RECONCILING LEGAL AND TECHNICAL APPROACHES TO ALGORITHMIC BIAS

ALICE XIANG*

---

* Senior Research Scientist (AI Ethics Lead), Sony AI, and formerly Head of Fairness, Transparency, and Accountability Research, Partnership on AI. I would like to thank Andrew Selbst, Jon Penney, Alexandra Chouldechova, Daniel E. Ho, Jessica Hwang, and Heather Wong for their immensely helpful comments and feedback. Thanks also to the audiences at WeRobot 2020, the NeurIPS Workshop on Fair AI in Finance, Stanford Law School, Seoul National University Law School, Simons Institute for the Theory of Computing, University of Waterloo, and Tsinghua University. I would also like to thank McKane Andrus for his excellent research assistance. Thanks to the editors at the *Tennessee Law Review* for their edits.

In recent years, there has been a proliferation of papers in the algorithmic fairness literature proposing various technical definitions of algorithmic bias and methods to mitigate bias. Whether these algorithmic bias mitigation methods would be permissible from a legal perspective is a complex but increasingly pressing question at a time when there are growing concerns about the potential for algorithmic decision-making to exacerbate societal inequities. In particular, there is a tension around the use of protected class variables: most algorithmic bias mitigation techniques utilize these variables or proxies, but anti-discrimination doctrine has a strong preference for decisions that are blind to them. This Article analyzes the extent to which technical approaches to algorithmic bias are compatible with U.S. anti-discrimination law and recommends a path toward greater compatibility.

This question is vital to address because a lack of legal compatibility creates the possibility that biased algorithms might be considered legally permissible while approaches designed to correct for bias might be considered illegally discriminatory. For example, a recent proposed rule from the Department of Housing and Urban Development ("HUD"), which would have established the first instance of a U.S. regulatory definition for algorithmic discrimination, would have created a safe harbor from disparate impact liability for housing-related algorithms that do not use protected class variables or close proxies. An abundance of recent scholarship has shown, however, that simply removing protected class variables and close proxies does little to ensure that the algorithm will not be biased. In fact, this approach, known as "fairness through unawareness" in the machine learning community, is widely considered naive. While the language around algorithms was removed in the final rule, this focus on the visibility of protected attributes in decision-making is central in U.S. anti-discrimination law.

Causal inference provides a potential way to reconcile algorithmic fairness techniques with anti-discrimination law. In U.S. law, discrimination is generally thought of as making decisions "because of" a protected class variable. In fact, in Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc., the case that motivated the HUD proposed rule, the Court required a "causal connection" between the decision-making process and the disproportionate outcomes. Instead of examining whether protected class variables appear in the algorithm, causal inference would allow for techniques that use protected class variables with the

intent of negating causal relationships in the data tied with race. While moving from correlation to causation is challenging—particularly in machine learning, where leveraging correlations to make accurate predictions is typically the goal—doing so offers a way to reconcile technical feasibility and legal precedence while providing protections against algorithmic bias.

INTRODUCTION

In recent years, there has been growing public awareness around the issue of biases in algorithms. Algorithms are increasingly being deployed to make or aid decisions in high-stakes contexts, like criminal justice,[1] child welfare,[2] employment,[3] healthcare,[4] and credit,[5] raising concerns about their potential flaws and unintended consequences.

On the policy side, this outcry has led policymakers to propose new rules and legislation.[6] The proposed 2019 Algorithmic Accountability Act, for example, would require algorithmic impact assessments, including an evaluation of potential biases, for algorithmic decision-making systems in high stakes contexts.[7] The

---

1. *See, e.g.*, Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803, 809 (2014); 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.

2. *See, e.g.*, Alexandra Chouldechova et al., *A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions*, *in* PROCEEDINGS OF MACHINE LEARNING RESEARCH OF THE 2018 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 3 (2018) (discussing an algorithmic system deployed to evaluate child abuse and neglect risk in Allegheny County).

3. *See, e.g.*, Jeffery Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G; *Pre-Employment Assessments*, HIREVUE, https://www.hirevue.com/products/assessments (last visited Feb. 25, 2020).

4. *See, e.g.*, Fei Jiang et al., *Artificial Intelligence in Healthcare: Past, Present and Future*, 2017 STROKE & VASCULAR NEUROLOGY, 230, 230.

5. *See, e.g.*, Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 148 (2017).

6. *See, e.g.*, Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019).

7. *Id.*

proposed Act, however, does not define algorithmic bias or specify permissible methods for addressing such bias. The closest a federal government entity has come to defining algorithmic bias was the Department of Housing and Urban Development's ("HUD") proposed rule on disparate impact.[8] The proposed rule, which sought to codify the Court's decision in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*,[9] would have created a safe harbor from disparate impact liability for algorithms that do not make use of protected class attributes or close proxies.[10] "Protected class variables" here refer to data on the protected class membership of individuals. Protected classes are demographic categories subject to special protections under anti-discrimination law.[11] The proposed HUD rule would have been the first U.S. regulation defining illegal bias or discrimination specifically in an algorithmic context, such that it might be regarded as an example for other statutory or regulatory efforts to address algorithmic bias. Although the final rule removed the safe harbor and specific language around algorithms, concluding that "it is premature at this time to more directly address algorithms," the approach of the proposed rule suggests that the presence or absence of protected class variables might be used as evidence for or against illegal algorithmic bias.[12]

---

8. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42854 (proposed Aug. 19, 2019).

9. 135 S. Ct. 2507 (2015).

10. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. at 42862 ("(c) *Failure to allege a prima facie case*. A defendant, or responding party, may establish that a plaintiff's allegations do not support a prima facie case of discriminatory effect under paragraph (b) of this section, if . . . (2) Where a plaintiff alleges that the cause of a discriminatory effect is a model used by the defendant, such as a risk assessment algorithm, and the defendant: (i) Provides the material factors that make up the inputs used in the challenged model and shows that these factors do not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act and that the model is predictive of credit risk or other similar valid objective . . . ."). Note that the stipulation that the model must be predictive of a valid objective is trivial without a specific threshold of accuracy.

11. For example, Title VII of the Civil Rights Act of 1964 provides protection against discrimination in employment on the basis of specific protected class attributes. 42 U.S.C. § 2000e-2(a) (2018).

12. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 85 Fed. Reg. 60288, 60290 (Sept. 24, 2020) (to be codified at 24 C.F.R. pt. 100).

In fact, concerns that the use of protected class variables might lead to discrimination liability have encouraged practitioners to avoid taking action to measure or mitigate algorithmic bias.[13] Outside of domains where demographic data collection is mandatory, algorithmic fairness practitioners rarely have access to demographic data to check for algorithmic bias, let alone to address bias concerns in their algorithms.[14] This has led to an unfortunate stalemate whereby calls by the public and policymakers to address algorithmic bias are ironically being stymied by concerns that taking action would increase rather than decrease legal liability.

In contrast, on the research side, an entire subfield of machine learning ("ML") called "algorithmic fairness" has emerged to address the issue of algorithmic bias[15] and has proposed many different methods for measuring and mitigating bias in algorithms that *require* the use of protected class variables or proxies.[16] The intuition

---

13. McKane Andrus et al., *"What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness*, *in* PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 249, 250 (2021).

14. *Id.*

15. *See, e.g.*, 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, AND DATA MINING 797, 798 (2017) ("Existing approaches to algorithmic fairness typically proceed in two steps. First, a formal criterion of fairness is defined; then, a decision rule is developed to satisfy that measure, either exactly or approximately.").

