The Ohio State Technology Law Journal

ALGORITHMS AND FAIRNESS

VIRGINIA FOGGO,[1] JOHN VILLASENOR,[2] & PRATYUSH GARG[3]

Algorithms, including those used in artificial intelligence ("AI"), are experiencing rapid adoption in a growing array of applications, including policing, criminal justice, employment, financial services, education, healthcare, and many others. While algorithms offer many potential benefits, they also pose the risk of propagating, amplifying, or generating bias.

This Article addresses a set of critically important questions that legal scholars, policymakers, technologists, companies, civil society groups, and consumers will be facing with increasing frequency in the coming years: How should algorithmic bias—or conversely, fairness—be measured? To what extent are different measures of fairness mutually exclusive or compatible? And how do measures for algorithmic fairness relate to—and stand to inform or be informed by—anti-discrimination law?

This Article presents a set of new results that have not been addressed before in legal academic publications regarding different approaches to measuring algorithmic fairness and the nature and extent of mutual incompatibilities among those measures. It also examines algorithmic

[1] PhD Student, UC Berkeley Department of Philosophy.

[2] Professor of engineering, law, and public policy, UCLA; director of the UCLA Institute for Technology, Law, and Policy; non-resident senior fellow, the Brookings Institution. The authors thank Jane Bambauer, Katie Eyer, Stacy Hawkins, Aziz Huq, Dan Linna, Sandra Mayson, Richard Re, Andrew Selbst, and Rebecca Wexler for their valuable feedback on this Article. The work leading to this Article was supported by a grant from the Micron Foundation.

[3] Research Assistant, UCLA Samueli School of Engineering.

fairness in the context of the antidiscrimination law, providing analysis on how statutes and case law relating to disparate treatment and disparate impact will impact algorithm design, and how in turn those legal frameworks can be informed by perspectives gained through experience with algorithms. It concludes with a set of recommendations and observations aimed at providing both (1) practical guidance on how to choose among multiple options for measuring fairness and (2) a foundation for broader discussions about updating approaches to algorithm design and conceptions of anti-discrimination law in ways that can promote algorithmic equity.

CONTENTS

I. Introduction

Algorithms should be fair. But what, exactly, does that mean? This question is profoundly important given the increasing incorporation of algorithms into healthcare, policing, criminal justice, marketing, education, finance, and nearly every other sector of society. Answering it has proven elusive due not only to the many different possible definitions of algorithmic fairness but also because of a shortage of sufficient information regarding the relationships among the definitions and the degree to which they are compatible or mutually exclusive. An additional complicating factor arises from inconsistencies in terminology. Depending on the source, multiple *different* terms are sometimes used to describe the *same* underlying method of evaluating fairness. An example of this is the method referred to variously as "test-fairness" and "calibration."[4] To add another wrinkle, sometimes the *same* term is used to describe *different* underlying methods of evaluating fairness. For instance, the metric known as "predictive parity" is applied in multiple different ways in the academic literature, with the result that whether or not an algorithm achieves predictive parity can depend on which version of the definition is used.[5]

The issue of fairness measures[6] is further complicated by important questions regarding the broader context of discrimination law frameworks and related efforts to identify and mitigate patterns of discrimination. The trajectory of discrimination law in recent decades reflects the intersection of an evolving statutory framework with a growing body of case law. Unsurprisingly, this evolution has occurred with a strong focus on the legal questions that arise in relation to identifying and combating discrimination—and less attention to the multiplicity of mathematical approaches to measuring it that, when properly contextualized, can provide a valuable complement to a purely legal analysis. As algorithms for making predictions and decisions become more widely deployed and are rightly subject to scrutiny regarding whether and to what extent they discriminate, it will be important to evaluate fairness with a perspective informed by an understanding of both the legal and technical issues involved.

[4] For example, what Verma and Rubin and (separately) Chouldechova refer to as "test-fairness," Corbett-Davies and Goel refer to as "calibration." *See* Sahil Verma & Julia Rubin, *Fairness Definitions Explained*, *in* PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS 1, 5 (2018), available at https://dl.acm.org/doi/pdf/10.1145/3194770.3194776?casa_token=9Rdkcqf4D8QAAAAA:ocMfzWRvoDXu07FuavioYbsD2ZWFL5fNzEZrvY7JhRiVeRsFynnPBNBAPphlDvD5DJe_7i4QPJs; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 52 BIG DATA 153, 156 (2017); Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning 6 (Aug. 14, 2018) (unpublished manuscript), https://arxiv.org/pdf/1808.00023.pdf. The use of multiple different terms to the same underlying fairness measure can, for obvious reasons, impede dialog among people who are not aware of the terminology duplication.

[5] Mayson states that "overall predictive parity" occurs when an algorithm achieves parity across groups in both positive predictive value and negative predictive value. *See* Sandra Mayson, *Bias in, Bias out*, 128 YALE L.J. 2218, 2243 (2019). By contrast, Verma and Rubin use a definition of "predictive parity" requiring only that positive predictive value be equal across groups, with no constraint on equality of negative predictive value." Verma & Rubin, *supra* note 4, at 3.

[6] We use the phrases "fairness *measures*" and "fairness *metrics*" interchangeably.

These issues are not merely theoretical: Companies, governments, courts, civil society groups, algorithm designers, consumers, academic researchers, and others involved in identifying and mitigating algorithmic bias[7] all have an interest in constructive legal and policy outcomes. Making effective progress in promoting algorithmic fairness will require a clear understanding of how to define, measure, and discuss it, as well as an understanding of how the various options for measuring fairness differ, the tradeoffs involved when choosing among them, and how decisions regarding which fairness measures to use should be influenced by—or should influence—discrimination law.

This Article aims to help facilitate these objectives in several ways. First, it provides a broad treatment of algorithmic fairness that both incorporates and extends the results of previous technical and legal scholarship. In doing so, it uses a series of examples to explain various fairness metrics and presents a set of new results regarding the relationships *among* fairness measures. While previous legal academic publications have discussed the fact that fairness measures can be mutually incompatible (in the sense that satisfying one can make it mathematically impossible to satisfy another), this Article explains that the landscape is in fact more nuanced. In some cases, under constraints that we discuss and that have important policy implications, it *is* possible to achieve fairness under more than one fairness measure.

Second, the article is written with the goal of being accessible and useful to a broad range of readers in the law and policy communities, including those who may not have an extensive background in mathematics. This is important because much of the literature on technical solutions for algorithmic fairness has been published in journals in technical fields

[7] As used herein, the term "bias" is generally intended to refer to bias in relation to legally and/or (in places where no formal legal protection yet exists) ethically problematic consideration of attributes such as race, gender, sexual orientation, religion, gender identity, etc. in association with decisions regarding hiring, financial services, housing, etc. Of course, there are also contexts in which "bias" has an innocuous meaning—such as, for example, if coaches holding tryouts for a soccer team make team roster decisions based on "bias" in favor of skilled soccer players. "Bias" herein is also not intended to refer to bias as that term is sometimes used in a highly technical sense in machine learning; e.g., in the "bias-variance tradeoff."

and is written for audiences of those journals. While there are also a growing number of law review publications on this topic, they often focus on a smaller subset of fairness measures than we consider here. There is thus a need for an article that discusses a broader array of fairness measures than has been addressed in much of the previous legal academic scholarship on this topic.

A third aim of this Article is to provide information that can be useful in the inevitable and necessary dialog on how best to apply and potentially update the legal and policy frameworks that will be used for assessing and addressing algorithmic bias. While much of the growth in attention to algorithmic fairness and algorithmic bias is recent, the more general challenge of bias has been the subject of decades of research and case law. This has created an important legal foundation reflected in statutes and case law relating to disparate treatment and disparate impact. As important as these frameworks are, in future years it will be important to improve and extend them to address the risks and opportunities created by algorithms. There is thus a need for a broad, technically-informed and legally-informed view of the algorithmic fairness landscape that is useful and accessible to scholars and practitioners in law, policy, and beyond.

A. Measuring Fairness: Historical Context

While recent years have seen rapid growth in the number of publications in both the legal and technical academic press addressing algorithmic fairness, interest in the broader issue of fairness measures dates back decades. As Hutchinson and Mitchell observe in a 2019 article titled *50 Years of Test (Un)fairness: Lessons for Machine Learning*:

> [t]he period of time from 1966 to 1976 in particular gave rise to fairness research with striking parallels to ML [machine learning] fairness research from 2011 until today, including formal notions of fairness based on population subgroups, the realization that some fairness criteria are incompatible with one another, and pushback

> on quantitative definitions of fairness due to their limitations.[8]

During the second half of the twentieth century, researchers devoted significant attention to identifying and addressing bias in testing in areas such as education and employment.[9]

Then, as now, researchers recognized the challenges inherent in defining bias and in determining how broader ethical, legal, and social considerations should inform efforts to address it. In a paper published in 1971 in the *Journal of Educational Measurement*, Thorndike described an "increase in concern about 'culture-fairness' of tests and testing procedures," observing that "the problems that are involved are partly problems of empirical fact, but partly problems of definition."[10] In a 1976 article in *Psychological Bulletin*, Hunter and Schmidt "describe[d] three distinct ethical positions" as well as "five statistical definitions of test fairness," and "show[ed] how each is based on one of these ethical positions."[11] They also considered the "technical, social, and legal advantages and disadvantages of the various ethical positions and statistical definitions,"[12] concluding that "any purely statistical approach to the problem of test bias is doomed to rather immediate failure."[13] Scheuneman wrote in 1979 that "[i]n the past few years, the

[8] Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)fairness: Lessons for Machine Learning*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 49, 49 (2019), available at https://dl.acm.org/doi/pdf/10.1145/3287560.3287600?casa_token=MzAzCR3bbssAAAAA:UFrz8jSFbdBQ3agjnRDDn8dF4aYUOhH-KTVSXwD-BzBZK6UPWYJk1Td5AFmm2QhAS4VMikcx9pU.

[9] *See, e.g.*, Hillel J. Einhorn & Alan R. Bass, *Methodological Considerations Relevant to Discrimination in Employment Testing*, 75 PSYCHOL. BULL. 261 (1971); RONALD L. FLAUGHER, BIAS IN TESTING: A REVIEW AND DISCUSSION*,* TM REPORT NO. 36 (1974); Robert M. Guion, *Employment Tests and Discriminatory Hiring*, 5 INDUS. REL. 20 (1966).

[10] Robert L. Thorndike, *Concepts of Culture-Fairness*, 8 J. EDUC. MEASUREMENT 63, 63 (1971).

[11] John E. Hunter & Frank L. Schmidt, *Critical Analysis of the Statistical and Ethical Implications of Various Definitions of Test Bias*, 83 PSYCH. BULL. 1053, 1053 (1976).

[12] *Id.*

[13] *Id.* at 1069.

issue of test bias with its far-reaching political and social implications has been the subject of much controversy."[14]

In the context of employment, Congress took action to address bias in Title VII of the Civil Rights Act of 1964, which prohibited employers, employment agencies, and labor organizations from discrimination based on "race, color, religion, sex, or national origin."[15] With respect to testing, the Act also provided that

> nor shall it be an unlawful employment practice for an employer to give and act upon the results of any professionally developed ability test provided that such test, its administration or action upon the results is not designed, intended, or used to discriminate because of race, color, religion, sex, or national origin.[16]

The Supreme Court considered employment testing in *Griggs v. Duke Power Company* in 1971, concluding in relation to the use of racially exclusionary tests that "[n]othing in [Title VII of] the [Civil Rights] Act [of 1964] precludes the use of testing or measuring procedures; obviously they are useful. What Congress has forbidden is giving these devices and mechanisms controlling force unless they are demonstrably a reasonable measure of job performance."[17]

Professional organizations also undertook efforts to develop and promulgate testing standards, including the *Standards for Educational and Psychological Testing* developed jointly by the American Educational Research Association, the American Psychological

[14] Janice Scheuneman, *A Method of Assessing Bias in Test Items*, 16 J. EDUC. MEASUREMENT 143, 143 (1979).
[15] Civil Rights Act of 1964, Pub. L. No. 88-352, § 703, 78 Stat. 241, 255 (codified as amended at 42 U.S.C. § 2000e–2). Of course, the Civil Rights Act of 1964 also addressed discrimination in domains beyond employment, including public accommodations, voting rights, and public education.
[16] Civil Rights Act of 1964, Pub. L. No. 88-352, § 703(h), 78 Stat. 241, 257 (codified as amended at 42 U.S.C. § 2000e–2(h)).
[17] Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).

Association, and the National Council on Measurement in Education.[18] These three organizations have been collaboratively publishing editions of the *Standards for Educational and Psychological Testing* since 1966, most recently in 2014.[19] The Title VII provision regarding testing remains an important topic of legal scholarship as well.[20]

B. Recent Attention to Algorithmic Fairness

Against the backdrop of this historical context, we can now turn to a discussion of the growth in the last decade or so of algorithmic approaches to make or inform all manner of decisions. For instance, algorithmic risk assessments are used in the criminal justice system in decisions regarding bail, sentencing, and parole.[21] Algorithms are also used to evaluate applications for jobs and loans, predict future healthcare expenses, identify potentially problematic content on social media sites, produce search results in search engines, perform facial recognition, and facilitate online commerce.[22] In all of these applications—and many more—there is recognition of the potential for algorithmic bias and a desire to avoid or at least mitigate it. This has led to a growing set of recent publications on algorithmic fairness and, with

[18] *See generally* AM. EDUC. RSCH. ASS'N, APA & NAT'L COUNCIL ON MEASUREMENT IN EDUC., STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING (2014). The webpage for the *Standards for Educational and Psychological Testing* notes that "[t]he Testing Standards are a product of the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education" and that editions of these standards have been "[p]ublished collaboratively by the three organizations since 1966." *See The Standard for Educational and Psychological Testing*, APA, https://www.apa.org/science/programs/testing/standards [https://perma.cc/2DYF-BFRS].

[19] *Id.*

[20] *See, e.g.*, Kimberly West-Faulcon, *Fairness Feuds: Competing Conceptions of Title VII Discriminatory Testing*, 46 WAKE FOREST L. REV. 1035 (2011).

[21] *See, e.g.,* Jason Tashea, *Rick-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, ABA (Mar. 1, 2017, 1:30 AM), https://www.abajournal.com/magazine/article/algorithm_bail_sentencing_parole [https://perma.cc/72VN-GJGL].

[22] *See, e.g.,* Evanthia Faliagka et al., *Application of Machine Learning Algorithms to an Online Recruitment System*, *in* PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON INTERNET & WEB APPLICATIONS & SERVICES 215 (2012); Public Affairs, UC Berkley, *Mortgage Algorithms Perpetuate Racial Bias in Lending, Study Finds*, BERKLEY NEWS (Nov. 13, 2018), https://news.berkeley.edu/story_jump/mortgage-algorithms-perpetuate-racial-bias-in-lending-study-finds/ [https://perma.cc/6G3T-36KF].

it, a growing recognition of the complexities involved.[23] For example, in *Bias in, Bias out*, Mayson focused primarily on racial inequities in prediction but also addressed algorithmic fairness more broadly.[24] In *Measuring Algorithmic Fairness*, Hellman advocated using a metric termed "error ratio parity," which occurs when the ratio of the false positive rate to the false negative rate is the same across groups.[25] Huq suggested using technical solutions that have "the distinctive feature of aligning racial equity with social efficiency."[26] Among these is a suggestion to give less priority to false positive rates.[27]

In an echo of Hunter and Schmidt's 1976 warning against a "purely statistical approach,"[28] modern scholars have voiced the importance of not attempting to address algorithmic fairness in isolation.[29] Solow-Niederman et al. have written that "looking at issues such as fairness or bias in a tool in isolation elides vital bigger-picture considerations about the institutions and political systems within which tools are developed and deployed."[30] Ajunwa has written that "the current framing of

[23] *See, e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249 (2008); Brandon L. Garrett & John Monahan, *Judging Risk*, 108 Calif. L. Rev. 439 (2020); Leah Wisser, *Pandora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 Am. Crim. L. R. 1811 (2019); Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, Sci. Advances, Jan. 17, 2018, at 1; Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, *in* Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 797 (2017), available at https://dl.acm.org/doi/pdf/10.1145/3097983.3098095?casa_token=PDW79AE8l8QAAAAA:q7HHlV7s5WWhbRsXzf-mzKPfUsBl6tp4jFuTnweQbzkHXRnBXgyjZ8aOEA9T66LObF3kM0CKPEM; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671 (2016); Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 Hastings L.J. 1389 (2019); Mark MacCarthy, *Standards of Fairness for Disparate Impact: Assessment of Big Data Algorithms*, 48 Cumb. L. Rev. 102 (2017); Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. Pa. L. Rev. Online 189 (2017).

