hand and Kim other 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 the same Justice that authored the opinion for the Court in *Ricci* specifically endorses such approaches, like choosing to site schools where they will enroll a racially diverse cohort of students,[125] 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 II has explored how the costs of privileging either predictive parity or error rate balance can be mitigated. It begins by considering how to minimize the costs of choosing predictive parity over error rate balance and concludes that by changing the consequences of being labeled high risk, the costs of false positives will be lowered which will, in turn, make the unfairness of differential error rates less significant. This Part then goes on discuss how the costs of privileging error rate balance over predictive parity might be lowered. As this is the option recommended by Part I, most of the discussion focuses here. This Part

---

[124] Cite Kim

[125] *See* Parents Involved in Community Schools v. Seattle School District, cite, (parenthetical). However, as Justice Kennedy is no longer on the Supreme Court, his own views about these issues are less important going forward.

argues that consideration of race within algorithms will lessen the cost of the information loss produced by forgoing predictive parity. In addition, this Part argues against the consensus view that consideration of race within algorithms is always impermissible. Instead, it presents a picture of constitutional equal protection jurisprudence that would render this issue an open question.

CONCLUSION