16. *See* Jon Kleinberg et al., *Algorithmic Fairness*, *in* 108 AEA PAPERS & PROCEEDINGS 22, 23 (2018) ("Using nationally representative data on college students, we underline how the inclusion of a protected variable—race in our application—not only improves predicted GPAs of admitted students (efficiency), but also can improve outcomes such as the fraction of admitted students who are black (equity)."). *See generally* Zachary C. Lipton et al., *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, *in* 31 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (2018) (analyzing whether preventing disparate impact requires disparate treatment in the algorithmic context). Most proposed definitions of algorithmic fairness seek improvements with the use of protected class variables. *See* Corbett-Davies et al., *supra* note 15, at 798–99 (discussing the use of protected class variables to satisfy statistical parity and predictive parity); Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, *in* 29 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 1, 4 (2016) (discussing the use of protected class variables for satisfying equal opportunity criterion); Muhammad Bilal Zafar et al., *Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment*, *in* PROCEEDINGS OF THE 26TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 1171, 1171–73 (2017) (discussing the use of protected class variables for addressing disparate mistreatment); *see also* Michael Feldman et al., *Certifying and Removing Disparate Impact*, *in* PROCEEDINGS OF THE 21ST ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE, DISCOVERY, AND

behind this necessity is that measuring how the performance of an algorithm differs across protected class groups requires knowing to which groups different individuals belong. Similarly, actively correcting for differences in how the algorithm treats different groups or correcting for disproportionate outcomes across groups requires taking into account the group membership of individuals. Due in part to concerns about the legality of using protected class variables in algorithmic decision-making, however, methods proposed by the ML literature are not being widely used in practice.[17]

As this Article will illustrate, there are strong technical motivations for using protected class variables to mitigate algorithmic bias, but efforts to use protected class variables evoke the long-standing tension between anti-classification and anti-subordination principles in anti-discrimination law. Anti-classification is the principle that classification or treatment that differs based on protected class attributes is discriminatory.[18] This principle is often interpreted as prohibiting decision-making that is

---

DATA MINING 259, 259–61 (2015) (discussing the use of protected class variables to satisfy the 80% rule adopted by the U.S. Equal Employment Opportunity Commission).

17. *See* Andrus et al., *supra* note 13, at 253. In fact, even simply accessing the protected class variables themselves is a prohibitive first step. *See* Kenneth Holstein et al., *Paper 600: Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, *in* PROCEEDINGS OF THE 2019 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 1, 8 (2019) ("Although most auditing methods in the fair ML literature assume access to sensitive demographics (such as gender or race) at an individual level, many of our interviewees reported that their teams are only able to collect such information at coarser levels, if at all. For example, companies working with K-12 student populations in the US are typically prohibited from collecting such demographics by school or district policies and FERPA laws. A majority (70%) of survey respondents, out of the 69% who were asked the question, indicated that the availability of tools to support fairness auditing without access to demographics at an individual level would be at least 'Very' useful."); *see also* Miranda Bogen et al., *Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination*, *in* PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 492, 497 (2019) (describing how companies, especially in highly regulated industries, are unlikely to collect data on sensitive attributes for antidiscrimination efforts due to concerns about exposing themselves to liability).

18. Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 10 (2003) ("Roughly speaking, this principle holds that the government may not classify people either overtly or surreptitiously on the basis of a forbidden category: for example, their race.").

conscious of protected classes.[19] Anti-subordination, on the other hand, is the principle that the law should seek to dismantle hierarchies between protected class groups, even if doing so involves some degree of consciousness of these group classifications.[20]

This Article further illustrates how adopting causal frameworks for evaluating potential algorithmic discrimination can help to reconcile these legal and technical approaches to mitigating algorithmic bias. Already in anti-discrimination law, the concept of causality is key to determining whether a decision-making process is discriminatory. In fact, *Texas v. Inclusive Communities*, the case that the proposed HUD rule discussed above sought to reflect, requires "robust causality" between the disproportionate outcomes and the decision-making rule or policy in order to establish liability.[21] Moreover, anti-discrimination statutes define discrimination in causal terms, as disparities or decisions made "because of" a protected class attribute.[22] From a technical perspective, a causal framing would make it possible to distinguish between different sources of bias and different uses of protected class variables where the effect is to exacerbate differences between how different groups are treated versus ameliorate such differences. One of the contributions of this Article is thus to show that not only is causality a potential lens through which to assess the fairness of an algorithm, but also it is vital to achieve legal compatibility.

While some biases may be introduced intentionally—past papers have identified, for instance, the ways in which discriminatory actors might be able to "mask" their discriminatory behavior using

---

19. *Id.* Though, Balkin and Siegel make the argument that the issues addressed using the anti-classification principle in court often go through a sieve of anti-subordination understanding before arriving in court. *See* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* 2 (Stan. Univ., Working Paper, 2018), https://arxiv.org/abs/1808.00023 ("*[A]nti-classification[]* stipulates that risk assessment algorithms not consider protected characteristics—like race, gender, or their proxies—when deriving estimates." (footnote omitted)).

20. Balkin & Siegel, *supra* note 18, at 9.

21. Texas Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2523 (2015).

22. *See, e.g.*, Age Discrimination in Employment Act of 1967, 29 U.S.C. § 623(a) (2018); Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2018). According to a recent Eighth Circuit case, under Title VII, "[t]o prevail on a hostile work environment claim," a plaintiff must prove five elements. Hales v. Casey's Mktg. Co., 886 F.3d 730, 735 (8th Cir. 2018). One of these elements requires the plaintiff to demonstrate "that a causal nexus existed between the harassment and protected group status." *Id.*

algorithms[23]—this Article focuses on the unintentional algorithmic biases that result from historical discriminatory trends reflected in the training data rather than discriminatory intent on the part of the algorithm's developers or deployers. The key issue this Article tackles is the extent to which well-meaning algorithm developers can use protected class variables to address algorithmic bias in light of existing anti-discrimination law jurisprudence. The key risk that this Article seeks to mitigate is the possibility that technical and legal approaches to mitigating bias will diverge so much that laws prohibiting algorithmic bias will fail in practice to weed out biased algorithms, while technical methods designed to address algorithmic biased will be deemed illegally discriminatory.

This Article is situated within the broader legal academic discourse around how anti-discrimination law might apply to the algorithmic context. Although prior work has surveyed the extent to which U.S. law permits the collection of sensitive attributes that would facilitate the measurement of algorithmic bias,[24] there remains the issue of what can be done from both a legal and technical perspective to mitigate bias. Past literature has noted that there are challenges with reconciling disparate treatment and disparate impact jurisprudence in the context of mitigating bias in algorithms[25] and has examined equal protection implications through the statutory lens of Title VII,[26] and scholars have touched briefly on the constitutional issues at play.[27]

The contributions of this Article are: (1) to illustrate the key role of protected class variables in mitigating algorithmic bias from a technical perspective; (2) to assess potential concerns from anti-discrimination law doctrine with the use of protected class variables

---

23. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 692 (2016).

24. *See generally* Bogen et al., *supra* note 17 (discussing how and when companies collect sensitive attribute data for antidiscrimination purposes).

25. *See* Barocas & Selbst, *supra* note 23, at 672; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 642 (2017); Lipton et al.., *supra* note 16, at 1–19.

26. *See* Barocas & Selbst, *supra* note 23, at 729–31; Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J., 803, 809 (2020); Zach Harned & Hanna Wallach, *Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer*, 47 FLA. ST. U. L. REV. 617, 617 (2020).

27. *See* sources cited *supra* notes 25–26. The closest a paper has come to evaluating these constitutional issues is examining anti-classification in the context of criminal justice risk assessment tools. *See* Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1094–1101 (2019).

for bias mitigation; and (3) to propose causality as a unifying concept between legal and technical approaches to addressing bias that would enable the legal use of protected class variables in bias mitigation techniques. Moreover, this Article illustrates how the mathematization that comes with addressing algorithmic bias highlights inconsistencies in existing anti-discrimination jurisprudence, which litigation regarding algorithmic bias might force courts to grapple with more directly.

Part I provides definitions of "algorithmic fairness" and "algorithmic bias" and discusses examples of algorithmic bias that have prompted public outcry over this issue and motivated efforts to mitigate algorithmic bias. Part II discusses why protected class variables or proxies are necessary from a technical perspective in order to mitigate algorithmic bias. Part III situates this tension between consciousness and blindness of protected class variables in algorithmic decision-making within the broader tension in anti-discrimination law between anti-classification and anti-subordination. Although anti-classification has become the dominant framework embraced by the Court in recent years, this Article discusses how this approach could inadvertently preclude most technical methods to address algorithmic bias. Part IV discusses one of the core underlying issues driving this tension—the way the Court determines responsibility for historical discrimination that would enable race-conscious remedial action. Part V discusses relevant jurisprudence governing race-conscious remedial government action, focusing on affirmative action jurisprudence. Part VI addresses relevant jurisprudence that would also be applicable to private sector actors, focusing on disparate impact and disparate treatment doctrines. Part VII proposes causal inference as a potential resolution to tensions between efforts to mitigate algorithmic bias and existing anti-discrimination jurisprudence.