[24] *See* Mayson, *supra* note 5, at 2222-23.

[25] Deborah Hellman, *Measuring Algorithmic Fairness*, 106 Va. L. Rev. 811, 835 (2020).

[26] Aziz Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L.J. 1043, 1134 (2019).

[27] *Id.* at 1126.

[28] Hunter & Schmidt, *supra* note 11, at 1053.

[29] Lee Rainie & Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age*, Pew Rsch. Ctr. (Feb. 8, 2017), http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age [https://perma.cc/2HRU-5ETY].

[30] Alicia Solow-Niederman et al., *The Institutional Life of Algorithmic Risk Assessment*, 34 Berkeley Tech. L.J. 705, 708 (2019).

algorithmic bias as a technical problem rather than as a legal problem is misguided."[31]

An example of the complexity of the questions that can arise when attempting to measure algorithmic bias is illustrated by the debate that followed ProPublica's May 2016 publication regarding Northpointe's COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) criminal risk assessment software.[32] The article, titled *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, asserted that "Prediction Fails Differently for Black Defendants," stating that:

> Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.[33]

This led to significant coverage in the broader press.[34]

[31] Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1617, 1707 (2020).

[32] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/9G6L-MYML].

[33] *Id.*

[34] *See, e.g., The Hidden Discrimination in Criminal Risk-Assessment Scores*, NPR (May 24, 2016, 4:32 PM), https://www.npr.org/2016/05/24/479349654/the-hidden-discrimination-in-criminal-risk-assessment-scores [https://perma.cc/TUS8-VUSK]; John Naughton, *Even Algorithms Are Biased Against Black Men*, GUARDIAN (June 26, 2016, 4:00 AM), https://www.theguardian.com/commentisfree/2016/jun/26/algorithms-racial-bias-offenders-florida [https://perma.cc/KE4R-A3AA]; Rachael Revesz, *Criminal Justice Software Algorithm Used Across the U.S. Is Biased Against Black Inmates, Study Finds*, INDEPENDENT (June 27, 2016, 5:45 PM), https://www.independent.co.uk/news/world/americas/northpointe-algorithm-propublica-biased-black-white-defendants-reoffend-a7106276.html [https://perma.cc/A44R-N4E6].

It also led to a response from Northpointe, which in July 2016 published a paper arguing that:

> ProPublica focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites. Their use of these statistics resulted in false assertions in their article that were repeated subsequently in interviews and in articles in the national media.[35]

The Northpointe authors also asserted that "[t]he results demonstrate *predictive parity* for blacks and whites."[36] In an article published later in 2016, Flores et al. wrote that "COMPAS does not predict outcome differently across groups of Black and White defendants—a given COMPAS score translates into roughly the same likelihood of recidivism, regardless of race."[37] Flores et al. also noted "the existence of standards for educational and psychological testing put forth by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (2014)"[38] and asserted that ProPublica "failed to test for bias within these existing standards."[39]

In a subsequent paper considering both ProPublica's assertions as well as those of its critics, Chouldechova showed "that the differences in false positive and false negative rates cited as evidence of racial bias by [ProPublica] are a direct consequence of applying an RPI [recidivism prediction instrument] that that [sic] satisfies predictive parity to a population in which recidivism prevalence differs across groups."[40] Chouldechova also noted that "fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one.

[35] WILLIAM DIETERICH ET AL., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 1 (2016).
[36] *Id.* at 2 (emphasis in original).
[37] Anthony W. Flores et al., *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*," FED. PROBATION, Sept. 2016, at 38, 44.
[38] *Id.*
[39] *Id.*
[40] Chouldechova, *supra* note 4, at 2.

A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used."[41] Other authors have also noted the impact of choosing different fairness measures. In an article containing a detailed analysis of the assessments of COMPAS provided by ProPublica and its critics, Hamilton noted that "when base rates between groups differ, the algorithm cannot achieve equal false positive rates and equal positive predictive values at the same time because only the latter statistic is heavily influenced by base rate differentials."[42]

As the above examples illustrate, algorithmic fairness is a complicated and potentially controversial topic. The existence of different ways to measure fairness and different conceptions of what it means to be "fair" inevitably means that there is no perfect way to assess algorithmic bias. However, the dialog about algorithmic bias will be more effective—and more likely to lead to positive outcomes—if it is informed by greater awareness regarding the tools available for measuring bias and the relationships of those tools to each other and to existing antidiscrimination law.

The remainder of this Article is organized as follows. Part II provides background on algorithmic predictions that will facilitate the subsequent discussion of fairness metrics. It then presents a set of algorithmic fairness measures, providing definitions, examples, and discussions of the extent to which they are mutually compatible. Part III provides a review of discrimination law with particular attention to the ways in which courts have addressed allegations of bias in relation to civil rights statutes. Part IV examines approaches to promoting algorithmic fairness, including the potential tensions that can arise in light of discrimination law frameworks. It also argues that algorithm developers should provide transparency regarding which fairness measure(s) they are using and offers guidance on factors that can influence metric selection. Conclusions are presented in Part V.

[41] *Id.*

[42] Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC DAVIS L. REV. ONLINE 261, 269 (2019).

Before proceeding to the remainder of the Article, a few caveats are in order. First, while we discuss some of the most commonly cited fairness metrics here, we do not claim that we cover *all possible* fairness metrics. Indeed, we do not address "fairness through awareness,"[43] the "threshold test,"[44] "treatment equality,"[45] or metrics based on causal reasoning.[46] Second, because we focus primarily on binary decisions, we do not address the full set of issues that arise when arriving at those binary decisions; for example, those based on applying a thresholding function to a score.[47] Third, we underscore that fairness is not and should not be a purely mathematical endeavor, and that as important and useful as mathematical measures of fairness might be, they are applied in a broader context that admits different and sometimes incompatible views on how to define fairness. And, as we discuss, there are multiple incompatible metrics for assessing fairness. Thus, the overarching goal of this Article is not to identify any single optimal solution to algorithmic fairness, for such a solution does not exist. Rather, it is to promote a fuller understanding of some of the tools available to measure fairness, with the recognition that application of those tools will be highly context dependent.

[43] *See* Cynthia Dwork et al., *Fairness Through Awareness*, *in* PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214, 215 (2012), available at https://dl.acm.org/doi/pdf/10.1145/2090236.2090255?casa_token=IKU5yfVTeb8AAAAA:b3oWWJkmblTwNGKMz3GKHRh254YEqzBdVpMfolp13tzW6pDT4Xpeafc7RyaEuXLQjYPEdZThcr8 ("We capture fairness by the principle that any two individuals who are similar *with respect to a particular task* should be classified similarly. In order to accomplish this *individual-based* fairness, we assume a distance metric that defines the similarity between the individuals").

[44] Camelia Simoiu et al., *The Problem of Infra-Marginality in Outcome Tests for Discrimination*, 11 ANNALS APPLIED STAT. 1193, 1193 (2017).

[45] Treatment equality considers errors, requiring that the ratio of false negatives to false positives be equal across groups. *See* Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RSCH., July 2018, at 1, 14.

[46] *See, e.g.*, Verma & Rubin, *supra* note 4, at 6.

[47] *See, e.g.*, Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, *in* PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 3323 (Daniel D. Lee & Ulrike von Luxburg eds. 2016) (discussing receiver operating characteristics ("ROC") and their relation to thresholding).

II. Measuring Fairness

We consider fairness in the context of algorithms that make predictions.[48] For simplicity, in much of the discussion that follows we will assume that these predictions are binary,[49] a framing that applies to a long list of applications. For example, a financial institution may wish to predict whether or not a loan applicant would repay a loan. A court may wish to predict whether or not a parolee will get rearrested within a certain time frame. A healthcare provider may wish to predict whether a patient will develop a particular medical condition within the next five years. A company doing a job search may wish to predict whether applicants would be able to perform effectively if hired. A credit card company needs to predict whether a requested transaction is legitimate or fraudulent. In scenarios like these and many more, algorithms can be used to generate a binary prediction.[50]

For each of these scenarios, there is also the separate question of outcomes (as opposed to predictions), which again in the interest of simplicity we will also treat as binary. If a person is given a loan, he or she will either repay it or default. A parolee will either get rearrested or not. A patient will either develop a particular medical condition or not.

[48] While in this Article we focus the discussion on algorithmic fairness in the context of predictions, we note that algorithmic fairness issues also arise in relation to statistical inferences.

[49] We focus herein on binary predictions, though we note that there are also many algorithms that produce non-binary outputs. For instance, an algorithm that predicts how likely a consumer is to purchase a product might produce a score in the range from 1 to 10, with 1 representing least likely and 10 representing most likely. Or it might provide a non-numerical output expressed by choosing one of multiple categories such as "very likely," "likely," "uncertain," "unlikely," and "very unlikely." Optionally, outputs such as these can be converted to binary form (e.g., by applying a threshold to a score, such that all scores above the threshold are grouped into one category and all scores below it are grouped into a second category), though whether doing so will be beneficial will depend on the context.

[50] For simplicity, in this portion of the discussion we are assuming that it is the algorithm itself that makes a binary prediction and associated decision. In some applications (e.g., when an algorithm used by a credit card company is deciding whether to flag an attempted transaction as fraudulent), this is what occurs. In other environments—such as criminal justice—an algorithmic output is not necessarily binary, and it is only one of multiple factors that are considered by a human who holds the actual power to make a decision on questions such as whether a person should be granted bail, granted parole, etc.

A job applicant will either be capable of performing the job or not. A credit card purchase will either be legitimate or fraudulent.

It is important to emphasize that making algorithmic predictions, observing outcomes, and evaluating the fairness of those predictions in light of the outcomes are different (though often related) tasks. Sometimes the same entity will do all three. Algorithm developers are increasingly recognizing the need to include fairness considerations as part of the design process.[51] As the developers are engaged in the work of formulating an algorithm to make a prediction (e.g., regarding whether a loan applicant will repay a loan), before placing the algorithm into commercial service they typically test it on historical data to evaluate the extent to which its decisions might inadvertently reflect bias.[52] This provides an opportunity to revise and improve the algorithm to address these issues before it is used for real-world applications.

There are also many scenarios in which people or groups other than the developers of a predictive algorithm will assess its fairness. In fact, as algorithm-based decisions become more widely deployed, after-the-fact fairness evaluation by groups independent of the design team will likely become the norm.[53] People involved in evaluating an algorithm generally will seek to get as much information about the predictions and outcomes as possible and then to use those data as input to evaluate

[51] *See* Jairo Mejía & Roberto Maestre, *Fairness by Design in Machine Learning Is Going Mainstream*, BBVA DATA (Aug. 10, 2018), https://www.bbvadata.com/fairness-by-design-in-machine-learning-is-going-mainstream/ [https://perma.cc/A8Z6-KP7Z]; *see also* Sally Ward-Foxton, *Reducing Bias in AI Models for Credit and Loan Decisions*, EE TIMES (Apr. 30, 2019), https://www.eetimes.com/reducing-bias-in-ai-models-for-credit-and-loan-decisions/# [https://perma.cc/D5T6-385B].

[52] *See* Kathryn Hume & Alex LaPlante, *Managing Bias and Risk at Every Step of the AI-Building Process*, HARV. BUS. REV. (Oct. 30, 2019), https://hbr.org/2019/10/managing-bias-and-risk-at-every-step-of-the-ai-building-process [https://perma.cc/X3P3-HYAQ] ("Train and test the model (or several potential model variants). Gauge the impact of fairness and privacy enhancements on accuracy.").

[53] In general, after-the-fact evaluations provide an important feedback mechanism to identify and address potential algorithmic bias. It can also serve as a way to help identify problems posed by "concept drift," which in the context of machine learning refers to changes over time in the underlying statistics of data of interest, which could cause a predictive model developed using the outdated statistics to be less accurate. For example, policy changes with respect to policing could have a significant impact on the accuracy of an algorithm used to predict crime, if that algorithm uses data collected before that policy change was implemented.

fairness according to one or more measures. Thus, predictions and outcomes are a vital input to algorithmic fairness measures. With that in mind, we will first present background on predictions and outcomes, as these concepts and the related terminology will form the foundation for the subsequent discussion of fairness metrics.

A. Background: Predictions and Outcomes

Before presenting measures of algorithmic fairness, it is important to lay the groundwork by noting some metrics that are commonly used for measuring the relationship between binary predictions and binary outcomes, and that are commonly used as inputs to algorithmic fairness calculations. To facilitate this explanation, we will initially consider a scenario involving predictions regarding whether or not students will pass a particular test.

Every student can be associated with both a prediction made before the test is administered and an outcome observed after the test is given. The presence of both a binary prediction and a binary outcome leads to four possibilities. A *true positive* refers to cases in which the algorithm correctly predicts that a student will pass the test. A *true negative* refers to cases in which the algorithm correctly predicts that a student will fail the test. A *false positive* (also known as a Type I error in statistics) refers to cases in which the algorithm predicts that a student will pass the test, while in fact the student fails it. A *false negative* (also known as a Type II error in statistics) refers to cases in which the algorithm predicts that a student will not pass the test, while in fact the student passes it.

In the context of this scenario, true positives, true negatives, false positives, and false negatives are evaluated at the level of individual. Across a group, it is also possible to compute probabilities associated with each of these measures. The *true positive rate* is the probability that students who pass the test are correctly predicted by the algorithm to do so. The *false negative rate* is the probability that students who pass the test are incorrectly predicted not to do so.[54] The *true negative rate* is the probability that students who fail the test are correctly predicted

[54] The sum of the true positive rate and false negative rate will always be 1.

by the algorithm to fail. The *false positive rate* is the probability that students who fail the test are incorrectly predicted to pass it.[55]

Another metric is the *positive predictive value* of the prediction algorithm, which is the probability that students who are predicted to pass the test do in fact end up passing it. Analogously, the *negative predictive value* of the algorithm is the probability that students who are predicted to fail the test do in fact end up failing it.

These measures are best illustrated with the aid of a specific example. We consider two groups of 100 students, denoted as the orange group and blue group, respectively. Prior to the test date, the algorithm produces a binary prediction regarding whether each student in each group will pass the test. Then the test is administered, and it turns out that 70 students in the orange group and 52 students in the blue group pass the test. This means that the *base rates* for the orange group and blue group are 0.7 and 0.52, respectively.[56] As will be discussed in more detail below, the presence of different base rates across groups has important consequences in relation to measures of algorithmic fairness. The two tables below provide an example of how the various metrics described above are computed.[57]

[55] The sum of the false positive rate and true negative rate will always be 1.

[56] This example, and the other examples herein, are provided to help illustrate the meaning of various metrics and prediction approaches. In the interest of simplicity and clarity, we are not addressing statistical significance, which would provide guidance on the accuracy to which numerical results computed based on the relatively small sample sizes discussed in the examples are more broadly representative. For simplicity in the discussion herein, we assume that the observed probabilities are statistically meaningful, though in practice larger sample sizes would be required for that assumption to hold true.