I. BACKGROUND ON ALGORITHMIC BIAS

A. Algorithmic Fairness, Bias, and Decision-making

For the purposes of this Article, I define "algorithmic fairness" as the literature that explores the technical, legal, and ethical concerns with algorithmic decision-making, in particular issues of algorithmic bias. Throughout this Article, "algorithm" will specifically refer to

ML algorithms (i.e., statistical models trained on data).[28] The definition of "algorithmic bias" is a hotly contested topic in the ML literature, so here I will define it broadly as when ML algorithms systematically perform less well for or penalize certain subgroups. While "algorithmic bias" often encompasses disparities along attributes that are not protected classes, such as socioeconomic class or education level, this Article will focus on protected class attributes and their proxies due to the legal concerns around using those variables in decision-making. While most of the examples in this Article will use race[29] or sex[30] as the protected class variable, national origin,[31] sexual orientation,[32] age,[33] disability,[34] and other attributes[35] are also protected classes. In addition, this Article will focus on biases that stem from the data used to train the model, including both features and outcome variables; prior legal literature has examined biases related to model design decisions.[36]

---

28. "Algorithm" more generally refers to "a process or set of rules to be followed in calculations or other problem-solving operations, esp. by a computer." *Algorithm*, NEW OXFORD AMERICAN DICTIONARY (3d ed. 2010). ML "refers to the automated detection of meaningful patterns in data." SHAI SHALEV-SHWARTZ & SHAI BEN-DAVID, UNDERSTANDING MACHINE LEARNING: FROM THEORY TO ALGORITHMS xv (2014). For this Article, I choose a narrower definition of "algorithm," because algorithmic bias is primarily relevant for ML algorithms given the close relationship between algorithmic bias and the biases reflected in the training data and model development process.

29. *See* Brown v. Bd. of Educ., 347 U.S. 483, 495 (1954) (holding that racial segregation in public education violates the equal protection guarantee of the Fourteenth Amendment).

30. *See* Cannon v. Univ. of Chi., 441 U.S. 677, 709 (1979) (holding that Title IX provides a private right of action for victims of discrimination on the basis of sex).

31. *See* DiCarlo v. Potter, 358 F.3d 408, 415–16 (6th Cir. 2004) (noting, in the Title VII claim analysis, derogatory comments the plaintiff's supervisor made about the plaintiff's "Italian-American heritage").

32. *See generally* EQUAL EMP. OPPORTUNITY COMM'N, WHAT YOU SHOULD KNOW: THE EEOC AND PROTECTIONS FOR LGBT WORKERS (2020).

33. *See* Age Discrimination in Employment Act, 29 U.S.C. § 623(a) (2018).

34. *See* Americans with Disabilities Act of 1990, 42 U.S.C. § 12101 (2018).

35. *See, e.g.*, Vietnam Era Veterans' Readjustment Assistance Act of 1974, 38 U.S.C. § 4212 (2018) (prohibiting discrimination on the basis of veteran status); Genetic Information Nondiscrimination Act of 2008, Pub. L. No. 110–233, 122 Stat. 881 (prohibiting discrimination on the basis of genetic information); Pregnancy Discrimination Act of 1978, Pub. L. No. 95–555, 92 Stat. 2076 (prohibiting discrimination on the basis of pregnancy status); Civil Rights Act of 1968, Pub. L. No. 90-284, 82 Stat. 73 (prohibiting discrimination on the basis of familial status).

36. *See, e.g.*, Barocas & Selbst, *supra* note 23, at 699.

"Algorithmic decision-making" refers to decisions made using an ML or other statistical model trained on data.[37] These decisions can be completely automated or used to inform human decision-makers. A resume filtering algorithm would be an example of the former because it automates initial employment screening decisions; a recidivism risk assessment algorithm would be an example of the latter because it provides recommendations to judges for detention and parole decisions.

While algorithmic bias can refer to a wide variety of harms, for the purposes of this Article, I focus on bias associated with allocative harms. Allocative harms are those that result from a resource being unfairly distributed among individuals of different demographic groups.[38] For example, in the employment context, an example of an allocative harm would be not receiving a job or promotion. In the criminal justice context, an example of an allocative harm would be not being released in the pretrial context. Allocative harms are most relevant in an anti-discrimination law context, where standing requires a clearly identifiable harm to an individual.[39] Other harms related to algorithmic bias, like representational harm, whereby certain groups are represented in less flattering or proportionate ways, are less relevant to this context. For example, a commonly cited example of a representational harm is that Google image results for "CEO" previously disproportionately showed images of men rather than women.[40] While this kind of harm raises important questions about the role of algorithms in shaping people's stereotypes, this type of harm would not generally be legally actionable.

While algorithmic decision-making has many potential benefits, the algorithmic fairness literature tends to examine the ways in which algorithms can have negative side effects, specifically on

---

37. In practice, these "decisions" are often predictions about what might happen to an individual (e.g., probability of recidivism) or classifications of an individual into different groups (e.g., high or low risk for recidivism).

38. *See generally* SOLON BAROCAS ET AL., *Introduction*, *in* FAIRNESS IN MACHINE LEARNING (2019) (explaining allocative harms).

39. Standing in federal courts requires the plaintiff's "injury be concrete, particularized, and actual or imminent; fairly traceable to the challenged action; and redressable by a favorable ruling." Monsanto Co. v. Geertson Seed Farms, 561 U.S. 139, 149 (2010) (citing Horne v. Flores, 557 U.S. 433, 445 (2009)).

40. Only 11% of image results featured women even though 27% of CEOs at the time were women. Jennifer Langston, *Who's a CEO? Google Image Results Can Shift Gender Biases*, UW NEWS (Apr. 9, 2015), https://www.washington.edu/news/2015/04/09/whos-a-ceo-google-image-results-can-shift-gender-biases/.

demographic subpopulations.[41] Nonetheless, to begin this discussion about algorithmic bias, it is important to understand some of the benefits of algorithmic decision-making because these drive the adoption of algorithmic tools in the first place. First, algorithms are inherently data-driven and can take into account a large number of factors about individuals,[42] whereas humans might only be able to consider several distinct pieces of information simultaneously.[43] Second, automation allows decisions to be made much more quickly and scalably, thus potentially reducing costs in the long run.[44] Third, algorithmic decision-making provides consistency. While studies have raised concerns that human judges might make different decisions depending on idiosyncratic factors like how long it has been since they last ate,[45] such concerns are not applicable to

---

41. *See* Anne L. Washington, *How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate*, 17 COLO. TECH. L.J. 131, 150–51 (2019); *see, e.g.*, sources cited *infra* note 47; *see also* Chelsea Barabas et al., *Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, *in* PROCEEDINGS OF THE 2018 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 2–3 (2018); Ben Hutchinson & Margaret Mitchell, *50 Years of Test (Un)fairness: Lessons for Machine Learning*, *in* PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 49, 49–54 (2019).

42. *See, e.g.*, RHEMA VAITHIANATHAN ET AL., DEVELOPING PREDICTIVE MODELS TO SUPPORT CHILD MALTREATMENT HOTLINE SCREENING DECISIONS: ALLEGHENY COUNTY METHODOLOGY AND IMPLEMENTATION 31 (2017) (over 800 variables were able to be used per individual in developing a child welfare risk prediction algorithm).

43. Miller's Law is a famous example of this principle and asserts that humans can only hold seven, plus or minus two, objects of information simultaneously in short-term memory. George A. Miller, *The Magical Number Seven, Plus or Minus Two Some Limits on Our Capacity for Processing Information,* 63 PSYCH. REV. 81, 91 (1956).

44. *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY 39–84, 127–74 (2018) (providing case studies of an algorithm for determining welfare eligibility in Indiana and an algorithm used to predict child risk scores in Allegheny county, both of which were intended to cut costs and staffing needs).

45. *See* Shai Danziger et al., *Extraneous Factors in Judicial Decisions*, *in* 108 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 6889, 6889 (2011) ("Our findings suggest that judicial rulings can be swayed by extraneous variables that should have no bearing on legal decisions."). This famous study analyzed parole decisions made by Israeli judges and found that "the percentage of favorable rulings drops gradually from ≈65% to nearly zero within each decision session and returns abruptly to ≈65% after a break." *Id. But see* Andreas Glöckner, *The Irrational Hungry Judge Effect Revisited: Simulations Reveal that the Magnitude of the Effect is Overestimated*, 11 JUDGMENT & DECISION MAKING 601, 601–08 (2016) (finding that the disparate outcomes could be simulated by judges that simply try to finish shorter, less-promising cases before breaks); K. Weinshall-Margel & J. Shapard, *Overlooked Factors in the Analysis of Parole Decisions*, *in* 108 PROCEEDINGS OF THE

algorithms. Finally, there is the potential for algorithms to centralize decision-making, which can make auditing decisions easier, as a single algorithm has the potential to play the role of hundreds or thousands of human decision-makers.

That said, some of these benefits are double-edged swords. For example, the fact that algorithms can be deployed to automate a large number of decisions implies that even subtle biases can create large systematic effects, further entrenching and perpetuating inequality.[46] Moreover, while algorithms are often seen as more objective and evidence-driven than humans, this perception can lead to the phenomenon of "automation bias," whereby humans give excessive weight to algorithmic decisions and ignore contrary information.[47]

Algorithms can also obfuscate decision-making processes. They require expertise and access to information to understand, but

---

NATIONAL ACADEMY OF SCIENCES E833, E833 (2011) (suggesting that the results in the Danziger study are a result of case order presentation, with prisoners without legal representation having lower success rates and also being considered right before breaks).

46. In addition, having a single algorithm replace many independent decision-making processes means that any limitations (including not-so-subtle ones) in the algorithm will be amplified. For example, if a risk assessment algorithm has a cut-off point of 10, whereby people with scores ≤ 10 are denoted low risk and people with scores > 10 are denoted high risk, then people with scores of 10.5 will always have much worse categorizations than those with scores of 10 even though the two groups have very similar risk profiles. In the case of human decision-making, or decentralized decision-making more generally, there are unlikely to be such sharp cut-offs.

47. Automation bias occurs when humans ascribe excessive value to automated decisions or predictions, often ignoring contradictory information. This in turn limits the extent to which having a "human in the loop" can actually improve algorithmic decisions. *See, e.g.*, Linda J. Skitka et al., *Accountability and Automation Bias*, 52 INT'L J. HUM.-COMPUT. STUD. 701, 701–04 (2000); *see also* Mary L. Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, *in* AIAA 3RD INTELLIGENT SYSTEMS CONFERENCE 1, 1–5 (2004). Even before AI became a hot button issue, this issue was studied in the context of automated systems in aviation and healthcare. *See, e.g.*, Kate Goddard et al., *Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators*, 19 J. AM. MED. INFORM. ASSOC. 121, 121 (2012). Automation bias is generally explained as being the result of the human tendency to seek the paths of least cognitive exertion, especially when faced with tight time constraints, and the tendency of humans to think too highly of the expertise of automated systems. There are also some studies, however, that suggest that in practice inertia or "foot-dragging" might limit the adoption of technological systems, such that they might not have as much influence on decision-making processes. *See, e.g.*, Angèle Christin, *Algorithms in Practice: Comparing Web Journalism and Criminal Justice*, 4 BIG DATA & SOC'Y 1, 1–3 (2017).

people affected by their decisions are unlikely to have much information about the data or methods used to train the algorithm,[48] given that such information is typically the proprietary intellectual property of the entity developing the tool.[49] Even if such information were open-sourced by the entity or made available during the discovery process, it is challenging to use such information at face value to evaluate whether there are problems with the algorithm.[50] In fact, even the developers of algorithms often regard them as "black boxes."[51] The subfield of explainable ML, which seeks to develop techniques for explaining how algorithmic predictions or decisions were made, is still nascent, and the techniques that are available are rarely deployed in practice to provide more transparency to end users.[52] As a result, it is less likely that problems with algorithms will be detected by end users or those affected by algorithmic decisions, making such problems potentially more insidious. And even when problems are detected, it is often unclear who in the network of actors behind the system should be liable.[53]

---

48. *See generally* EUBANKS, *supra* note 44 (investigating and analyzing the impact of data-based technology on America's poor and working-class people).

49. Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1, 3–4 (2016).

50. Some authors have made the point that with appropriate transparency measures in place, algorithms be more transparent than human decision-making. Such transparency gains, however, will only be realized if there are policy changes to make the components of an algorithm available and open to inspection. Indeed, the authors caution that "without the appropriate safeguards, the prospects for detecting discrimination in a world of unregulated algorithm design could become even more serious than they currently are." *See* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 114 (2018).

51. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 106–07 (2015); Zachary C. Lipton, *The Mythos of Model Interpretability*, 2018 ACM QUEUE 1, 6.

52. *See, e.g.*, Umang Bhatt & Alice Xiang et al., *Explainable Machine Learning in Deployment*, *in* PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 648, 648 (2020). *See generally* Tim Miller et al., *Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*, *in* PROCEEDINGS OF THE 17TH ANNUAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING (2017) (explaining that attempts are being made to explain algorithmic methods, but they only open the black box for a small set of experts due to narrow understandings of explainability).

53. *See generally* Madeleine Clare Elish, *Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction*, 5 ENGAGING SCI., TECH., & SOC'Y 40, 55 (2019) ("This article has proposed the concept of a moral crumple zone as a provocation to

It is worth noting, however, that algorithmic tools are not a new technology. Credit scores, which use data to predict a person's likelihood of paying their debts, have been used in the United States since the 1960s.[54] What is distinct, however, is the extent to which algorithms have proliferated, increasingly affecting a wide variety of important decisions. Although algorithmic decision-making was popularized under the assumption that algorithms would be more objective than human decision-makers, there is a growing realization that they might suffer from many of the same biases, given that they are often trained on past decisions or determinations made by humans.[55]

B. Examples of Algorithmic Bias

Given the difficulty of detecting algorithmic bias as an end user, the literature on algorithmic fairness has been inspired by a few high-profile examples where algorithmic bias was unearthed. The initial explosion in algorithmic fairness papers came after ProPublica reported in 2016 that the COMPAS risk assessment tool, one of the most popular algorithmic tools used by court systems across the country, was biased against black defendants.[56]

---

rethink how, why, and with what implications responsibility will be assigned when automated, autonomous, or 'intelligent' systems fail.").

54. Noel Capon, *Credit Scoring Systems: A Critical Analysis*, 46 J. MKTG. 82, 84 (1982) ("Since the early 1960s the use of credit scoring systems has expanded enormously, as journals serving practitioners have been filled with articles extolling their virtues. Further, passage of the Equal Credit Opportunity Act Amendments offered further endorsement of credit scoring systems when instructions regarding their use were specifically included in Regulation B, which implements the Act." (citations omitted)).

55. *See, e.g.*, Barocas & Selbst, *supra* note 23, at 671–76. *See generally* "RAW DATA" IS AN OXYMORON (Lisa Gitelman ed., 2013*)* (advancing an understanding that data is anything but a raw, unfiltered view of the world and that systems built upon data will reflect the biases of that data); Nathan Kallus & Angela Zhou, *Residual Unfairness in Fair Machine Learning from Prejudiced Data*, *in* PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (2018) ("We study how prejudicial biases in a dataset can lead to *residual unfairness*, which persists even after fairness adjustment if error parity metrics assessed from the censored dataset are used. We show that the residual unfairness that remains even after adjustment will disadvantage the same group that was prejudiced against before, in the training data. This proves that even after fairness adjustment, fair machine learning still has a 'bias in, bias out' property.").