[57] Each of these tables is constructed by building off of four key numbers that in the aggregate are often presented as a confusion matrix, which is a 2x2 matrix in which the rows represent the predictions and the columns represent the outcomes. In Table 1, there are 60 true positives, 20 false positives, 10 false negatives, and 10 true negatives. All of the other numbers shown in and immediately below Table 1 can be derived from those four numbers.

	Total	**Pass**	**Fail**	
All	100	70	30	Base rate = 0.7
Predicted to pass	80	60	20	Positive predictive value = 60/80 = 0.75
Predicted to fail	20	10	10	Negative predictive value = 10/20 = 0.5

Table 1: Orange Group (100 members total)

Orange group true positive rate = 60/70 = 0.86
Orange group true negative rate = 10/30 = 0.33
Orange group false positive rate = 20/30 = 0.66
Orange group false negative rate = 10/70 = 0.14

	Total	**Pass**	**Fail**	
All	100	52	48	Base rate = 0.52
Predicted to pass	56	42	14	Positive predictive value = 42/56 = 0.75
Predicted to fail	44	10	34	Negative predictive value = 34/44 = 0.77

Table 2: Blue Group (100 members total)

Blue group true positive rate = 42/52 = 0.81
Blue group true negative rate = 34/48 = 0.71
Blue group false positive rate = 14/48 = 0.29
Blue group false negative rate = 10/52 = 0.19

B. Fairness Measures: Framing the Issue

Given two groups, what does it mean for an algorithm to make fair decisions?[58] One possible answer is that a fair algorithm will make a positive prediction at equal rates for both groups. If the two groups are loan applicants and the prediction concerns loan repayment, this means that, on average, an equal fraction of members from each group will be deemed likely to repay a loan (and will presumably thus be granted a loan). But suppose that, due to historical patterns of discrimination, the members of one group have lower average incomes, higher debts, and

[58] For other papers that discuss fairness measures, *see generally* Corbett-Davies & Goel, *supra* note 4; Mayson, *supra* note 5; Melissa Hamilton, *The Sexist Algorithm*, 38 BEHAV. SCI. & L. 145 (2019).

more expenses and therefore less repayment capacity than members of a second group—meaning that the base rate for successful repayment would differ across the groups.

Would it still be "fair" to grant loans to an equal percentage of loan applicants from each group? It could be argued that doing so is fair, as it helps to mitigate the historical impact of discriminatory social structures that have contributed to the variation in repayment capacity across groups. Or would fairness require granting loans at a higher rate to members of the group with the higher repayment capacity, with the result that the percentage of applicants receiving loans would differ across groups? It could be argued that making loan decisions purely based on individual repayment capacity is fair in the sense of treating similar individuals similarly, regardless of group membership, though it would fail to mitigate the negative effects of such loan decisions on the disadvantaged group.

This example illustrates that concepts of fairness can be in tension. Tensions of this sort are sometimes described in terms of "individual fairness" and "group fairness."[59] However, such designations can tend to oversimplify, as there can be hidden complexities and assumptions in the choice of how to define "individual fairness" and "group fairness."[60] An additional concern is that prediction algorithms will often not operate in ways that cleanly comport with (or fail to comport with) either label, however they may be defined.

We believe it is preferable to evaluate and discuss fairness with the aid of clearly defined mathematical metrics. Of course, we are not suggesting that the discussion starts and ends with mathematics: As noted earlier, fairness is a concept that also involves policy, law, and social norms, all of which should be considered in addition to mathematics, though not in a manner that evicts mathematics entirely from the discussion. And a more holistic treatment of fairness that

[59] Reuben Binns, On the Apparent Conflict Between Individual and Group Fairness (Dec. 14, 2019) (unpublished manuscript), https://arxiv.org/pdf/1912.06883.pdf.

[60] Mayson addressed this issue as well, writing that "[m]uch recent work in algorithmic fairness has categorized measures of equality as either 'group fairness' or 'individual fairness' metrics. This dichotomy, however, can be misleading." Mayson, *supra* note 5, at 2239 n. 65.

involves all such considerations will be more productive if it is informed by a common vocabulary regarding the various possible mathematical fairness measures, even if there remains disagreement regarding their relative or absolute utility in a particular context.

In general, there are mathematical incompatibilities among different fairness measures. Thus, at the heart of discussions surrounding fairness are questions regarding whether or not the various statistical measures of fairness can be simultaneously satisfied. And, if they cannot, which ought to be prioritized in any given context?

In the subsequent discussion, we will assume that there are two groups of people characterized by different base rates with respect to an outcome of interest (e.g., passing a test, repaying a loan, being re-arrested after being released from prison, etc.). We then consider how information on binary predictions and binary outcomes is used in the following fairness measures: equality of opportunity, equalized odds, predictive parity, and statistical parity. These are briefly explained in the following table, with more detailed explanations to follow.

Metric name	**Definition**
Equality of opportunity	True positive rate is equal for both groups
Equalized odds	True positive rate is equal for both groups *and* the false positive rate is equal for both groups
Predictive parity	Positive predictive parity is equal for both groups
Statistical parity	The probability of making a positive (or negative) decision is equal for both groups

Table 3: Fairness Measures

While this is far from complete list of all possible fairness measures, it is diverse enough to effectively illustrate some of the complexities that can arise when measuring fairness yet compact enough to keep the discussion straightforward and tractable. A key point that we will return to repeatedly is that when the base rates across groups differs, and we lack perfect prediction, it is typically possible to simultaneously satisfy more than one metric only under highly constrained conditions, if at all.

1. “Equality of Opportunity”

When the true positive rate is equal across the two groups, a prediction algorithm is said to satisfy equality of opportunity.[61] In the example in Tables 1 and 2, the true positive rates for the orange and blue groups differ: Of the 70 people in the orange group who did pass, 60 of them had been predicted to pass. This corresponds to a TPR (true positive rate) of 60/70, which is (approximately) equal to 0.86. For the blue group, of the 52 people who did pass, 42 of them had been predicted to pass, corresponding to a TPR of 42/52, which is (approximately) 0.81. Thus, because 0.86 and 0.81 differ, the predictor shown in those tables does not satisfy equality of opportunity.[62] Note that “equality of opportunity” refers to a scenario in which, for the people who actually do belong to the positive class—such as students who are going to pass a test, or people who would pay back a loan if given one—the likelihood of being identified by the algorithm as being in that class is the same for both groups.[63] It does not necessarily comport with “equal opportunity” as that term might be used in a broader normative or policy sense, and for this reason we have used quotation marks in the title of this subsection above.

Assuming in this case that a positive outcome results in favorable or desirable treatment, equality of opportunity (in the technical sense) comports with one possible (of multiple) moral understanding of equal opportunity in the sense that individuals who merit that outcome have the same probability (or *opportunity*) of actually obtaining it, regardless of group membership (equal true positive rate). By the same token, since the false negative rate is always equal to one minus the true positive rate, they also have the same probability of being *denied* this outcome—that is, they have the same probability of a missed opportunity (equal

[61] *See, e.g.*, Hardt et al., *supra* note 47, at 8.

[62] As noted, since 0.86 is not equal to 0.81, the predictor does not satisfy equality of opportunity. However, it could be argued that since 0.86 and 0.81 are not *that* different, the predictor comes close to satisfying equality of opportunity. When using quantitative measures like this, there are interesting questions that could be asked about what level of difference is necessary to constitute a “disparate” impact.

[63] *See* Ziyuan Zhong, *A Tutorial on Fairness in Machine Learning*, TOWARDS DATA SCI. (Oct. 21, 2018), https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb [https://perma.cc/JW77-V6SS].

false negative rate). However, there are also a number of reasons why this definition of fairness may *not* align with normative understandings of equality of opportunity. To give one example, equality of opportunity in the technical sense does not account for external social barriers that may be the cause of differing base rates for the observed behavior among groups. For instance, in using a predictive algorithm to decide who will be granted loans, it may be possible to achieve mathematical equality of opportunity despite the fact that many women, as a result of discriminatory social structures, have not had an equal opportunity to build credit. In this case, a prediction deemed "fair" under this metric could nonetheless be deemed unfair when viewed more holistically.

2. Equalized Odds

As noted above, when a prediction algorithm produces equal true positive rates across the two groups, it is said to satisfy equality of opportunity.[64] If, in addition, the predictive algorithm also has equal false positive rates across the two groups, it satisfies equalized odds.[65] Thus, equalized odds adds an additional condition over and above equality of opportunity. A predictive algorithm that satisfies equalized odds will of necessity satisfy equality of opportunity, though a predictor that satisfies equality of opportunity may—but does not necessarily—satisfy equalized odds.

More formally, in a predictor that satisfies equalized odds, two conditions must both hold. First, the TPR must be equal for the two groups, and second, the false positive rate ("FPR") must be equal for the two groups.[66] As discussed earlier in relation to the student example in Tables 1 and 2, the TPR for the orange and blue groups are 0.86 and 0.81, respectively. As a result, the predictor in Tables 1 and 2 does *not* satisfy equalized odds, since the TPR is not equal for the two groups. Note that even if the TPR *had* been equal across both groups, that alone would not have been sufficient to satisfy equalized odds. Equalized odds requires that the TPR be equal across both groups, *and* that the FPR be

[64] *Id.*
[65] *Id.*
[66] Hardt et al., *supra* note 47, at 3.

equal across the two groups. In this case, there is no need to evaluate whether both groups have the same FPR, since the lack of equality in TPR means that equalized odds is not satisfied.[67] While equalized odds requires equality among TPR as well as equality among FPR, it is also possible to construct a less stringent fairness metric that would require equality only among FPR (without placing any constraints on TPR). Such a metric is sometimes called "equal specificity."[68]

3. Predictive Parity

Predictive parity is satisfied when the positive predictive values are equal for both groups.[69] This means that, of the people who are *predicted* to be in the positive class, the percentage who are *actually* in the positive class is the same for both groups. Recall that the positive predictive value is the probability that a prediction of a positive outcome will turn out to be correct.[70] Consider again the prediction of test outcomes shown in Tables 1 and 2. For the orange group, 80 of the students were predicted to pass, and 60 of those 80 actually did pass. Thus, the positive predictive value for the orange group is 0.75. Analogously, for the blue group, of the 56 students who were predicted to pass, 42 of them actually did pass. This means that the positive predictive value for the blue group is 42/56 = 0.75. Since the positive predictive value for both the orange and the blue group is 0.75, this predictor satisfies predictive parity.

It is also possible to construct a related and more stringent definition of predictive parity called "overall predictive parity" in which not only must the positive predictive values be equal across both groups, but the negative predictive values must also be equal.[71] Under that more restrictive definition, the example in Tables 1 and 2 would not satisfy

[67] The FPR for the orange and blue groups are 0.66 and 0.29 respectively, so this lack of equality is alone sufficient to show that the predictor does not satisfy equalized odds.

[68] Mayson, *supra* note 5, at 2243.

[69] *See, e.g.*, Verma & Rubin, *supra* note 4, at 3.

[70] *Id.* at 2.

[71] Mayson, *supra* note 5, at 2243 ("If the algorithm's rearrest forecasts are correct at an equal rate for each group, the algorithm achieves parity in positive predictive value. If the no-rearrest forecasts are correct at an equal rate for each group, the algorithm achieves parity in negative predictive value. And if both are true, it achieves overall predictive parity.").

predictive parity because the negative predictive values of the orange and blue groups, 0.5 and 0.77 respectively, are not equal.

4. Statistical Parity

In contrast with equalized odds, equality of opportunity, and predictive parity, statistical parity is obtained solely based on predictions without any consideration of outcomes.[72] More specifically, statistical parity requires that the probability of making a positive (or equivalently, negative) decision is the same across both groups.[73] In Tables 1 and 2, this means that statistical parity would occur if an algorithm predicted that the same fraction of students in each group would pass the test. This clearly does not occur. For the orange group, 80 of the 100 members (80%) are predicted to pass while, for the blue group, only 56 of the 100 members (56%) are predicted to pass.

A. Relationships Among Fairness Measures

As previously noted, it is not generally possible to satisfy all possible definitions of fairness. It is therefore interesting to ask which fairness metrics *can* be simultaneously satisfied, and conversely, which metrics are mutually exclusive. Moreover, when it *is* possible to satisfy more than one fairness metric, is doing so advisable as a matter of policy? To answer the latter question, it will be important to consider mathematical constraints required when simultaneously satisfying multiple metrics, as well as the policy consequences of such constraints. This section illustrates and explores some of those relationships, including incompatibilities that can arise when aiming to achieve more than one metric at the same time. A detailed set of mathematical derivations is available in a related technical paper, with some of the key results summarized in the present Article.[74]

[72] *See id.* at 2242.

[73] *See, e.g.*, Verma & Rubin, *supra* note 4, at 3.

[74] *See generally* Pratyush Garg, John Villasenor & Virginia Foggo, *Fairness Metrics: A Comparative Analysis*, *in* PROCEEDINGS OF THE 2020 IEEE INTERNATIONAL CONFERENCE ON BIG DATA 3662 (2020).

Statistical parity and predictive parity: Putting aside the specific example in Tables 1 and 2, we now consider whether it is mathematically possible for a predictor to simultaneously satisfy both predictive parity and statistical parity when base rates differ.[75] Under certain circumstances, the answer is yes. This can be illustrated by way of an example. Consider two groups of ten people, which we will call the square group and the circle group. If given a loan, eight members of the square group would repay it, while two members of that group would default. By contrast, if all the members of the circle group were given a loan, eight members would default and two members would repay the loan. This is illustrated in Figure 1 below with the letters "R" (for "repay") and "D" (for "default") indicating what each member of each group would do if given a loan:

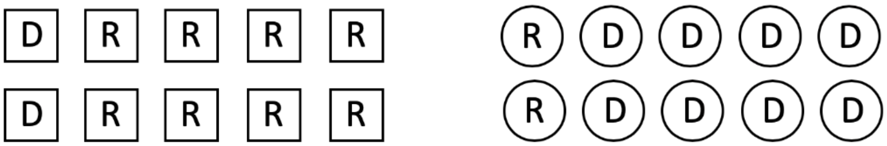

Figure 1: Two groups of ten members each. In the square group, 8 of the 10 members would repay ("R") a loan, and 2 members would default ("D"). In the circle group, the opposite is true.

As the "D" and "R" indications in Figure 1 make clear, the base rates differ strongly across the groups: 80% of the square group would repay a loan, while only 20% of the circle group would do so. Now consider an algorithm that identifies the individuals contained within the rounded rectangles shown in Figure 2 as being creditworthy, i.e., the algorithm predicts that they would repay a loan.[76] Based on this assessment, a loan is provided.

[75] This statement is true under the definition of predictive parity that we have used here: that the positive predictive value ("PPV") is equal across the two groups.

[76] Figure 2 is constructed to illustrate a prediction that is mathematically possible. Whether an algorithm would in fact (e.g., in relation to the square group) recommend granting a loan to two people who would end up defaulting while recommending denying a loan to six other people who would not end up defaulting is of course a separate matter.

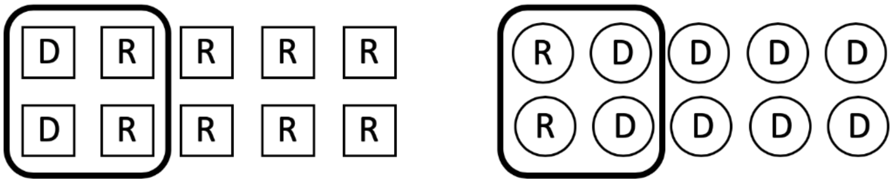

Figure 2: The groups from Figure 1, additionally annotated to show an example prediction of group members likely to repay a loan.

Note that the predictor in Figure 2 satisfies *statistical parity*, since for each group 40% of the members were given a loan. It also satisfies *predictive parity* since, for both groups, of the four people predicted to repay the loan, two of them actually did, which corresponds to a positive predictive value ("PPV") of 0.5 for both groups. It might be argued that this prediction is particularly "fair" because it satisfies not only one but two different fairness metrics.

But in other ways, the algorithm is clearly unfair. After all, for the circle group in Figure 2, all of the people who would repay the loan were given one (i.e., the true positive rate is 100%), while for the square group only two of the eight people who would repay the loan were provided with one (i.e., the true positive rate is 25%). In achieving statistical and predictive parity, the predictive algorithm has very different true positive rates, denying a loan to 75% (6 of the 8) of the members of the square group who would have repaid it. One way to bring the true positive rates into alignment would be to alter the predictive algorithm so that it provides loans to *all* the members of the square group as illustrated in Figure 3:

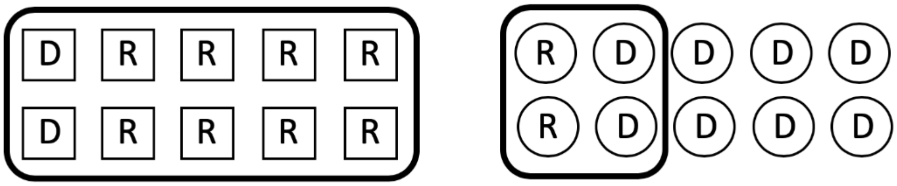

Figure 3: The groups from Figure 1, annotated to show an alternate prediction.

In Figure 3 the true positive rate is 100% for both groups, and the prediction thus satisfies equality of opportunity. But an algorithm equalizing the true positive rates in this manner would no longer satisfy statistical parity (because 100% of the square group would get loans, compared with only 40% of the circle group) or predictive parity (because the positive predictive value would be 0.8 for the square group and 0.5 for the circle group).

The preceding example offers a clear illustration of the policy challenges involved in attempting to simultaneously achieve statistical parity and predictive parity, despite the fact that it is mathematically possible to do. These sorts of trade-offs among various fairness measures are not limited to the particular pairwise combination of statistical parity and predictive parity—they will arise (though in different mathematical form) when satisfying fairness under other pairs of measures as well. However, there are also important mathematical considerations involved specifically in attempting to satisfy these two particular metrics.

Statistical parity and predictive parity can only simultaneously hold *when the ratio of the true positive rates of the two groups is the inverse of the ratio of the base rates of the two groups*.[77] As already discussed, Figure 2 shows a prediction satisfying both statistical parity (because 40% of each group is predicted to repay and thus is given a loan) and predictive parity (since within each group, 50% of those who are predicted to pay do in fact repay). For the square group the base rate is 0.8 (because of all the members of the group, 80% would repay) and for the circle group the base rate is 0.2 (because of all the members of the group, 20% would repay). This corresponds to a base rate ratio of 4:1. The true positive rate for the square group is 0.25 and the true positive rate for the circle group 1.0, corresponding to a true positive rate ratio of 1:4. Thus, the base rate ratio is the inverse of the true positive rate ratio.

This illustrates a general disadvantage of requiring both statistical parity and predictive parity when there are significant differences in base rates across the two groups: a large base rate ratio means that the true positive

[77] Garg, Villasenor & Foggo, *supra* note 74.

rate for one of the groups must be low.[78] And a low true positive rate is often problematic from a policy perspective.[79] In the example scenario presented above, the true positive rate of 25% for the group on the left meant that 75% of the members of that group who would have repaid a loan were not given one.

Equalized odds and predictive parity: When the base rates across the two groups are unequal, and when (as is nearly always the case) the prediction is imperfect,[80] *it is not possible to simultaneously satisfy both equalized odds and predictive parity*.[81] Those wishing to choose among the two will have to consider context-dependent costs and benefits involved in satisfying one over the other.

Equalized odds and statistical parity: As noted earlier, equalized odds occurs when both (1) the true positive rates are equal across the two groups, and (2) the false positive rates are equal across the two groups.[82] When the base rates are different across the two groups, *equalized odds and statistical parity can be satisfied only when the true positive rate*

[78] For example, if the base rate ratio is 5:1, this means that the TPR ratio will be 1:5, meaning that the lower TPR can be no higher than 20% (because the higher TPR is five times higher than the lower TPR, and cannot exceed 100%).

[79] In noting that a low true positive rate can be problematic, we are not suggesting that the true *negative* rate will be unimportant, or less important. For example, in predictions used to make parole decisions for non-violent offenders, if a "positive" outcome designates a future arrest for a non-violent crime, it may be more desirable to focus on ensuring a high true negative rate (i.e., correctly predicting who will not be arrested in the future) to avoid unnecessarily subjecting individuals to continued incarceration, regardless of the effect on the true positive rate. However, it is also important to keep in mind that a low true positive rate corresponds to a high false negative rate. In the case of predicting future arrests, a high false negative rate involves the risk of exposing people to criminal activity that could potentially have been prevented. A context-dependent analysis of the comparative value of true negatives and true positives will always be necessary.

[80] "Perfect prediction" describes a predictor that is always correct. In practice, this almost never occurs. In other words, when making binary predictions about binary outcomes, there are nearly always some false positives and/or some false negatives. In the case of a perfect predictor, equality of opportunity and equalized odds would be satisfied since the true positive rate for both groups would be 100% and the false positive rate for both groups would be zero. A perfect predictor would also satisfy predictive parity, since the positive predictive value would be 100% for both groups. However, a perfect predictor would not satisfy statistical parity if the two groups had different base rates.

[81] Garg, Villasenor & Foggo, *supra* note 74.

[82] *Id.*

equals the false positive rate.[83] While this is mathematically feasible, it is not particularly attractive from a policy standpoint, since we generally want the true positive rate to be higher than the false positive rate. In other words, with respect to positive predictions, we want the predictor to be correct more often than it is incorrect. This can be illustrated using the example in Figure 4, which involves two groups of twelve people who are seeking a loan, where "R" (repay) and "D" (default) again indicate what each person would do if given a loan, and the rectangles indicate the people in each group who were predicted to repay:[84]

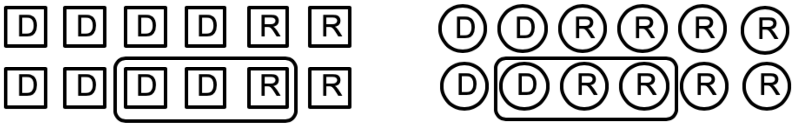

Figure 4: An example satisfying both equalized odds and statistical parity.

In Figure 4, the base rates are different across the groups: 1/3 of the members of the square group would repay a loan if given one, while 2/3 of the members of the circle group would do so. As the rectangles indicate, in each group three members (i.e., a fraction 0.25 of the members in each group) are predicted to repay, so statistical parity is satisfied. In addition, the true positive rate is equal to 0.25 for both groups (obtained as 1/4 for the square group and 2/8 for the circle group) and the false positive rate of 0.25 is equal for both groups (obtained as 2/8 for the square group and 1/4 for the circle group). This means that equalized odds is satisfied. But achieving both equalized odds *and* statistical parity requires equal true and false positive rates. Again, this is problematic since it is typically desirable to have a *high* true positive rate and *low* false positive rate, which is not possible when those rates are equal. In the example of Figure 4, the true positive rate for both groups is only 0.25, meaning that 75% of the people who would repay a loan would be incorrectly predicted to default.

[83] *Id.*

[84] As was the case with Figure 2, Figure 4 is constructed to illustrate a prediction that is mathematically possible. Whether an algorithm would in fact recommend granting a loan to the three members of each group indicated by the rectangles is of course a separate question.

While the foregoing discussion has not explored all possible pairwise combinations of fairness metrics, it illustrates the general nature of the challenges that can arise when attempting to simultaneously satisfy fairness according to multiple metrics. More generally, there are always three possible answers to the question of whether it is possible to simultaneously satisfy two fairness metrics in the presence of unequal base rates across two groups.[85] For some metric pairs (e.g., the combination of equalized odds and predictive parity), it is mathematically impossible to satisfy both metrics in the case of an imperfect predictor and different base rates.[86] For some metric pairs, it is mathematically possible to satisfy both metrics, but only at the cost of a generally unattractive policy outcome (e.g., it is possible to satisfy both equalized odds and statistical parity, but only if the true positive rate equals the false positive rate).[87] Finally, there are some metric pairs that can both be satisfied at the cost of constraints that are not necessarily problematic from a policy standpoint.[88] An example of this is the pair consisting of equality of opportunity and predictive parity, which can be simultaneously satisfied if—among other constraints—there is a difference in false positive rates.[89]

Thus, as a practical matter, when avoiding problematic constraints like requiring equal true and false positive rates, it will generally only be possible to satisfy fairness according to one metric (or, subject to highly specific constraints, two metrics), leaving the question of which metric(s) should be used.[90] Relatedly, choosing to design an algorithm to satisfy a particular fairness metric will often mean choosing *not* to achieve fairness under other metrics. Given the impossibility of achieving across-the-board fairness, what should be done? The answer

[85] *Id.*

[86] *Id.*

[87] *Id.*

[88] *Id.*

[89] *Id.* A difference in FPR alone is not sufficient for satisfying equality of opportunity and predictive parity. But if the FPRs are identical, then those two metrics cannot be simultaneously met.

[90] An exception under which it *is* in fact possible to simultaneously satisfy two metrics despite the presence of different base rates occurs when one metric implies the other. For example, equalized odds (which requires that true positive rate be equal across groups and that the false positive rate be equal across groups) also implies equality of opportunity (which requires only that true positive rate be equal across groups).

depends not only on mathematics but also on policy and on the law regarding discrimination. To provide a more complete framing, we review some key aspects of discrimination law and then use that as a framing to explore interactions between fairness measures and discrimination law.

III. Legal Frameworks Relating to Discrimination

Discrimination law is a complex and evolving field that has been shaped by the Constitution, statutes, case law, legal scholarship, politics, and trends in the broader public discourse. The past half century has seen the passage of multiple federal anti-discrimination statutes. For instance, Title VII of the Civil Rights Act of 1964 ("Title VII") provided that it is unlawful for an employer "to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin."[91] There are also federal anti-discrimination statues addressing credit lending, housing decisions, administration of public funds, and more.[92] It is also important to note recent scholarship in which some scholars have questioned the

[91] Civil Rights Act of 1964, Pub. L. No 88-352, §703(a)(1), 78 Stat. 241, 255 (codified as amended at 42 U.S.C. § 2000e–2(a)(1)). In June 2020, in *Bostock v. Clayton Cty.* the Supreme Court addressed the interpretation of the Title VII prohibition on employment discrimination based on "sex." 140 S. Ct. 1731, 1741 (2020). The Court held that "[a]n employer who fires an individual merely for being gay or transgender defies the law." *Id.* at 1754. While that decision applies to employment discrimination under Title VII, a reasonable inference is that prohibitions against discrimination based on "sex" in other anti-discrimination statutes will be similarly interpreted.

[92] Civil Rights Act of 1968, Pub. L. No. 90-284, Title VII, 82 Stat. 73, 81-89 (codified as amended at 42 U.S.C. §§ 3601-19) ("It shall be unlawful To refuse to sell or rent after the making of a bona fide offer, or to refuse to negotiate for the sale or rental of, or otherwise make unavailable or deny, a dwelling to any person because of race, color, religion, sex, familial status, or national origin."); Equal Credit Opportunity Act Amendments of 1976, Pub. L. No. 94-239, 90 Stat. 251, 251 (codified as amended at 15 U.S.C. § 1691) ("It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction . . . on the basis of race, color, religion, national origin, sex or marital status, or age (provided the applicant has the capacity to contract)"); Civil Rights Act of 1964, Pub. L. No. 88-352, Title VI, 78 Stat. 252, 252-53 (1964) (codified as amended at 42 U.S.C. § 2000d) ("No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance.").

sufficiency of applying traditional interpretations of antidiscrimination law to intersectional discrimination claims. For instance, Pappeo writes that

> Title VII's failure to acknowledge and recognize intersectional discrimination claims disproportionately affects Black female plaintiffs by leaving them with no adequate remedy. I urge courts to adopt intersectionality theory to develop an analytical framework to interpret Title VII to adequately address Black women's claims based on two or more protected categories.[93]

This section focuses on exploring the common legal tests for discrimination, and how those tests inform—or potentially stand to be informed by—approaches to engaging with the various technical and mathematical notions of fairness.[94] Current anti-discrimination law recognizes two general categories of unlawful discrimination that will likely be cited most often in relation to algorithm bias: disparate treatment and disparate impact.[95] It is helpful to first consider the frameworks used for evaluating disparate treatment claims before moving on to a discussion of disparate impact. While the discussion in this Part addresses anti-discrimination statutes and their interpretation, it is important to note that the Equal Protection Clause of the Fourteenth Amendment has also played a key role in addressing discrimination.[96]

[93] Yvette N. A. Pappoe, *The Shortcomings of Title VII for the Black Female Plaintiff*, 22 U. PENN. J.L. & SOCIAL CHANGE 1 (2019).

[94] *See* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017); s*ee also* Barocas & Selbst, *supra* note 23.

[95] There is also what might be considered a third category of anti-discrimination law relating to failure to provide reasonable accommodations. However, algorithm bias will more commonly implicate one or both of disparate impact or disparate treatment as opposed to implicating failure to provide reasonable accommodation.

[96] *See, e.g.*, Brown v. Bd. of Educ., 347 U.S. 483, 495 (1954) (concluding that "in the field of public education the doctrine of 'separate but equal' has no place. Separate educational facilities are inherently unequal. Therefore, we hold that the plaintiffs and others similarly situated for whom the actions have been brought are, by reason of the segregation complained of, deprived of the equal protection of the laws guaranteed by the Fourteenth Amendment."); Loving v. Virginia, 388 U.S. 1, 2 (1967) (considering "whether a statutory scheme adopted by the State of Virginia to prevent marriages between persons solely on the basis of racial classifications violates the Equal Protection and Due Process Clauses of the Fourteenth Amendment" and concluding "these statutes cannot stand consistently with the Fourteenth

The Supreme Court has interpreted the Equal Protection Clause to prohibit disparate treatment, but not disparate impact.[97]

A. Disparate Treatment

Disparate treatment doctrine prohibits differential treatment of individuals based on, or because of, protected attributes such as race, sex, and religion, among others. As the Supreme Court explained in 2003 in *Hazen Paper Co. v. Biggins*:

> In a disparate treatment case, liability depends on whether the protected trait (under the [Age Discrimination in Employment Act], age) actually motivated the employer's decision. The employer may have relied upon a formal, facially discriminatory policy requiring adverse treatment of employees with that trait. Or the employer may have been motivated by the protected trait on an ad hoc, informal basis. Whatever the employer's decisionmaking process, a disparate treatment claim cannot succeed unless the employee's protected trait actually played a role in that process and had a determinative influence on the outcome.[98]

Amendment"); Obergefell v. Hodges, 576 U.S. 644, 675 (2015) (concluding "that the right to marry is a fundamental right inherent in the liberty of the person, and under the Due Process and Equal Protection Clauses of the Fourteenth Amendment couples of the same-sex may not be deprived of that right and that liberty").

[97] *See, e.g.*, Washington v. Davis, 426 U.S. 229, 242 (1976) (explaining that "we have not held that a law, neutral on its face and serving ends otherwise within the power of government to pursue, is invalid under the Equal Protection Clause simply because it may affect a greater proportion of one race than of another. Disproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution"); Vil. of Arlington Hts. v. Metro. Hous. Dev., 429 U.S. 252, 264-65 (1977) ("Our decision last Term . . . made it clear that official action will not be held unconstitutional solely because it results in a racially disproportionate impact. 'Disproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination.' Proof of racially discriminatory intent or purpose is required to show a violation of the Equal Protection Clause." (quoting Washington v. Davis 426 U.S. 229, 242 (1976))).