56. Notably ProPublica was only able to assemble its dataset through extensive use of the Freedom of Information Act, further illustrating the challenges of

ProPublica illustrated that COMPAS' false positive rates—the rate at which defendants who did not recidivate were incorrectly predicted to be at high risk for recidivism—were roughly twice as high for black defendants as for white defendants.[57] Equivant, the company behind COMPAS, countered that its tool had predictive parity for white and black defendants, meaning that among those classified as high risk, the recidivism rate was similar between white and black defendants (positive predictive parity), and among those classified as low risk, the recidivism rate was similar between white and black defendants (negative predictive parity).[58] ML scholars took note and soon proved an impossibility theorem showing that when different demographic groups have different baselines in the underlying data (in this case, different rates of re-arrest), and the model is not a perfect predictor,[59] it is impossible for a well-calibrated score[60] to have equal average scores for both the positive and negative classes.[61] In the case of a binary risk assessment classification, this means that it is impossible to simultaneously have predictive parity and equalized false positive and false negative rates across black and white defendants. Given that arrest data show highly disparate rates of re-arrest for black and white defendants, well-calibrated tools will generally have different false

---

demonstrating algorithmic bias. *See* Washington, *supra* note 41, at 148–60; Angwin et al., *supra* note 1.

57. *See* Washington, *supra* note 41, at 148–60; Angwin et al., *supra* note 1.

58. WILLIAM DIETERICH ET AL., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 9–19 (2016) (demonstrating the predictive parity of the COMPAS tool).

59. A perfect predictor is a model that can perfectly make predictions. For example, in the risk assessment context, this would mean a model that can perfectly predict whether an individual will recidivate. In practice, this is impossible.

60. Calibration implies that among those with a predicted probability of recidivism of X%, X% actually do indeed recidivate.

61. Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, *in* PROCEEDINGS OF THE 8TH CONFERENCE ON INNOVATIONS IN THEORETICAL COMPUTER SCIENCE 1, 3 (2016) ("Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal—that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously. Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to approximate versions of the conditions as well.").

positive or false negative rates for black and white defendants.[62] This counterintuitive result sparked significant interest in the technical community around developing new definitions for algorithmic bias, which are often then applied to the ProPublica COMPAS dataset.[63]

In another high-profile example, Amazon revealed in 2018 that it had scrapped efforts to build a resume filter algorithm after it found that the algorithm was biased against women.[64] The training data for the algorithm consisted of past resumes that Amazon had received, tagged based on whether the individual was ultimately hired, so the algorithm could learn to distinguish between "successful" vs. "unsuccessful" candidates.[65] Because female applicants historically had a lower success rate, however, the algorithm learned that features associated with being a woman were negative signals for hiring success.[66] The algorithm learned, for example, to penalize the graduates of women's colleges, and while "chess club" was a positive signal, "women's chess club" was a negative one.[67]

More recently, New York insurance regulators launched an investigation into an algorithm deployed by Optum,[68] part of UnitedHealth Group.[69] A study had found that the algorithm, which

---

62. *See, e.g.*, Geoff Pleiss et al., *On Fairness and Calibration*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 1–2 (2017).

63. *E.g.*, Corbett-Davies et al., *supra* note 15, at 798–804; Corbett-Davies & Goel, *supra* note 19, at 3. *See generally* Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017) (discussing several emerging fairness criteria in recidivism prediction instruments); James R. Foulds et al., *An Intersectional Definition of Fairness*, *in* PROCEEDINGS OF THE 36TH IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (2018) (using case studies on census data and the COMPAS criminal recidivism dataset to demonstrate the utility of new definitions of fairness in ML and providing a learning algorithm attuned to the researchers' intersectional fairness criteria).

64. Dastin, *supra* note 3.

65. *Id.*

66. *Id.*

67. *Id.*

68. Melanie Evans & Anna Wilde Mathews, *New York Regulator Probes UnitedHealth Algorithm for Racial Bias*, WALL ST. J. (Oct. 26, 2019, 7:00 AM), https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601.

69. UnitedHealth Group is the largest healthcare company in the world by revenue. Sterling Price, *Largest Health Insurance Companies of 2021*,

assigned risk scores to patients reflecting their predicted health level, was biased against black patients: black patients with a given risk score were in fact less healthy than white patients with the same score.[70] The authors of the study attributed the bias to the fact that the algorithm was actually predicting healthcare costs as a proxy for health, and "[l]ess money is spent on Black patients who have the same level of need, [so] the algorithm thus falsely concludes that Black patients are healthier than equally sick white patients."[71]

As these examples illustrate, algorithmic bias has increasingly become a concern in many high-stakes contexts, like criminal justice, employment, and healthcare. While examples like these have motivated companies, jurisdictions, and regulators to start examining the issue of algorithmic bias, it remains challenging to address in practice, as the following Parts will discuss.

II. THE TECHNICAL NECESSITY OF USING PROTECTED CLASS ATTRIBUTES OR PROXIES

There are many technical challenges that complicate efforts to address algorithmic bias in practice, but for the purpose of this Article, I focus specifically on ones related to the use of protected class variables because these challenges are most related to legal compatibility issues. The Sections below will connect the ML literature to legal concepts in discussing why removing protected class variables or close proxies does little to mitigate bias and why using protected class variables is vital for actively mitigating bias.

A. Proxy Variables, Omitted-variable Bias, and the Rashomon Effect

One of the most intuitive and straightforward ways to think about approaching algorithmic fairness is to simply exclude protected class variables or close proxies from the model's training data. This approach is known in the ML literature as "fairness through unawareness"[72] and is analogous to legal concepts of

---

VALUEPENGUIN, https://www.valuepenguin.com/largest-health-insurance-companies (last updated May 24, 2021).

70. *See* Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447, 447 (2019).

71. *Id.* at 454.

72. Hardt et al., *supra* note 16, at 1. The name comes from one of the seminal algorithmic fairness articles: *Fairness Through Awareness*. *See* Cynthia Dwork et al.,

blindness or neutrality toward the protected class. In fact, the Equal Credit Reporting Act prohibits creditors from considering information about protected attributes in any aspect of a credit transaction.[73] This was also the strategy that was suggested by HUD in its proposed rule-making[74] implementing the disparate impact standard articulated in *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc.*[75] While the HUD proposed rule did not provide a specific definition for "close proxy," in general, a proxy variable is one that is correlated with the variable of interest (in this case, the protected class variable).[76] In the United States, for example, zip codes and neighborhoods are famously proxies for race and have been used for redlining.[77] Due to this correlation, proxies can be seen as standing in for the protected class variable, providing much of the same information. A "close" proxy is presumably one that has a particularly high correlation with the protected class variable (though how high was not specified in the HUD rule).

The fundamental problem with "fairness through unawareness" is that eliminating the use of protected class variables and close proxies does not necessarily reduce algorithmic bias and can, ironically, exacerbate the issue. Removing protected attributes and close proxies will do little if there are still sufficient weak proxies in the data: combining many variables that are weakly correlated with a protected classification can yield strong predictions for that protected classification.[78] Although this is, in theory, an issue with

---

Fairness Through Awareness, *in* PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTING SCIENCE CONFERENCE 214, 214–25 (2012).

73. Rules Concerning Evaluation of Applications, 12 C.F.R. § 1002.6(b)(9) (2021).

74. HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42,854, 42,857–58 (Aug. 19, 2019) (to be codified at 24 C.F.R. pt. 100).

75. 135 S. Ct. 2507 (2015).

76. *See* HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. at 42857–58. The issue of proxy variables is also known in the algorithmic fairness literature as redundant encodings. Lipton et al., *supra* note 16, at 8.

77. *See* Frank Rene Lopez, *Using the Fair Housing Act to Combat Predatory Lending*, 6 GEO. J. POVERTY L. & POL'Y 73, 75 (1999) ("In the 1960's, banks began engaging in a practice known as redlining, circling minority and low-income communities with red ink to indicate areas of the map upon which a general denial of credit would be enforced." (citations omitted)).

78. Anupam Datta et al., *Proxy Discrimination in Data-Driven Systems*, *in* PROCEEDINGS OF THE 2017 ACM SIGSAC CONFERENCE ON COMPUTER AND

any decision-making process, part of what distinguishes algorithmic decision-making from human decision-making is the sheer amount of data being used. Algorithms can process a far larger number of variables than humans can, increasing the number of potential proxies.