[98] Hazen Paper Co. v. Biggins, 507 U.S. 604, 610 (1993) (internal citations omitted). *Hazen Paper v. Biggins* addressed a claim under the Age Discrimination in Employment Act (ADEA), a statute under which mixed-motive disparate impact claims are not available. *See* Gross v. FBL Financial Services, Inc., 557 U.S. 167 (2009). As discussed *infra*, in other areas of discrimination law in which mixed-motive claims are permitted, the question of whether

Put simply, the Supreme Court has tended to define unlawful disparate treatment as *intentional* discrimination based on protected attributes.[99] The Supreme Court's 1973 ruling *McDonnell Douglas Corp. v. Green*[100] articulated a burden-shifting framework for examining liability for disparate treatment in employment cases in which there is no direct evidence of intent.[101] The three-step analysis the Court provided in *McDonnell Douglas* for addressing race-based employment discrimination—often referred to as the *McDonnell Douglas* framework or pretext analysis—has since been applied in various contexts beyond employment law, and to cases involving age- and gender-based discrimination in addition to race-based discrimination.[102] The *McDonnell Douglas* analysis requires plaintiffs to first present a *prima facie* case of discrimination.[103] Defendants are then provided the opportunity to present a legitimate, non-discriminatory reason for the allegedly discriminatory practice or policy. Plaintiffs in turn are asked to demonstrate that the offered reason was merely a pretext for discrimination.[104]

consideration of a protected trait must be determinative for a claim to succeed is more complex.

[99] However, it is worth noting that the "disparate treatment" of individuals—i.e., differential treatment of persons based on protected characteristics—may be conceptually distinct from "discriminatory intent," and conflating the two may serve to inhibit antidiscrimination objectives. Eyer, for example, writes that "it is readily apparent why the 'intent' proxy that the Supreme Court often uses for disparate treatment is inadequate." Katie Eyer, *The New Jim Crow Is the Old Jim Crow*, 128 YALE L.J. 1005, 1010 (2018).

[100] 411 U.S. 792 (1973).

[101] *Id.* Questions regarding what exactly constitutes the sort of "direct evidence" necessary to justify diverging from the traditional *McDonnell Douglas* analysis has historically been a matter of significant debate. *See, e.g.*, Brian W. McKay, *Mixed Motives Mix-Up: The Ninth Circuit Evades the Direct Evidence Requirement in Disparate Treatment Cases,* 38 TULSA L. REV. 503, 504 (2003) ("Exactly what Justice O'Connor meant by direct evidence has generated considerable debate, and there have been conflicting interpretations of the requirement among lower courts.") (citation omitted).

[102] Best Med. Intern., Inc. v. Wells Fargo Bank, N.A., 937 F. Supp. 2d 685, 694-95 (E.D.Va. 2013) ("While the *McDonnell Douglas* opinion only addressed Title VII claims of race discrimination, courts apply the burden-shifting framework to other forms of discrimination such as that based on age or gender . . . [c]ourts also apply *McDonnell Douglas* to claims brought under other federal statutes directed at curbing discriminatory practices.").

[103] *McDonnell Douglas Corp.*, 411 U.S. at 802.

[104] A federal district court in 2013 explained the *McDonnell Douglas* steps as follows:

Under a pretext analysis of disparate treatment, a protected trait must be the *determinative* factor in the defendant's decision. An example of such pretextual discrimination would be a case in which an employer has a policy that it will only hire applicants from certain zip codes that it knows are predominantly occupied by white residents. In this example, the employer implements this hiring policy *because of* its racially discriminatory effects. Although zip code is a facially neutral attribute, its use in hiring decisions in this example would constitute disparate treatment due to the employer's discriminatory intent, and the resulting racially discriminatory effects.[105] Notably, disparate treatment—treating individuals differently on the basis of a protected attribute—will not always be unlawful.[106] For example, programs aimed at increasing female participation in STEM careers, while typically involving intentional recruitment of participants in part based on gender, are not generally deemed to be unlawful, because such programs serve a legitimate, non-discriminatory goal. (However, that might change in light of recent Department of Education investigations spurred by complaints that these programs violate Title IX.[107])

McDonnell Douglas provides a potentially useful way to analyze the legality of algorithms that *explicitly* consider protected attributes or their proxies (including when an algorithm developer might argue that such

> (1) [T]he plaintiff establishes a prima facie case of discrimination, which gives rise to a presumption of discrimination; (2) if a prima facie case is shown, the defendant bears the burden of rebutting this presumption of discrimination by offering legitimate, nondiscriminatory reasons for taking action; and (3) if a defendant does so, the burden shifts back to the plaintiff who must show, by a preponderance of the evidence, that the defendant's reasons are pretextual.

Best Med. Intern., 937 F. Supp. 2d at 694.

[105] Of course, in this example, plaintiffs would be required to provide evidence of the defendant's discriminatory intent. Without such evidence, a discrimination claim would be more successfully pursued under a disparate *impact* framework.

[106] *See* Anita M. Alessandra, *When Doctrines Collide: Disparate Treatment, Disparate Impact, and Watson v. Fort Worth Bank & Trust,* 137 U. PA. L. REV. 1755, 1758 (1989).

[107] *See* Teresa Watanabe, *Women-Only STEM College Programs Under Attack for Male Discrimination*, L.A. TIMES (Aug. 20, 2019), https://www.latimes.com/california/story/2019-08-20/women-only-science-programs-discrimination-complaints [https://web.archive.org/web/20190823185534/https://www.latimes.com/california/story/2019-08-20/women-only-science-programs-discrimination-complaints].

consideration is undertaken precisely to *counteract* bias embedded in the input data).[108] However, it will often not be possible to identify the *determinative* factor in a given decision, as is required under a pretext analysis. Moreover, when explicit consideration of protected attributes is lacking, but proxies act as substitutes, it will be difficult to show discriminatory intent, adding a further difficulty for plaintiffs seeking to establish liability for disparate treatment.

Identifying and assigning liability for discrimination becomes more complicated when a decision is made based on a combination of legitimate and illegitimate motives, i.e., when the defendant's justification of its policy or practice is not purely pretextual, and a protected attribute alone is not the determinative factor. Such cases are likely to be assessed under the mixed-motive framework[109] (sometimes called the motivating-factor method of proof)[110] which requires only that the plaintiff prove that a protected attribute was a motivating or substantial factor in determining an action.[111] If discriminatory intent is

[108] This approach to bias mitigation has been advocated by scholars including Mayson: "[I]ncluding race as an input variable would promote accuracy and racial equity at the same time." Mayson, *supra* note 5, at 2265. Other scholars have suggested that if a sufficiently large number of other inputs are available, then the impact of considering information about a protected class could have less relevance. *See, e.g.* Barocas & Selbst, *supra* note 23, at 695 ("The use of protected class as an input is usually irrelevant to the outcome in terms of discriminatory effect, at least given a large enough number of input features.").

[109] Notably, mixed-motive liability is one example of the ways in which "disparate treatment" may differ in important ways from "discriminatory intent." Mixed-motive analysis has historically been reserved for those cases in which discriminatory intent is not the sole driving factor in a decision, and yet there is clear, *direct* evidence of discrimination. In such cases, it is not obvious that the allegedly discriminatory decision/policy (i.e., termination of employment) amounts to discriminatory *intent*, though the direct evidence certainly suggests that disparate treatment of individuals based on protected attributes *is* involved.

[110] *See, e.g.,* Tristin K. Green, *Making Sense of the McDonnell Douglas Framework: Circumstantial Evidence and Proof of Disparate Treatment Under Title VII*, 87 CALIF. L. REV. 983, 984 (1999) ("Title VII in light of Supreme Court Doctrine and the Civil Rights Act of 1991 . . . proposes that the two currently recognized inferential methods of proof for proving intentional discrimination, termed here the 'motivating-factor' method of proof and the 'falsity-of-proffered-reason' method of proof, are available as alternate methods."); Univ. of Tex. Sw. Med. Ctr. v. Nassar, 570 U.S. 338, 354 (2013) ("If Congress had desired to make the *motivating-factor standard* applicable to all Title VII claims . . . ." (emphasis added)).

[111] *See, e.g.*, Price Waterhouse v. Hopkins, 490 U.S. 228, 250 (1989) ("In saying that gender played a motivating part in an employment decision, we mean that, if we asked the employer at the moment of the decision what its reasons were and if we received a truthful response, one of those reasons would be that the applicant or employee was a woman."); William R. Corbett,

one non-determinative factor among multiple factors involved in choosing a particular course of action, the question then becomes to what extent the defendant should be held liable.

There has been a great deal of variation, both in the courts and in legal scholarship, in not only the understanding of the extent of liability for mixed-motive discrimination in various contexts, but also *when* (if at all) a mixed-motive analysis is available to defendants. To give one example, in *Mhany Management v. Incorporated Village of Garden City*, a federal court in the Eastern District of New York explained that:

> in this Circuit, although a defendant in an FHA case can escape liability entirely if it proves it would have rendered the same decision had it not considered impermissible reasons, a defendant in a Title VII case can only reduce monetary damages and avoid certain injunctive relief based on liability if it makes the same showing. To be sure, this more defendant-friendly standard under the FHA cuts against the broad remedial interpretation typically accorded to the FHA . . . and the general rule that Title VII and the FHA be construed in a similar manner.[112]

This suggests (at least in the Second Circuit) an interpretation that, although defendants may be subject by law to varying degrees of liability for discrimination depending on the context, there is little normative or intuitive support for this context-dependent variation. Due to the growing complexity of algorithmic tools, the decisions and recommendations that they make will sometimes be made based on a *combination* of legitimate and (usually, but not always unintentionally) illegitimate factors. For this reason, understanding when and to what extent a mixed-motive analysis applies will be a consideration in addressing algorithmic discrimination in the future.

Fixing Employment Discrimination Law, 62 SMU L. REV. 81, 85 (2009) ("First, the plaintiff must prove that the protected characteristic was a motivating or substantial factor . . . .").

[112] Mhany Mgmt. v. Inc. Vill. of Garden City, 985 F. Supp. 2d 390, 422-23 (E.D.N.Y. 2013).

B. Disparate Impact

Disparate impact is implicated when discriminatory intent or motive is lacking (or present but impossible or impractical for plaintiffs to show), yet the practice in question has a material adverse impact on a protected group.[113] In contrast with disparate treatment, disparate impact does not require showing discriminatory intent.[114] Disparate impact case law first developed in the context of litigation regarding employment discrimination.[115] In *Griggs v. Duke Power Co*. in 1971, the Supreme Court concluded that under the Civil Rights Act of 1964, "practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the *status quo* of prior discriminatory employment practices."[116] The *Griggs* Court also wrote that the Civil Rights Act of 1964 "proscribes not only overt discrimination, but also practices that are fair in form, but discriminatory in operation."[117] In explaining disparate impact in 1988 in *Watson v. Ft. Worth Bank & Trust*, the Court wrote that some "practices, adopted without a deliberately discriminatory motive, may in operation be functionally equivalent to intentional discrimination."[118]

To demonstrate disparate impact in violation of antidiscrimination law, it is not sufficient for a plaintiff to simply point to a statistical disparity.[119] A plaintiff must also show a causal relationship tying a challenged policy to a resulting disparate impact.[120] As the *Watson* Court wrote:

> [o]nce the employment practice at issue has been identified, causation must be proved; that is, the plaintiff

[113] Barocas & Selbst, *supra* note 23, at 701.
[114] *Id.* at 676.
[115] *See* Michael Selmi, *The Evolution of Employment Discrimination Law: Changed Doctrine for Changed Social Conditions* 1-2 (GWU L. Sch., Public Law Research Paper No. 2014-8, 2014), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2430378.
[116] 401 U.S. 424, 430 (1971) (emphasis in original).
[117] *Id.* at 431.
[118] 487 U.S. 977, 987 (1988).
[119] Marcel C. Garaud, *Legal Standards and Statistical Proof in Title VII Litigation: In Search of a Coherent Disparate Impact Model,* 139 U. PA. L. REV. 455, 473 (1990).
[120] *Id.*

> must offer statistical evidence of a kind and degree sufficient to show that the practice in question has caused the exclusion of applicants for jobs or promotions because of their membership in a protected group.[121]

This emphasis on the importance of causation in establishing disparate impact liability in the context of employment was echoed by the Court a year later in *Wards Cove v. Antonio*.[122] More recently, the Supreme Court in 2015 articulated similar causation requirements for disparate impact cases arising under the Fair Housing Act. The Court explained that "a disparate-impact claim that relies on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity."[123] The Court underscored the importance of establishing a "robust causality" linking a defendant's policies with disparate impact, explaining that a "robust causality requirement ensures that '[r]acial imbalance . . . does not, without more, establish a prima facie case of disparate impact' and thus protects defendants from being held liable for racial disparities they did not create."[124]

The Civil Rights Act of 1991 specified the burden of proof for disparate impact cases in the context of employment, providing, for example, that "an unlawful employment practice based on disparate impact" can be established if

> a complaining party demonstrates that a respondent uses a particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin and the respondent fails to demonstrate

[121] *Watson*, 487 U.S. at 994.

[122] 490 U.S. 642, 656-58 (1989).

[123] Tex. Dept. of Hous. & Cmty. Aff. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 542 (2015); it is worth noting that the requirement that a plaintiff point to a particular policy responsible for the observed disparate impact does not always apply to discrimination cases under Title VII. Instead, "if the complaining party can demonstrate to the court that the elements of a respondent's decisionmaking process are not capable of separation for analysis, the decisionmaking process may be analyzed as one employment practice." Civil Rights Act of 1991 §105(B)(i) (codified as amended at 42 U.S.C. 2000e-2).

[124] *Inclusive Cmtys. Project,* 576 U.S. at 542 (quoting *Wards Cove Packing*, 490 U.S. at 653). The Court also cited 42 U.S.C. §2000e–2(k), which addresses the "burden of proof in disparate impact cases" and was enacted as part of the Civil Rights Act of 1991 (see §105).

> that the challenged practice is job related for the position in question and consistent with business necessity.[125]

The statute further provides that a "demonstration that an employment practice is required by business necessity may not be used as a defense against a claim of intentional discrimination under this subchapter."[126]

Title VI of the Civil Rights Act of 1964, which prohibits discrimination "on the ground of race, color, or national origin . . . under any program or activity receiving Federal financial assistance," also bars practices with unjustified disparate impact.[127] As guidance from the U.S. Department of Justice explains, under Title VI as a first step, "to establish an adverse disparate impact, the investigating agency must (1) identify the specific policy or practice at issue; (2) establish adversity/harm; (3) establish significant disparity; and (4) establish causation."[128] As a second step, "if the evidence establishes a prima facie case of adverse disparate impact . . . courts then determine whether the recipient has articulated a 'substantial legitimate justification' for the challenged policy or practice."[129] Moreover, "the discriminatory policy or practice must also bear a demonstrable relationship to that goal."[130] If the defendant succeeds in this showing, the court may still rule in favor of the plaintiff if the plaintiff can demonstrate (as a third step) that "there are alternative practices that may be comparably effective with less disparate impact."[131]

The question of whether disparate impact claims could be brought in relation to housing discrimination was not definitively settled until 2015. In *Texas Department of Housing and Community Affairs v.*

[125] Civil Rights Act of 1991, Pub. L. No. 102-66, §105(a), 105 Stat. 1071, 1075 (1991) (codified as amended at 42 U.S.C. § 2000e–2(k)(1)(A)(i)). This is one of two ways to establish disparate impact under this statute; the other is articulated in 42 U.S.C. § 2000e–2(k)(1)(A)(ii).
[126] 42 U.S.C. § 2000e–2(k)(2).
[127] Civil Rights Act of 1964, Pub. L. No. 88-352, § 601, 78 Stat. 241 (codified as amended at 42 U.S.C. § 2000d).
[128] CIVIL RIGHTS DIVISION, U.S. DEPARTMENT OF JUSTICE, TITLE VI LEGAL MANUAL § VII (2019).
[129] *Id.*
[130] *Id.* § VII(C)(2).
[127] *Id.*

Inclusive Communities Project, the Court held that "[r]ecognition of disparate-impact claims is consistent with the [Fair Housing Act's] central purpose"[132] and also addressed the associated burden-shifting framework.[133]

Federal courts have applied similar burden-shifting frameworks in disparate impact litigation in relation to other antidiscrimination statutes as well (e.g., Title VII of the Civil Rights Act of 1964[134]), though, interestingly, not with respect to the Age Discrimination in Employment Act.[135] Thus, burden-shifting is used both in relation to disparate treatment (i.e., *McDonnell Douglas* and its progeny) as well as disparate impact.