Proxy variables influence algorithmic decision-making through omitted-variable bias.[79] Omitted-variable bias occurs when not all relevant variables are present in a statistical model.[80] Due to the omissions, the weights on the included variables reflect not only the effects of those variables but also the effects of any excluded variables that are correlated with them, giving proxies an outsized influence on the decision-making process.[81] This in turn can lead not only to biased decision-making but also to improper conclusions or inferences based on the algorithm's output.[82]

Take, for example, an algorithm that predicts mortality in order to inform doctors or insurance providers whether it would be more appropriate to refer a patient to treatment or hospice.[83] If the algorithm were trained on data where women had a much lower mortality rate than men, and the algorithm did not have access to

---

COMMUNICATIONS SECURITY 1193, 1194 (2017); *see also* Bin Bi et al., *Inferring the Demographics of Search Users: Social Data Meets Search Queries*, *in* PROCEEDINGS OF THE 22ND INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 131, 133–39 (2013) (developing a technique for inferring demographic traits from Facebook likes and search queries); Till Speicher et al., *Potential for Discrimination in Online Targeted Advertising*, *in* PROCEEDINGS OF THE 2018 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 2–14 (2018) (identifying multiple techniques for leveraging proxies available in Facebook data to target members of protected classes). *See generally* SOLON BAROCAS ET AL., *Classification*, *in* FAIRNESS IN MACHINE LEARNING, *supra* note 38 (explaining how a large number of weak proxies can approximate the protected classification).

79. Jongbin Jung et al., *Omitted and Included Variable Bias in Tests for Disparate Impact*, *in* PROCEEDINGS OF THE 22ND INTERNATIONAL CONFERENCE ON WORLD WIDE WEB 1, 2 (2019).

80. *See id.*

81. *See id.*

82. *See id.* at 15.

83. These kinds of algorithms have been developed. *See, e.g.*, Anand Avati et al., *Improving Palliative Care with Deep Learning*, *in* IEEE BIBM INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM) 2017: MEDICAL INFORMATICS AND DECISION MAKING 57, 57–63 (2018) (developing a deep learning model using electronic health records to predict mortality within the next three–twelve months); Kris Newby, *Compassionate Intelligence: Can Machine Learning Bring More Humanity to Health Care?*, STAN. MED. (2018), https://stanmed.stanford.edu/2018summer/artificial-intelligence-puts-humanity-health-care.html (explaining the emerging use of AI in palliative care).

data about the sex of the patient but did have data on weight, then (given that women on average weigh less than men ceteris paribus) the algorithm would attribute more importance to the weight variable than is warranted. If someone then examined the algorithm to see what factors are predictive of mortality, they would infer that weight is more important than it actually is. This is especially worrisome in contexts like this hypothetical given that weight is not clearly *unrelated* to mortality—some portion of the correlation with mortality is potentially causal, but the other portion is attributable to the missing sex variable.

Although methods have been developed that do not directly use protected class variables,[84] these methods are less effective at achieving fairness objectives.[85] They implicitly leverage proxies of protected class variables, such that the decision-making is still driven by correlations with the protected class variables.[86] Moreover, if the proxies are imperfect, these approaches can lead to biased decision-making within the protected class groups.[87] To illustrate this, if the goal were to address racial bias in an algorithm, and zip code were used as an imperfect proxy for race, then bias mitigation would improve the algorithm's performance for minorities who live in predominantly minority zip codes but might worsen it for minorities who live in predominantly majority zip codes. This effort to even out performance across racial groups through using zip code could thus ironically harm certain minorities.[88]

Moreover, prohibiting the use of protected class variables and close proxies does little to prevent intentional obfuscation of biased

---

84. *See, e.g.*, Toshihiro Kamishima et al., *Fairness-Aware Classifier with Prejudice Remover Regularizer*, *in* PROCEEDINGS OF THE 2012 EUROPEAN CONFERENCE ON MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES 35, 35–50 (2012) (proposing an approach that uses the protected attribute as a regularizer); Dino Pedreshi et al., *Discrimination-Aware Data Mining*, *in* PROCEEDINGS OF THE 14TH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE, DISCOVERY, AND DATA MINING 560, 560 (2008) (proposing an approach that uses the protected attribute in selecting acceptable rules); Muhammad Bilal Zafar et al., *Fairness Constraints: A Flexible Approach for Fair Classification*, 20 J. MACH. LEARNING RSCH. 1, 8–15 (2019) (proposing a framework to construct fairness constraints that indirectly use protected attributes).

85. *See, e.g.*, Lipton et al., *supra* note 16, at 1–2, 9 (analyzing the shortcomings of disparate learning processes, which use sensitive features at training but not at prediction time, in terms of accuracy and impact parity).

86. *Id.* at 1.

87. *Id.* at 9.

88. *See id.*

algorithms.[89] Studies have shown that there are techniques to conceal whether an algorithmic decision-making process is driven by protected class variables.[90] Due to what is called the Rashomon or multiplicity effect, it is possible to derive very different explanations of model behavior from functionally equivalent models.[91] As a result, if a developer creates a biased model leveraging protected class variables, but is prohibited by the law from using that algorithm, the developer can simply develop a functionally equivalent one that does not appear to base decision-making on protected class variables but achieves the same biased outcomes.[92] Thus, while the ML literature has resoundingly concluded that removing protected class variables does little to promote fairness due to proxy variables, strict anti-classification stances illustrate an important disconnect between the ML and legal/policy communities. [93]

The Court itself has also observed the potential for proxy variables to be used to disguise improper racially motivated policies. In the 2016 case *Fisher v. University of Texas* ("*Fisher II*")[94] the Court dismissed the notion that the Texas Ten Percent Plan, which leveraged school district as a proxy for race, would be beyond reproach simply because it is facially race neutral:

> [T]he Top Ten Percent Plan, though facially neutral, cannot be understood apart from its basic purpose,

---

89. *See generally* Botty Dimanov et al., *You Shouldn't Trust Me: Learning Models Which Conceal Unfairness from Multiple Explanation Methods*, *in* PROCEEDINGS OF THE 24TH EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE (2020) (showing how unfairness can easily be disguised from explanation methods that seek to determine whether protected class variables are used in algorithmic decision-making).

90. *See generally id.* (explaining these concealing techniques).

91. *See* Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199, 206 (2001); Lesia Semenova et al., *A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning*, 2020 ARXIV 1, 5.

92. *See generally* sources cited *supra* note 91 (describing the Rashomon effect).

93. *See, e.g.*, Corbett-Davies & Goel, *supra* note 19, at 8, 9 (discussing the limitations of anti-classification); Hardt et al., *supra* note 16, at 1 ("A naïve approach might require that the algorithm should ignore all protected attributes such as race, color, religion, gender, disability, or family status. However, this idea of 'fairness through unawareness' is ineffective due to the existence of *redundant encodings*, ways of predicting protected attributes from other features." (citation omitted)); Kristian Lum & James E. Johndrow, *A Statistical Framework for Fair Predictive Algorithms*, 2016 arXiv 1, 1 (describing how the exclusion of protected variables in an analysis is insufficient to avoid discrimination).

94. 136 S. Ct. 2198 (2016).

> which is to boost minority enrollment. Percentage plans are 'adopted with racially segregated neighborhoods and schools front and center stage.' 'It is race consciousness, not blindness to race, that drives such plans.' Consequently, petitioner cannot assert simply that increasing the University's reliance on a percentage plan would make its admissions policy more race neutral.[95]

Although the Court's dicta in *Fisher II* on this point was not binding given that the Top Ten Percent Plan was not challenged by the plaintiffs, the Court's reasoning suggested that such systems that rely on proxy variables are still suspect due to their racial motivations.[96] This in turn raises the question of what should be considered race-neutral in the context of algorithms; Part VII will propose a possible framing based on the causal relationship between race and the algorithmic decisions.[97]

B. Protected Class Variables as Context

While removing protected class variables does not necessary mitigate bias, including them in algorithms can actually improve both fairness and accuracy, depending on how they are used.[98] Consider a car insurance provider that is designing an algorithm to predict whether people will get into car accidents in order to set premiums. For the purpose of this example, let us assume that people who drive red cars are more likely to get into car accidents as the flashiness of a red car tends to attract drivers who are more reckless. Let us also assume, however, that because red is a lucky color in some Asian cultures, Asian drivers are less likely to choose red cars because they are flashy and more likely to choose them because they are lucky. If this were the case,[99] then Asian red-car drivers would be less likely to get into car accidents than non-Asian red-car drivers. In this scenario, an algorithm should flag a non-Asian red-car driver as being at higher risk of a car accident than an Asian red-car driver. This would not only be fairer, in that it would