To make matters more complex, avoiding disparate *treatment* liability may sometimes inevitably result in disparate outcomes for different groups. Similarly, actions taken to avoid disparate *impact* liability may require treating individuals differently based on group membership, thereby implicating disparate treatment. Since both disparate treatment and disparate impact are forms of unlawful discrimination, how can such decisions be made? The Supreme Court partially addressed this issue in the context of employment discrimination in 2009 in *Ricci v. Destafano,* a case that arose after a group of "New Haven firefighters

[132] 576 U.S. 519, 539 (2015).

[133] *See id.* at 521-42. That said, a new pending rule from the Department of Housing and Urban Development stands to alter how FHA disparate impact cases are adjudicated. *See* HUD's Implementation of the Fair Housing Act's Discriminatory Impact Standard, 85 Fed. Reg. 60288 (Sep. 24, 2020) (to be codified at 24 C.F.R. pt. 100). The rule was scheduled to take effect on October 26, 2020, though it has been preliminarily enjoined. *See also* Virginia Foggo & John Villasenor, *Algorithms, Housing Discrimination, and the New Disparate Impact Rule*, 22 COL. SCI. TECH. L. REV (forthcoming 2020).

[134] *See, e.g.*, Watson v. Fort Worth Bank & Tr., 487 U.S. 977 (1988).

[135] For example, the Third Circuit has ruled that a similar framework applies, though it involves only two, less stringent steps. "[U]nder the ADEA, a plaintiff must (1) identify a specific, facially neutral policy, and (2) proffer statistical evidence that the policy caused a significant age-based disparity . . . . Once a plaintiff establishes a *prima facie* case, an employer can defend by arguing that the challenged practice was based on 'reasonable factors other than age'—commonly referred to as the 'RFOA' defense." Karlo v. Pittsburgh Glass Works, LLC, 849 F.3d 61, 69 (3rd Cir. 2017). The defendant need not demonstrate that such a practice is based on business necessity, and "the employer only needs to show that it relied on a 'reasonable' factor, not that 'there are [no] other ways for the employer to achieve its goals.'" *Id.* at 69-70.

took examinations to qualify for promotion to the rank of lieutenant or captain."[136] As the Court explained, when "the examination results showed that white candidates had outperformed minority candidates, the mayor and other local politicians opened a public debate that turned rancorous" [137] and the city ultimately "took the side of those who protested the test results [and] threw out the examinations."[138] This led to a lawsuit by "white and Hispanic firefighters who likely would have been promoted based on their good test performance."[139]

The *Ricci* Court noted that "certain government actions to remedy past racial discrimination—actions that are themselves based on race—are constitutional only where there is a 'strong basis in evidence' that the remedial actions were necessary."[140] The Court ultimately ruled in favor of the plaintiffs, concluding that disparate treatment in this case was not justified because the defendant failed to demonstrate with a "strong basis in evidence" that defendant would otherwise have faced disparate impact liability.[141] Although disparate treatment proved *not* to be permissible in the *Ricci* case, the Court's language makes clear that actions taken to avoid disparate impact and that as a result involve disparate treatment can sometimes be justified.[142]

Notably, the Court acknowledged that if "the City faces a disparate-impact suit, then in light of our holding today it should be clear that the City would avoid disparate-impact liability based on the strong basis in evidence that, had it not certified the results, it would have been subject to disparate treatment liability."[143] This indicates that an entity can

[136] 557 U.S. at 562 (2009).
[137] *Id.*
[138] *Id.*
[139] *Id.*
[140] *Id.* at 582 (quoting Richmond v. J. A. Croson Co., 488 U.S. 469, 500 (1989)).
[141] *Id.* at 592.
[142] *See id.*
[143] Note, however, that in 2011 in *Briscoe v. City of New Haven* the Second Circuit Court of Appeals vacated a judgment by the United States District Court for the District of Connecticut that ruled that the holding in *Ricci* necessarily precluded the possibility of disparate impact claims arising from the city's subsequent certification of test results. *See* Briscoe v. City of New Haven, 654 F.3d 200, 209 (2d Cir. 2011) ("Briscoe's claim is neither precluded nor properly dismissed. Ricci did not substantially change Title VII disparate-impact litigation or preclusion principles in the single sentence of dicta targeted at the parties in this action.").

sometimes avoid disparate impact liability if its action is necessary to avoid disparate treatment liability.

IV. Algorithmic Fairness in Context

A. Rethinking *Ricci*

The implications of *Ricci* on the legal viability of approaches to mitigating algorithmic bias that involve explicit consideration of protected attributes has received significant attention among legal scholars.[144] The *Ricci* Court explained that "Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race."[145] The Court underscored that the problem in *Ricci* was that

> after the tests were completed, the raw racial results became the predominant rationale for the City's refusal to certify the results. The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process.[146]

Multiple scholars have recognized that performing an after-the-fact adjustment of algorithmic outputs based on race or another protected attribute will result in a clearly identifiable adverse impact on specific individuals, and is thus likely to run afoul of *Ricci* on that basis. For example, according to Kim, *Ricci* "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."[147] Hellman has a similar interpretation, writing that

[144] *See, e.g.*, Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803 (2020).
[145] 557 U.S. at 585.
[146] *Id.* at 593.
[147] Kim, *supra* note 94, at 191.

> [t]he 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.[148]

Barocas and Selbst have written that requiring "results-focused balancing . . . will pose constitutional problems."[149] Kroll observed that if "an agency runs an algorithm that has a disparate impact, correcting those results after the fact will trigger the same kind of analysis as" in *Ricci*.[150]

Thus, post-*Ricci*, it will generally not be permissible to apply race-based (or gender-based, etc.) modifications to algorithmic *outputs* in an attempt to rectify disparate impact that has only become evident when the results appear. But *prospective* modifications to an algorithm that introduce explicit race-based (or gender-based, etc.) considerations in an attempt to ensure unbiased outputs, performed before there are any identifiable people impacted, are less likely to run afoul of *Ricci*. That does not mean, however, that they will in the end be more likely to pass legal muster. The question of *when* an adjustment is made to correct for bias is in some sense a distinction without a difference—at least from a logical standpoint, if not a legal one.

Consider an algorithm used in hiring that, due to biases in the input data, tends to produce higher scores for men than for equally qualified women.[151] An after-the-fact downward adjustment of the scores for men would create a set of identifiable individuals adversely impacted by the

[148] Hellman, *supra* note 25, at 864 ("The most promising way to enhance algorithmic fairness is to improve the accuracy of the algorithm overall. And we can do that by permitting them to use protected traits (like race and sex) within the algorithm").
[149] Barocas & Selbst, *supra* note 23, at 726.
[150] Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 694 (2017).
[151] In this example, rather than considering an algorithm that produces a binary output, we are considering an algorithm that produces a score.

adjustment, thus potentially running afoul of *Ricci*. Suppose that, instead, the algorithm developers had designed the algorithm so that its internal computations are performed in a manner that ensures that the eventual outputs have equal average scores for men and women. In this case (assuming in this example that average score was a metric used for comparison), there is no disparate impact visible at the level of output scores, but there was disparate treatment inside the algorithm. And, that disparate treatment could, if the internal workings of the algorithm were revealed, be asserted as the basis for adverse impact on men who are evaluated using the algorithm.

More generally, these questions have implications for *when* consideration of protected attributes can occur and are also important for thinking about the extent to which statistical fairness metrics may be used to monitor outcomes and update algorithms. There are three (non-mutually exclusive) points in time at which it would be possible to consider—and perhaps make adjustments based on—protected attributes. Pre-processing is used before data is input to an algorithm,[152] in-processing refers to techniques applied within an algorithm,[153] and post-processing refers to adjustments applied to algorithm outputs.[154] While pre-processing of historical data used to train an algorithm as an approach to mitigating bias is certainly a valuable approach, it will not always prevent unfair outcomes for individuals being contemporaneously assessed, especially when used in isolation. In-processing or post-processing approaches may sometimes be necessary to address fairness concerns in real-time.

That being said, does the fact that a pre-processing approach was taken prohibit the additional use of in- or post-processing approaches in light of *Ricci?* For example, could an observed lack of equalized odds—despite pre-processing of data—serve as a "strong basis in evidence" of disparate impact liability? And, if so, would this justify updating an algorithm to weigh some factors differently for different groups in future evaluations, given that doing so is an alternative practice that

[152] Brian d'Alessandro et al., *Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification*, 5 BIG DATA 120, 130 (2017).
[153] *Id.*
[154] *Id.*

creates a less disparate impact? Moreover, is it possible that a failure to implement in- or post-processing approaches that are known to increase accuracy and mitigate bias could be viewed as a "policy" that satisfies the robust causality requirement necessary to establish disparate impact liability?[155] Existing case law stops well short of answering these questions.

B. Rethinking Disparate Impact and Disparate Treatment

As discussed above, current antidiscrimination law rightly recognizes that a prohibition on disparate treatment alone is not always sufficient to prevent discrimination. Disparate impact doctrine provides a more expansive set of protections against discrimination because it removes the need to show intent. Disparate impact instead focuses the inquiry on whether a disparity among group outcomes can be attributed to specific policies of the entity (e.g., a company) making the allegedly discriminatory decisions, rather than to external disparities that arose independently of the entity making the evaluation. For instance, the *Inclusive Communities* Court wrote that "a disparate-impact claim that relies on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity"[156] and that a "robust causality requirement ensures that '[r]acial imbalance . . . does not, without more, establish a prima facie case of disparate impact' and thus protects defendants from being held liable for racial disparities they did not create."[157]

As well-intentioned as these frameworks are, it is important to consider how they might inadvertently impede efforts to introduce fairness in an algorithmic context. Consider an algorithm for criminal risk that uses employment status as one of the inputs. There is a well-documented correlation between unemployment rate and neighborhood of residence. As a 2018 article in the *Washington Post* noted, "the unemployment rate is 10 percent in impoverished neighborhoods, compared with the overall

[155] Foggo & Villasenor, *supra* note 133.

[156] Tex. Dep't of Hous. & Cmty. Aff. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 542 (2015).

[157] *Id.* (quoting Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 653 (1989)).

national rate of 4.1 percent."[158] It is also well-documented that one legacy of segregation is that the majority of residents in many impoverished urban neighborhoods are persons of color. A 2009 academic paper explained that, "a vast body of work has documented the patterns of racial segregation and concentrated poverty in U.S. metropolitan areas."[159] The combination of these two correlations means that, all else being equal, a criminal risk assessment algorithm that considers employment status *without* adjusting for neighborhood variations in employment opportunities would typically end up producing outputs that tend to give higher risk scores to persons of color.

But is this unlawful disparate impact in the legal sense? The company that created the algorithm might argue that the answer is "no," because it views consideration of employment status as necessary for more accurate prediction of crime, a goal which serves a legitimate aim. The company might further explain that the geographic and racial inequities in employment opportunities are exactly the sort of statistical disparities that the Supreme Court had in mind in emphasizing the importance, when conducting disparate impact inquiries, of "protect[ing] defendants from being held liable for racial disparities they did not create."[160]

Under current legal precedent, the company's argument might prevail. But suppose that the company decided that it wanted to revise the algorithm so that, instead of merely serving as a conduit to propagate externally created inequities, it would attempt to remove them. More specifically, the company might decide that treating employment status in a "fair" way in computing criminal risk would include an adjustment accounting for neighborhood variation in employment opportunities. Under this modified algorithm, the risk score penalty ascribable to being unemployed would be lower for people from impoverished neighborhoods than for those from better-off neighborhoods.

[158] *In Poor Neighborhoods, Unemployment Is 10%, Survey Says*, WASH. POST (Feb. 4, 2018), https://www.washingtonpost.com/national/in-poor-neighborhoods-unemployment-is-10percent-survey-says/2018/02/04/a107828e-09f8-11e8-8b0d-891602206fb7_story.html [https://perma.cc/4GYP-7QWX].

[159] Theresa L. Osypuk et al., *Quantifying Separate and Unequal: Racial-Ethnic Distributions of Neighborhood Poverty in Metropolitan Americ*a, 45 URB. AFF. REV. 25, 27 (2009).

[160] *Inclusive Cmtys. Project,* 576 U.S. at 542.

However well-intentioned such a step might be, it could result in unlawful racially disparate *treatment*. A criminal risk assessment algorithm that weighs employment status differently depending on neighborhood would end up weighing employment status in a manner that correlates with race. And what about the requirement of a disparate treatment enquiry to show intent? The company might argue that it had exercised intent only with respect to considering neighborhood, not with respect to considering race. But given the level of segregation in many cities, there would be a strong counterargument that consideration of neighborhood is merely a pretext for consideration of race, thus invoking a *McDonnell Douglas* disparate treatment analysis.[161] Thus, concern over the liabilities associated with exactly this outcome might lead a company to decide *not* to use its algorithm to attempt to counteract the biases contained in employment data.

It is also possible that a court might deem consideration of neighborhood to be necessary for avoiding disparate impact liability, thus absolving the company of disparate treatment liability. However, it is unclear when "policies" that involve consideration of protected attributes constitute the sort of "alternative practices" endorsed by disparate impact frameworks.[162] In relation to employment, for example, the Supreme Court has explained that "factors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals."[163] Such factors will not be particularly useful in assessing the comparative efficacy of algorithmic tools that explicitly consider

[161] *See*, *e.g.*, Ave. 6E Invs., LLC v. City of Yuma, 818 F.3d 493, 504 (9th Cir. 2016) ("The court analyzes whether a discriminatory purpose motivated the defendant by examining the events leading up to the challenged decision and the legislative history behind it, the defendant's departure from normal procedures or substantive conclusions, and the historical background of the decision *and whether it creates a disparate impact*." (emphasis added)).

[162] A respondent will only be held liable for disparate impact if there is an alternative practice available with a less disparate impact. When that alternative practice constitutes disparate treatment with benign intent, we must now assess whether or not it is necessary for the avoidance of disparate impact liability. Whether or not a company would have been liable for disparate impact depends on whether or not their decision could have been made using an alternative practice. Note the circularity of this process.

[163] Watson v. Ft. Worth Bank & Tr., 487 U.S. 977, 998 (1988).

protected attributes and those that do not, as consideration of additional attributes would not generally result in any monetary or otherwise tangible cost to the algorithm user.[164]

Algorithms that explicitly consider protected attributes like race, gender, or sex also implicate the complex legal questions around affirmative action. Affirmative action case law, which is most developed in relation to the admissions practices of public universities, is a distinct (but often overlapping) domain from the law relating to anti-discrimination in employment, housing, credit, etc. Supreme Court rulings on affirmative action to date have generally (though not exclusively) focused on the Equal Protection Clause of the Fourteenth Amendment.