---

95. *Id.* at 2213 (citations omitted).
96. *See id.*
97. *See infra* Part VII.
98. *See supra* note 16.
99. Please note that this example is purely for illustrative purposes, and there is not necessarily an empirical basis for any of the trends discussed.

prevent these algorithms from unduly penalizing Asian red-car drivers, but it would also allow the algorithm to more accurately predict accident risk. The intuition behind this example is that in this case, the cultural context associated with the protected class variable of race helps inform how other variables should be interpreted. Including an interaction variable that distinguishes between "Asian red-car driver" vs. "Non-Asian red-car driver" would thus allow the model to learn the differences based on cultural context.[100]

The important role that protected class attributes can play in enhancing both fairness and accuracy of algorithms was also noted by the Wisconsin Supreme Court in *State v. Loomis*.[101] The plaintiff, Eric Loomis, had challenged the use of sex in determining his COMPAS risk score, which a judge had considered in determining his sentence.[102] The court dismissed Loomis' due process argument on the ground that "any risk assessment tool which fails to differentiate between men and women will misclassify both genders."[103] The court found it compelling that "the inclusion of gender promotes accuracy," thus "serv[ing] the interests of institutions and defendants, rather than a discriminatory purpose."[104] "Notably, however, Loomis [did] not bring an equal protection challenge,"[105] so it is difficult to say how the court would have ruled on that basis.

Further emphasizing how protected class variables can provide important context, Barnes et al. observe in *Judging Opportunity Lost* that the anti-classification trend in recent affirmative action jurisprudence fails to account for the ways in which race shapes the

---

100. Of course, the ideal case would be to have direct data on how reckless the driver is or at least data on whether the driver is buying a red car because they see it as flashy or lucky. If the former data were available, we might not need car insurance premium algorithms at all because we could just charge people directly based on how reckless they are at driving. The latter data would also be difficult to get, as it would involve surveying people on their reasons for buying various cars, and the incentive to lie would be very high.

101. 881 N.W.2d 749, 766 (Wis. 2016) (citing Melissa Hamilton, *Risk–Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231, 255 (2015)).

102. *Id.* at 757.

103. *Id.* at 766.

104. *Id.* (citing Hamilton, *supra* note 101, at 255).

105. *Id.*

lived experiences of racial minorities in the United States.[106] They emphasize that race provides valuable context in considering students' achievements.[107] Knowing, for example, that a student is the first African American student body president in her predominantly white school is important context for understanding the barriers she confronted and the significance of her achievement.[108] In this case, taking into account the protected class status of the applicant can not only result in fairer decisions but also more accurate ones in terms of assessing the applicant's achievements, ability, and potential.[109] Thus, protected class variables can play an important role in yielding fairer and more accurate algorithms through providing additional contextual information.[110] As will be discussed in Part VII, one of the benefits of framing algorithmic discrimination issues in terms of causality is being able to distinguish between how the protected class variable is being used—as a factor in and of itself or as additional context.[111]

Thus, algorithmic fairness highlights the inability to simultaneously be conscious of and neutral to protected class variables.[112] As I will discuss in the following Part, the challenges in implementing algorithmic fairness methods in practice are analogous to the long-standing tension in anti-discrimination law

---

106. *See* Mario L. Barnes et al., *Judging Opportunity Lost: Assessing the Viability of Race-Based Affirmative Action after* Fisher v. University of Texas, 62 UCLA L. REV. 271, 303–04 (2015).

107. *Id.* at 292.

108. *See id.* at 293. *Ho Ah Kow v. Nunan*, 12 F. Cas. 252 (C.C.D. Cal. 1879) also illustrates the importance of context in considering the effects of a policy. In that case, a Chinese man challenged an ordinance requiring all men in city jail to cut their hair to be no longer than an inch. Ho Ah Kow v. Nunan, 12 F. Cas. 252, 252 (C.C.D. Cal. 1879). At the time, wearing a queue was an important part of Chinese culture, so for Chinese men, this was an especially offensive policy. *Id.* at 253. The ordinance was passed at a time when another city ordinance, the "Cubic Air Law" was in effect to target crowded living conditions common in Chinatown. *Id.* The ordinance required at least 500 cubic feet of air for each adult residing in a residence, and many Chinese men resisted the law, refusing to pay a fine, and thus crowded the city jail. *Id.* The court took into account this cultural and political context in ruling that the law violated the Fourteenth Amendment, concluding that forcing prisoners to cut their hair was a significant additional punishment for Chinese men and noting the "general feeling—amounting to positive hostility—prevailing in California against the Chinese, which would prevent their further immigration hither and expel from the state those already here." *Id.* at 256.

109. *See, e.g.*, *supra* notes 93–96.

110. *See, e.g.*, *supra* notes 93–96.

111. *See infra* Part VII.

112. *See, e.g.*, *supra* notes 83–87.

jurisprudence between anti-classification and anti-subordination. Although in recent years the Court has increasingly adopted an anti-classification framework,[113] algorithmic fairness could revitalize or at least cast anti-subordination into a new light. The fact that algorithms can easily be trained on millions of previous cases suggests that they are more susceptible to detecting and perpetuating societal biases and existing hierarchical structures. Moreover, the fact that algorithms can in turn be scaled to make decisions affecting millions of individuals suggests that simply accepting algorithms as they are without any effort to mitigate bias would be less acceptable than in the human decision-maker context.

III. ANTI-CLASSIFICATION VS. ANTI-SUBORDINATION

The fundamental tension highlighted in this Article—the desire to have decision-making processes be blinded to protected class attributes versus the necessity of considering such attributes or related attributes in order to proactively address discrimination—is analogous to a long-standing tension in anti-discrimination law between the principles of anti-subordination and anti-classification. This Part will discuss this tension and how the Court's recent shift toward anti-classification masks a history of distinguishing between benign and malicious uses of protected class attributes. Given that a strict anti-classification stance would preclude most methods proposed in the algorithmic fairness literature due to their reliance on protected class variables or proxies, in order to enable ML practitioners to actively mitigate algorithmic bias, distinguishing between different uses of such attributes is vital.[114] As will be discussed further in Part VII, from a technical perspective, causal inference can help to draw these distinctions.

"The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."[115] This famous quote by Justice Roberts in his majority opinion in the 2007 case *Parents Involved in Community Schools v. Seattle School District No. 1*, captures the Court's current skepticism of race-conscious interventions.[116] In that case, the school districts had voluntarily adopted student

---

113. *See* Balkin & Siegel, *supra* note 18, at 12–13.

114. *See, e.g.*, Lum & Johndrow, *supra* note 93.

115. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007).

116. *See id.*

assignment plans that took race into consideration in allocating students to school districts.[117] Citing *Brown v. Board of Education*,[118] the Court concluded that because the Seattle school district had never been segregated and the Jefferson County school district had "removed the vestiges of past segregation," neither district could justify its use of "race-based assignments."[119]

The Court has not always taken such a stark stance against race-conscious decisions. For example, *in Lau v. Nichols*,[120] the Court recognized that equal treatment and equal opportunity sometimes require providing additional resources for minority groups.[121] In that case, the Court found that Chinese-speaking students were discriminated against in public schools that did not provide Chinese language instruction or supplemental English lessons.[122] The Court reasoned that, "there is no equality of treatment merely by providing students with the same facilities, textbooks, teachers, and curriculum; for students who do not understand English are effectively foreclosed from any meaningful education."[123] Although the Court did not address the Equal Protection Clause and instead reached this decision on the basis of § 601 of the Civil Rights Act of 1964, 42 U.S.C. § 2000d, it is very notable that the Court's holding was based on its interpretation of the meaning of "discrimination."[124] The Court looked at the outcomes rather than simply the inputs into the system when concluding that "the Chinese-speaking minority receive fewer benefits than the English-speaking majority from respondents' school system which denies them a meaningful opportunity to participate in the educational program—all earmarks of the discrimination banned by the regulations."[125] Even though providing additional resources for Chinese-speaking students might be suspect under anti-classification, the Court used an anti-subordination argument, concluding that "[s]imple justice requires that public funds, to which all taxpayers of all races contribute, not be spent in any fashion