In *Regents of the University of California v. Bakke* in 1978, the Court concluded that race-based quotas violate the Equal Protection Clause.[165] At the same time, however, the Court explained that it was improper to completely prevent a university from considering race.[166] In *Grutter v. Bollinger* in 2003, the Court wrote that "student body diversity is a compelling state interest that can justify the use of race in university admissions,"[167] that "the Equal Protection Clause does not prohibit the [University of Michigan] Law School's narrowly tailored use of race in admissions decisions to further a compelling interest in obtaining the educational benefits that flow from a diverse student body,"[168] and that a school's use of race must "remain flexible enough to ensure that each applicant is evaluated as an individual and not in a way that makes an

[164] If algorithmic consideration of a protected attribute gave rise to a disparate treatment claim, the resulting litigation would of course involve monetary costs; however, this brings us back to our original point that it is unclear whether or not a plaintiff could succeed on such a claim of disparate treatment.

[165] 438 U.S. 265, 319-20 (1978) ("The fatal flaw in petitioner's preferential program is its disregard of individual rights as guaranteed by the Fourteenth Amendment.").

[166] *Id.* at 320 ("In enjoining petitioner from ever considering the race of any applicant, however, the courts below failed to recognize that the State has a substantial interest that legitimately may be served by a properly devised admissions program involving the competitive consideration of race and ethnic origin.").

[167] 539 U.S. 306, 325 (2003).

[168] *Id.* at 331.

applicant's race or ethnicity the defining feature of his or her application."[169]

In 2003 the Court ruled on *Gratz v. Bollinger*, a separate affirmative action decision also regarding the University of Michigan.[170] *Gratz* arose from a challenge to a policy used by the College of Literature, Science, and the Arts to "automatically distribute[] 20 points to every single applicant from an 'underrepresented minority' group, as defined by the University."[171] The Court held that "because the University's use of race in its current freshman admissions policy is not narrowly tailored to achieve respondents' asserted compelling interest in diversity, the admissions policy violates the Equal Protection Clause, Title VI [of the Civil Rights Act of 1964] and 42 USC §1981."[172]

In 2013 and 2016 respectively, the Court ruled on a pair of cases relating to undergraduate admissions at the University of Texas at Austin. In 2013 in *Fisher v. University of Texas* (sometimes called *Fisher I* to distinguish it from the 2016 decision of the same name), the Court explained that "[r]ace may not be considered unless the admissions process can withstand strict scrutiny"[173] and that under strict scrutiny a "university must clearly demonstrate that its 'purpose or interest is both constitutionally permissible and substantial, and that its use of the classification is 'necessary . . . to the accomplishment' of its purpose.'"[174]

In 2016 in *Fisher v. University of Texas* (often termed *Fisher II*), the Court upheld the consideration of race in admissions, explaining that "[t]he fact that race consciousness played a role in only a small portion of admissions decisions should be a hallmark of narrow tailoring, not evidence of unconstitutionality."[175]

[169] *Id.* at 337.

[170] 539 U.S. 244 (2003).

[171] *Id.*

[172] *Id.* at 275-76.

[173] 570 U.S. 297, 309 (2013).

[174] *Id.* at syllabus (quoting Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 305 (1978)).

[175] 136 S. Ct. 2198, 2206 (2016).

Several interesting observations directly relevant to algorithms flow from these cases. At least in the context of public university admissions decisions, *Gratz* indicates that algorithms including a point adjustment based on a protected characteristic would fail to pass constitutional muster. But *Fisher II* indicated that considering race in a narrowly tailored way *is* permissible. Given that algorithms are numerical by definition, that creates a potential paradox: while the Supreme Court has permitted consideration of race, doing so in an algorithm would often involve numerical scores (or their equivalent) specifically reflecting, in part, race—something that the Court has deemed impermissible.

Thus, existing affirmative action doctrine in relation to university admissions, like disparate impact doctrine in relation to employment, housing, and credit, fails to offer a clear answer to the question of the extent to which explicit consideration of protected attributes—where doing so contributes to the accuracy or lessens the disparate impact of an algorithm's outcomes for a particular group—is a legally justified practice, or if it would only serve to subject entities using the algorithm to disparate treatment liability.

This example illustrates how discrimination law frameworks can actually create a disincentive for decision-making entities such as companies and universities to use algorithms as a mechanism to counteract bias. And that disincentive could be even stronger than the above example suggests. To see why, consider a variant in which a company attempts to use an algorithm to correct for the neighborhood-level inequities in employment opportunities, but does not get the correction quite right: it either slightly under-corrects or slightly overcorrects. This would arguably create a situation implicating both disparate treatment *and* disparate impact.[176] Disparate treatment would arguably be present for the reasons explained above, i.e., the intentional decision to consider neighborhood—and, by implication, race—in

[176] A policy that constitutes disparate treatment in the legal sense may still have a disparate impact (in the colloquial sense) on a group of interest; however, the entity responsible for such a policy would only be found liable for one of the two, given the fact that the same policy would be in question with respect to both theories of liability. This is not to say that a company could not simultaneously be found liable for disparate treatment based on one particular policy, and disparate impact for a different policy.

computing risk scores. Disparate impact would be present because, unlike when the company made no effort to correct for the inequities, under this modified scenario in which the company attempted—with only partial success—to perform that correction, the racial disparities in the output of the algorithm *would* in part be directly attributable to a specific policy. In other words, the "reward" for attempting to correct for bias might be liability under disparate treatment doctrine that is not mitigated by the lower degree of adverse impact (relative to an approach that lacked the corrective steps described above) on a particular group.[177]

There is also the question of how the perspectives gained from studying algorithmic fairness can lead to new ways of thinking about anti-discrimination law. With that in mind, in the context of algorithms, the traditional categories of disparate impact and disparate treatment risk oversimplifying what will in fact be a more complex landscape. That does not mean that those frameworks lose their utility. It is easy to imagine scenarios in which one of the steps in investigating an allegedly discriminatory algorithm would be evaluating it to see whether disparate treatment and/or disparate impact are present.

But that should not be the end of the story. As noted in the earlier example in which a company makes an effort to include a correction for historical biases in its algorithm but does not get that correction exactly right, it is possible to act in complete good faith, to create an algorithm which is potentially far less biased than the human-based decision system it might replace or augment, yet *still* run afoul of either disparate treatment or disparate impact theories of liability.

Anti-discrimination law should leave sufficient flexibility to enable innovation using algorithms in ways that can reduce bias. It is not at all surprising that disincentives to such innovation exist, because the current legal landscape is the result of decades of legislation and litigation on discrimination in a non-algorithmic context. But given the inevitable increase over the coming decade in algorithmically-driven (or algorithmically-informed) decision making, a more flexible framework

[177] *Fisher II*, 136 S. Ct. at 2203.

is needed that can still hold algorithm designers accountable for intentional or inadvertent discrimination, while also giving them the opportunity to innovate to produce improved solutions.

To give a concrete example of how this might work, consider a company that makes an algorithm that is accused of being discriminatory. In accordance with traditional discrimination law, litigation could determine whether there is disparate impact and/or disparate treatment. If one or both of those are alleged, at some point in the proceedings the company should be given an opportunity to show that, had it made different design choices, the resulting algorithm would have been demonstrably *more* discriminatory. This is not to suggest that such a showing would automatically absolve an entity of all liability; however, it suggests that this sort of contextual information will be an important consideration.

C. The Importance of Metric Transparency

Where do algorithmic fairness measures fit into this picture, and how might they help resolve some of these tensions? As an initial matter, it is important to underscore that fairness metrics examine predictions (or scores) and outcomes, not the underlying algorithms used to make those predictions or the state of mind of algorithm developers. The presence of a statistical disparity—e.g., that members of a particular group are denied loans at a disproportionately high rate—is not sufficient to infer disparate treatment, as it would not show that the algorithm was designed with intent to create adverse outcomes for a particular group. But neither is it sufficient to show disparate impact, since it does not establish causality between any "policies" of the company that developed or implemented the algorithm and the observed statistical disparities.[178]

Moreover, those wishing to show causation between discriminatory outcomes and the inclusion of a particular attribute in the design of the algorithm will often have difficulties gaining access to code, which is

[178] A related but different question is whether a company (e.g., a financial institution that *uses* the algorithm) might be exposed to disparate impact liability.

often proprietary in nature.[179] Even when access to code is granted (e.g., under a protective order in litigation), the ways in which an algorithm makes use of a particular attribute will sometimes be sufficiently complex to make it difficult to establish clear causation.[180] To add yet another hurdle, in anticipation of a possible finding of causation, the code developers might invoke the business necessity defense.[181]

While fairness metrics alone are thus not sufficient to prove disparate impact or disparate treatment, they are nonetheless an extremely useful tool for identifying statistical disparities, which can, in turn, provide a basis for continued investigation. That raises the question of what constitutes a statistical disparity.[182] As discussed in Part II, when the base rates across two groups regarding the parameter of interest (e.g., whether a borrower will pay a loan, or whether a student will pass a test) are unequal, it will nearly always be possible to find a fairness metric that shows a statistical disparity.[183] For instance, if algorithm designers have optimized for positive predictive value, then the rates of false positives and false negatives will generally be different across the two groups. On the other hand, optimizing to minimize and equalize false positive and false negative rates will usually preclude satisfying

[179] *Cf.* Steven M. Bellovin et al., *Seeking the Source: Criminal Defendants' Constitutional Right to Source Code***,** 17 OHIO ST. TECH. L.J. 1 (2021) (discussing difficulty suffered by criminal defendants seeking source code as potentially exculpatory evidence).

[180] *See* Yavar Bathaee *The Artificial Intelligence Black Box and the Failure of Intent and Causation* 31 HARV. J.L & TECH. 890, 891-92 (2018).

[181] Stephanie Bornstein, *Antidiscriminatory Algorithms* 70 ALA. L. REV. 519, 525 (2018).

[182] For example, Equal Employment Opportunity Commission regulations provide guidance on this question. 24 C.F.R. § 1607.4(D) states that "[a] selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact." This is often referred to as the "eighty percent rule." While the eighty percent rule may be useful in cases of more obvious discriminatory impacts, it is arbitrary in terms of statistical significance and so has little utility in more nuanced cases of discrimination. 24 C.F.R. § 1607.4(D) acknowledges this in noting that "[s]maller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms or where a user's actions have discouraged applicants disproportionately on grounds of race, sex, or ethnic group." Identifying group outcomes that are sufficiently disparate to constitute discrimination will be a largely context-dependent endeavor, and it will be important to compare statistical discrepancies in algorithmic outcomes across various fairness metrics before drawing conclusions.

[183] *See supra* Part II.

statistical parity. Analogous incompatibilities can also be identified for other combinations of fairness metrics.

These sorts of incompatibilities are mathematical inevitabilities that will confront all algorithm developers, including (and perhaps especially) those who are particularly attuned to the goal of addressing bias. Developers will typically need to choose one fairness measure that they will seek to optimize, knowing that in doing so they are creating the impossibility of optimizing under the others.[184] Relatedly, a person wanting to criticize an algorithm will always be able to identify multiple metrics that show differences across groups that could be labeled "statistical disparities."

This highlights the importance of what might be called *metric transparency*: Companies that design algorithms should be transparent about which fairness metric(s) they are using for optimization, the reasons for choosing that metric, and the extent to which that choice leads to disparities under other commonly used fairness metrics. As a voluntary best practice,[185] metric transparency offers algorithm developers a mechanism to (1) convey information about the nature and extent of their efforts to achieve fairness, and (2) to provide important context that others, including the press, advocacy organizations, and courts, can use in their own evaluations. An additional advantage of metric transparency from the standpoint of algorithm designers is that it can be practiced without revealing the core aspects of the underlying algorithm that are typically closely held trade secrets.[186]

[184] Algorithm developers can also choose to simultaneously optimize under two fairness metrics, but as discussed earlier this will often lead to constraints that, while mathematically feasible, are problematic from a policy standpoint.

[185] It is also possible to contemplate *requiring* that companies disclose which metric(s) they have used in attempting to ensure algorithm fairness, though this would need to be done in a manner that would avoid negatively impacting the incentives and opportunities to innovate. For example, a company might choose not to optimize an algorithm for any one fairness measure and instead to be intentionally sub-optimal (but not too far from optimal) on two different measures that would be mathematically impossible to satisfy simultaneously. Any regulatory framework regarding metric disclosure would need to leave sufficient flexibility for companies to innovate and think outside the box as they decide how to address fairness during the algorithm design process.

[186] We note that transparency regarding the *metrics* that developers have used to assess fairness when creating an algorithm is different from transparency regarding the *underlying*

From the standpoint of civil society organizations and others aiming to scrutinize algorithms and promote algorithmic fairness, metric transparency can reduce the cost and time burden of inquiries, and can also make it easier to identify and focus on the algorithms most likely to be problematic.[187] This points to another advantage of metric transparency. Companies might initially be tempted to avoid metric transparency under the logic that it is always best to make potential adversaries work as hard as possible to obtain any information about the algorithm. But by being transparent, they might avoid costly litigation as a result of having identified in clear and specific terms the way in which they have evaluated fairness and the extent to which they achieve it.

In addition to disclosing *which* fairness metric(s) they chose to optimize for during the algorithm design process, companies could also take the further step of explaining *how* they made that choice. This could include providing information on who (e.g., civil society organizations, representatives of the communities with a stake in the performance of the algorithm, etc.) provided input to the decision process, what priorities motivated the decision, and what plans are in place for periodically evaluating and updating their approach to fairness in the future if appropriate in light of newly acquired data.

algorithm itself. Transparency regarding the underlying algorithm, while generally providing vastly more information than algorithmic secrecy, could still leave questions unanswered. *See, e.g.*, Kroll et al., *supra* note 150, at 657-58 (noting in a section titled "Transparency and its Limits" that while "full or partial transparency can be a helpful tool for governance in many cases . . . transparency alone is not sufficient to provide accountability in all cases."); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 187 (2017) (writing that "[t]ransparency, however, does not automatically lead to accountability,"); Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 974 (2016) (writing that "transparency is an inadequate way to understand—much less govern—algorithms.").

[187] Lee Rainie & Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm* Age, PEW RSCH. CTR. (Feb. 8, 2017), http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age [https://perma.cc/2HRU-5ETY].

D. Which Fairness Metrics Should be Used?

The choice of which fairness metric(s) to use is clearly a fundamental question when assessing whether or not an algorithm is biased. One issue that can guide the choice is what can be termed the observability challenge, which in some contexts will limit the fairness metrics for which data can reliably be gathered. The observability challenge will not always be present. In the example presented in Tables 1 and 2 in Part II regarding predictions of which students in two different groups will pass a test, it was possible to observe the outcome associated with each prediction, and thereby to compute all of the fairness metrics discussed in Part II.

But there are a large number of scenarios in which a prediction is used as the basis to make a decision that renders some outcomes unobservable. Consider a prediction regarding loan repayment. A financial institution obviously wants to grant loans to borrowers who it expects to repay. After the loan is given, it is straightforward to tally true positives (people who were deemed likely to repay and did repay), false positives (people who were deemed likely to repay but defaulted), and the positive predictive value (the fraction of people who were deemed likely to repay who did repay).

But if an algorithm predicts that a loan applicant would default, the financial institution will generally deny the loan, thereby rendering it impossible to measure false negatives (people who were deemed likely to default but who if given the loan would have repaid it), true negatives (people who were deemed likely to default and who if given the loan would indeed have defaulted), and the metrics derived from that information (such as the negative predictive value). It would also be impossible to observe the true positive rate, which is a fraction with an observable numerator but unobservable denominator. The numerator is the number of true positives and the denominator is the sum of the true positives and the false negatives; i.e., the total number of people who would have repaid a loan, if given one. By analogy, the true negative rate would also be unobservable.

A similar issue arises in relation to decisions regarding granting bail. If an arrestee is deemed a flight risk and denied bail on that basis, it becomes impossible to evaluate whether he or she actually would have failed to appear at trial if bail had been granted. If an algorithm used in hiring predicts that a job applicant will be unable to perform the duties of the job, the company will likely not make the hire, rendering it possible to assess the accuracy of that negative evaluation. This observability problem will likely be relevant whenever a prediction is used to make a decision regarding an action (e.g., granting a loan) or intervention (e.g., denying bail), rather than to merely collect data and observe outcomes.[188]

The literature on algorithmic fairness tends not to give sufficient attention to the observability challenge, despite its obvious practical significance. After all, there is little utility in recommending use of a fairness metric that requires data that will not be available in practice. With that in mind, it is noteworthy that *predictive parity* and *statistical parity* are two metrics that do not suffer from the observability challenge.[189] As explained earlier, predictive parity is satisfied when, of

[188] There are methods that can make it possible to obtain information on outcomes that would normally be unobservable. For instance, consider false negatives in the context of loans. Normally, a lender will not grant a loan to an applicant who is predicted to default. This makes it impossible to measure false negatives, i.e., circumstances in which a borrower who was predicted to default did not do so. However, a lender wishing to measure false negative rates (and true negative rates) could choose to grant loans to a small number of people who, as predicted by an algorithm, were expected to default. In gathering this information, it would be important to select randomly from the pool of people who would otherwise have been denied loans (as opposed to, for example, choosing only those whose scores placed them just below the threshold for loan approval.) Over time, this would allow accumulation of statistics on false negatives. Of course, this approach has drawbacks: to accumulate a statistically significant number of observations could involve a large cost in unpaid loans. Also, there are contexts in which this approach to data gathering is not an option. Consider an algorithm used to predict whether a patient can safely be given a particular medication. A negative prediction corresponds to an expectation that the medication cannot be safely administered. It would obviously not be acceptable to gather statistics on false negative rates by nonetheless administering the medication for a subset of people for whom dangerous reactions were predicted. Additionally, relationships between input data and output data may change over time as a result of external factors (i.e., public policy changes). When this occurs, collecting data points over long periods of time without sufficient sensitivity to relevant external factors can result in inaccuracies in the resulting statistical observations.

[189] With respect to predictive parity, this statement assumes that the positive predictions are associated with observable outcomes. A "positive" prediction can signify that a person has been deemed creditworthy and given a loan, or that an arrestee poses no danger to society and

the people who are predicted to be in the positive class (e.g., loan applicants who are deemed creditworthy and given a loan), the percentage who are actually in the positive class (e.g., those who pay back the loan) is the same across the two groups being compared.[190]

For statistical parity, the computation is even simpler, since there is no need to examine outcomes.[191] Again using the loan example, statistical parity is met when members of the two groups are predicted to be in the positive class with equal rates (e.g., an equal percentage of both groups are deemed creditworthy and given a loan). It should be noted, though, that in the literature on algorithmic fairness, statistical parity has often been viewed as an inadequate and overly stringent measure of algorithmic fairness, as it fails to account for differences in base rates among groups that contribute to disparities in outcomes and often comes at the expense of accuracy.[192]

While predictive parity may be desirable in cases where other measures of fairness are largely unobservable, satisfying it does not alone settle all fairness concerns, as it reflects only one of many possible ways to measure fairness.[193] For example, achieving predictive parity provides incomplete information on error rates (false positive and false negatives

should be granted bail. Of course, it is also possible to invert the terminology, and, for example, associate a "positive" prediction with an expectation that an arrestee *does* pose a danger to society. In that case, he or she will be denied bail, and it will not be possible to observe whether the arrestee would have re-offended had bail been granted.

[190] *See supra* Section II.

[191] *See supra* Section II.

[192] Dwork et al., *supra* note 43, at 218 ("Although in some cases statistical parity appears to be desirable . . . we now argue its inadequacy as a notion of fairness, presenting three examples in which statistical parity is maintained, but from the point of view of an individual, the outcome is blatantly unfair."); Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8 INNOVATIONS THEORETICAL COMPUT. SCI. CONF., 43:1, 43:3 (2017) ("[C]lassification and risk assessment are much broader activities where statistical parity is often neither feasible nor desirable."); Matthew Joseph et al., Rawlsian Fairness for Machine Learning 3 (Oct. 29, 2016) (unpublished manuscript) ("[G]roup-level definitions often fail at both fairness and accurate learning. If two groups actually have different proportions of individuals who are able to pay back their loans, then the accuracy of any learning algorithm will obviously suffer when constrained to predict an equal proportion of paybacks for the two groups.").

[193] *See supra* Section II.

rates), which may be impossible to accurately measure.[194] When an algorithm is being used in a context where error rates are particularly important for assessing fairness, observability issues relating to errors are a cause for concern and serve as a motivation for more careful scrutiny of algorithms.[195]

Of course, the observability challenge exists in non-algorithmic decision-making as well, but it is especially important to keep in mind when decisions are presented as data-driven and objective, and instead may serve to reinforce or even amplify past biases. For this reason, to the extent possible given the limitations on observability, context-dependent costs of incorrect predictions should be taken into account when determining the suitability of algorithmic decision-making in a particular application. Additionally, as Mayson notes, non-algorithmic policy will need to play an important role in reducing the costs of false predictions wherever possible.[196] Such policy action was of course needed in various contexts before the introduction of algorithmic decision-making; however, the nature of prediction and its sometimes-unobservable outcomes helps to highlight this issue.

And what about scenarios in which there is not an observability challenge, i.e., when all the data necessary to compute any of the

[194] More precisely, if predictive parity is satisfied, that means that PPV is equal across both groups. Since PPV is defined as the ratio of the number of true positives to the sum of the number of true positives and the number of false positives (i.e., TP/(TP+FP)), knowing the PPV makes it possible to infer the *ratio* of true positives to false positives (i.e., PPV/(1-PPV)), but not their actual values. Similarly, this information alone would not permit computation of the false positive *rate*. And the PPV does not contain information about false negatives, nor of false negative rates. That said the sum of true positives, false negatives, true negatives, and false positives is equal to the total number of predictions, which will typically be known. In addition, the sum of true positives plus false negatives will in some circumstances be known (e.g., in the case of observing whether students pass a test), an in other circumstances be unknown (e.g., in loan approvals, false negatives will not be observable since the loan applicants who are predicted not to repay will not be given a loan). In a system known to satisfy predictive parity, to the extent that some of this additional information is also known, that could impose some constraints on error rates.

[195] *See supra* Section II.

[196] Mayson, *supra* note 5, at 2287 (noting that "a supportive, needs-oriented response to risk would mitigate the immediate racial impact of prediction. If a high-risk classification meant greater access to support and opportunities, a higher false-positive rate among black defendants would be less of a concern").

fairness measures discussed earlier are available? As discussed in Part II, the combination of mathematics and policy considerations means that it will often only be practical to optimize an algorithm for fairness according to one metric (or, for some particular metric combinations and subject to the associated mathematical constraints, two metrics). Which metric(s) should be chosen? We do not believe that there is any one metric that is inherently superior to all the others. Rather, the choice should be context-dependent.

Imagine a medical study in which an algorithm is used to predict whether or not a particular individual will develop dementia (in this example, this is the "positive" class). Treatment used to slow (but not stop) the development of dementia is administered based on the results of the algorithmic assessment.[197] Since all people who would have developed dementia absent treatment will still *eventually* develop it given enough time—though it will happen more slowly for those who received a positive prediction—all fairness metrics will be observable over time.[198] The people running the study will want to be sure that the predictive algorithm does not unjustly result in the disproportionate allocation of treatment to one particular group over another. Which fairness metric should be used?

If we imagine that preventive treatment for dementia has no negative health effects if unnecessarily administered (i.e., to a person who was not destined to develop dementia), the cost of a false negative prediction is higher than that of a false positive prediction. (This example assumes that there is a modest financial cost to the preventative treatment, so although there are no negative health effects, there is still an incentive not to simply administer it to everyone.) Moreover, ensuring accuracy of positive predictions is far less important than ensuring the accuracy of a negative prediction. The people conducting the study should therefore aim for a high negative predictive value, without concern for how it may affect positive predictive value. A lack of parity in false

[197] We have intentionally constructed this example using a treatment that delays but does not prevent dementia. If the treatment were to prevent it, there would be no way to observe false positives (people who had been incorrectly predicted to develop dementia).

[198] For simplicity, we assume in this example that all the people in the study remain alive over the period of time concerned.

negative rates or in negative predictive value should be taken as an indication that the algorithm-based decisions are leading to greater health costs to one group than another. Luckily, parity in false negative rates (equality of opportunity[199]) and parity in negative predictive value are achievable simultaneously, and so are arguably the fairness metrics of choice.[200] This particular metric combination is one in which policy considerations do *not* necessarily prevent satisfying more than one metric.[201]

To take another example, imagine the goal is to analyze the fairness of assessments made by fraud detection tools used to instantaneously attempt to detect fraud when a debit card transaction is attempted. Whenever suspected fraud is detected, the customer must respond to a text message inquiry from the bank to confirm that the transaction is legitimate. If the customer responds in the affirmative, the transaction is permitted to proceed. In addition, in order to collect all relevant statistics for the various fairness measures previously discussed, a small fraction of account holders are randomly required, at the end of a single randomly selected month, to confirm the accuracy of all transactions during that past month. They will receive a list of all transactions made using the debit card during that period and will have to indicate whether or not each was legitimate. This example assumes that a positive

[199] Equality of opportunity usually refers to parity in true positive rates, but since the false negative rate is equal to 1–TPR, equality of opportunity will automatically result in parity in false positive rates, as well.

[200] Garg, Villasenor & Foggo, *supra* note 74. It can be inferred by rewriting equations (6) and (7) in Garg *et al*. in terms of *negative* predictive value instead of *positive* predictive value, that if the constraint of equal FPRs is removed, then it is possible to simultaneously have equal TPRs (i.e., equality of opportunity) and equal negative predictive values. The requirement of having different FPRs to achieve this is not necessarily problematic in this particular scenario due to the low costs of false positives.

[201] Equality of opportunity (parity in true positive rates or, equivalently, in false negative rates) and parity in negative predictive value are achievable simultaneously, but with the constraint that the true negative rates (and therefore the false positive rates) would necessarily be unequal. *See id.* at 6-7 (in particular section 3.1.3). In the example provided in the text regarding the prediction of dementia and a treatment with no medical downsides that can slow its progress, having unequal false positive rates across two groups would not necessarily be problematic, particularly since as the example is constructed, there would be little financial motivation to deny treatment to someone who has been identified as possibly standing to benefit from it.

prediction means a prediction that an attempted transaction is fraudulent.

The presence of data gathered from this randomly selected subgroup ensures that it will be possible to gather information on false negatives (a false negative in this scenario corresponds to a fraudulent transaction that is not detected by the algorithm, i.e., the prediction is that the transaction is not fraudulent, but in fact it is fraudulent) that would otherwise be difficult to observe, as account holders may not notice smaller fraudulent purchases quickly, or even at all. Because this transaction-specific customer verification only is performed for a fraction of customers for limited periods of time, the resulting statistics will be gathered more slowly than if all customers were forced to provide confirmation of each individual purchase.

Nonetheless, over time, this process will allow determination of the true and false positives, true and false negatives, and all the other metrics (including the fairness measures discussed previously) derived from that information. What measure should we use to assess the fairness of the algorithm with respect to a protected attribute such as gender? In this case, it can be argued that a good metric is equalized odds (as a reminder, this means that both the true positive rate and, separately, the false positive rate are the same across groups).

To see why, consider that one would want the assessment tool to be able to identify a case of fraud whenever one is present, and to be able to do so regardless of an individual's gender. Consider further what can happen if equalized odds is not satisfied. If the false negative rate is higher for women than for men (meaning also that the true positive rate would be lower for women than for men) this means that fraud is less likely to be detected for female debit card holders than their male counterparts, creating a potential disadvantage for a group that has historically experienced discrimination in relation to financial services. Additionally, if the false positive rate is higher for women than for men (meaning that the true negative rate would be lower for women than for men) this means that the fraud detection tool places a disproportionate burden on women, who would need to respond to more text inquiries from the bank to confirm transactions as legitimate. Notably, though, the burden involved in a false negative assessment is greater than the cost of a false positive assessment, since a false negative means that a fraudulent transaction goes undetected; this is a higher burden than having to respond to a text message to confirm a transaction.

If there is a difference in the base rates of fraudulent charges on men's accounts versus women's, satisfying equalized odds precludes the possibility of also satisfying predictive parity. Suppose that women have higher base rates of being targeted by account fraud than men.[202] Satisfying equalized odds would result in a lower PPV for men than for women.[203] That is, more unnecessary text message confirmation requests (on a percentage basis) on male accounts than female accounts would be sent. In this particular case, the trade-off is arguably worthwhile given the greater harm done in failing to identify fraud when it is present than by requiring someone to confirm a transaction unnecessarily. Moreover, this is more desirable than using a prediction that often fails to identify fraud, or one that produces outcomes that place a further gender-biased burden in relation to financial services. Thus, the desirability of this outcome stems in part from the fact that the reduced accuracy that accompanies satisfying equalized odds will likely result in relatively *minor* disproportionate costs—the hassle of responding more text messages from the bank to confirm a transaction's legitimacy—for the traditionally advantaged group (men), and will avoid further propagating historical disadvantages for women.

Finally, and distinctly from the specifics of the examples above, it is important to note that there may be circumstances in which it could be of interest to choose intentionally *not* to optimize an algorithm for fairness according to any one measure or (under limited circumstances) pair of measures, choosing instead to engage in what amounts to a compromise. To take one possible example, as discussed in Part II, it is generally impossible to simultaneously satisfy both equalized odds and

[202] We have no evidence that—and are not suggesting that—women actually are more likely to be targeted by debit card fraud than men; we are simply constructing this hypothetical example to illustrate some of the reasons why certain fairness measures might be deemed more important in a particular context.
[203] *See* Garg, Villasenor & Foggo, *supra* note 74 (in particular, this result follows from subtracting equations (6) and (7) and setting the TPRs to be equal and the FPRs to be equal).

predictive parity, i.e., choosing to achieve fairness under one of these measures makes it impossible to achieve fairness under the other.[204] If told that both of these measures are important, an algorithm developer could elect to satisfy neither, but in doing so to get as close as possible to satisfying both. There are, of course, interesting mathematical questions regarding how effective a compromise could be made. Thus, approaches to fairness include not only the option to pick just one measure (or sometimes two) and forego all interest in the others, but also the option to attempt to compromise among multiple measures.

V. Conclusions

As algorithms become more widely deployed, technologists and legal practitioners alike will be increasingly engaged in working to ensure fairness and to evaluate accusations of algorithmic bias. This Article has aimed to facilitate that process by exploring various algorithmic fairness measures and their relative compatibilities in a broadly accessible manner. In addition, the Article has described existing discrimination law frameworks and explored some of the key questions that will arise in applying—and perhaps updating—those frameworks in light of the growth of algorithms.

We believe there are important opportunities to create greater awareness in legal and policy circles regarding the alternatives available to those who wish to assess whether an algorithm and the predictions it makes are "fair." In addition, it is important to raise awareness regarding the existence of mathematical constraints that govern whether and in what manner more than one fairness measure can be simultaneously satisfied. While the specifics of these constraints will likely be of interest to only a subset of practitioners, the fact that these constraints exist is something that should be more widely known and discussed than is the case today.

[204] More specifically, when the base rates across the two groups are unequal and when the prediction is imperfect, it is not possible to simultaneously satisfy both equalized odds and predictive parity.