117. *Id.* at 709–10.
118. 349 U.S. 294 (1954).
119. *Parents Involved*, 551 U.S. at 733, 746–48 (emphasizing that *Brown* required school districts "'to achieve a system of determining admission to the public schools *on a nonracial basis*'" (quoting *Brown*, 349 U.S. at 300–01)).
120. 414 U.S. 563 (1974).
121. *See id.* at 566.
122. *Id.* at 568.
123. *Id.* at 566.
124. *See id.* at 566, 568.
125. *Id.* at 568 (citation omitted).

which encourages, entrenches, subsidizes, or results in racial discrimination."[126]

Indeed, prior to the Court's decision in *Anarand Constructors, Inc. v. Peña*,[127] the Court had distinguished between benign and malicious uses of race in decision-making. The Court in *Metro Broadcasting, Inc. v. F.C.C.*[128] held that *benign* racial classifications used by the federal government, such as those designed to address historical discrimination, should only be held to intermediate scrutiny instead of strict scrutiny.[129] The Court in that case upheld the F.C.C. policies at issue, even though they did not aim to remedy past discrimination, on the grounds that they were substantially related to the important government objective of enhancing broadcast diversity.[130] This distinction between benign and malicious uses of protected classifications was well-articulated by Judge Wisdom in *United States v. Jefferson County Board of Education*[131]:

> The Constitution is both color blind and color conscious. To avoid conflict with the equal protection clause, a classification that denies a benefit, causes harm, or imposes a burden must not be based on race. In that sense, the Constitution is color blind. But the Constitution is color conscious to prevent discrimination being perpetuated and to undo the effects of past discrimination. The criterion is the relevancy of color to a legitimate governmental purpose.[132]

This allowance for a color conscious conception of the Constitution would have been much more congruent with understandings of fairness in the algorithmic context, but the Court

---

126. *Id.* at 569 (quoting 110 CONG. REC. 6,543 (1964) (statement of Sen. Hubert Humphrey)).

127. 515 U.S. 200 (1995).

128. 497 U.S. 547 (1990).

129. *Id.* at 564–65.

130. *Id.* at 566.

131. 372 F.2d 836 (5th Cir. 1966).

132. *Id.* at 876. The idea that race consciousness could be used to dismantle historical discrimination was the prevailing view among courts in the 1960s. *See* Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles over Brown*, 117 HARV. L. REV. 1470, 1519–20 (2004) (citations omitted) (collecting cases from the 1960s).

in *Anarand* overturned the *Metro Broadcasting* rule, holding that all racial classifications by the federal government should be subject to strict scrutiny.[133] The Court reasoned that the Fifth and Fourteenth Amendments protect individuals instead of groups, such that "all governmental action based on race—a *group* classification long recognized as 'in most circumstances irrelevant and therefore prohibited'—should be subjected to detailed judicial inquiry to ensure that the *personal* right to equal protection of the laws has not been infringed."[134]

As discussed above,[135] however, algorithmic bias mitigation revolves around the idea that race and other protected class variables are often important for contextualizing data, such that protected class attributes often are not irrelevant.[136] For example,

---

133. Anarand Constructors, Inc. v. Peña, 515 U.S. 200, 227 (1995).

134. *Id.* at 227 (citation omitted) (quoting Hirabayashi v. United States, 320 U.S. 81, 100 (1943)).

135. *See supra* Part II.B.

136. The Court itself has also at times acknowledged the importance of considering the broader context around alleged discriminatory policies. In *Yick Wo v. Hopkins*, the Court established that race neutrality does not imply a lack of discrimination. *See* 118 U.S. 356, 373–74 (1886). That case examined a San Francisco ordinance, which made it illegal to operate a laundry in a wooden building without a special permit. *Id.* at 358. At the time, 95% of laundries operated in wooden buildings, and two-thirds of the wooden laundries were operated by Chinese people. *See id.* at 358–59. Most laundry owners applied for a permit, but out of the 200 applications from Chinese owners, only one was granted. *Id.* at 359. In contrast, virtually all non-Chinese applicants were granted a permit. *Id.* at 361. The Court ruled that, even if a law is impartial on its face, "if it is applied and administered by public authority with an evil eye and an unequal hand, so as practically to make unjust and illegal discriminations between persons in similar circumstances, material to their rights, the denial of equal justice is still within the prohibition of the [C]onstitution." *Id.* at 373–74. The kind of biased enforcement experienced by the plaintiffs, the Court concluded, amounted to "a practical denial by the state of that equal protection of the laws" and therefore violated the provision of the Fourteenth Amendment. *Id.* at 373.

This was a landmark case that came at a time when there were strong tensions between the Chinese and non-Chinese populations in California. Many Chinese immigrated to California in the 1850s during the Gold Rush, and Chinese economic success led them to be seen as a threat during the 1870s Recession. In 1875, the Page Act was passed, which officially prohibited the immigration of Chinese coolies and prostitutes. Page Act of 1875, ch. 141, § 5–6, 18 Stat. 477 (repealed 1974). But, in its application, the Act effectively prevented almost all immigration of Chinese women. *See* Alina Das, *Inclusive Immigrant Justice: Racial Animus and the Origins of Crime-Based Deportation*, 52 U.C. DAVIS L. REV. 171, 184–92 (2018). A few years later, in 1882, the Chinese Exclusion Act was passed, which prevented the immigration of all Chinese laborers. Chinese Exclusion Act, ch. 126, § 1, 22 Stat. 58 (1882) (repealed 1943). This was the law of the United States for sixty-one years until the passage of

the fundamental challenge with using arrest data to train risk assessment tools in the criminal justice system is disproportionate rates of arrest due to racially targeted policing practices, such that considering historical arrest records without any consideration of the influence of race leads to biased algorithms.[137]

Even the Court in *Adarand* conceded that there are distinctions between benign and malicious uses of racial classifications.[138] In fact, the Court sought to "dispel the notion that strict scrutiny is 'strict in theory, but fatal in fact'" and noted that "[t]he unhappy persistence of both the practice and the lingering effects of racial discrimination against minority groups in this country is an unfortunate reality, and government is not disqualified from acting in response to it."[139] Much of the Court's concern in overturning *Metro Broadcasting* stemmed from the challenges of distinguishing benign and malicious racial classifications rather than the notion that there is no distinction between benign and malicious classifications.[140] In fact, the Court stated that the "point of strict scrutiny is to 'differentiate between' permissible and impermissible governmental use of race."[141] Thus, even though the Court in recent

---

the 1943 Magnuson Act, when China became an ally against Japan in World War II. Magnuson Act, Pub. L. No. 78-199, § 1, 57 Stat. 600, 600 (1943). In addition to restrictions on immigration, see Chinese Exclusion Act §§ 1, 14. Chinese people in the United States were excluded from land ownership, voting, access to courts, employment (they could not be lawyers, doctors, teachers, pharmacists, barbers, hunters, etc.), naturalization, and interracial marriage. *See generally* David E. Bernstein, Lochner*, Parity, and the Chinese Laundry Cases*, 41 WM. & MARY L. REV. 211 (1999) (discussing, in part, the exclusion of Chinese workers from various occupations). As a result, many turned to the laundry business as one of the few available occupations. *See id.* at 220. The Court's reference to an "evil eye" was thus likely not only a reaction to the strong statistical evidence of discrimination but also to the broader context suggesting the law was targeted toward the Chinese community. *Yick Wo*, 118 U.S. at 373.

137. *See* Jennifer Skeem & Christopher Lowenkamp, *Using Algorithms to Address Trade-Offs Inherent in Predicting Recidivism*, 38 BEHAV. SCIS. & L. 259, 260 (2020) (finding that providing risk assessment algorithms with access to race variables can improve "both predictive accuracy and racial equity").

138. *See Anarand*, 515 U.S. at 228.

139. *Id.* at 237 (quoting Fullilove v. Klutznick, 448 U.S. 448, 519 (1980) (Marshall, J., concurring)).

140. *See id.* at 228 ("Justice Stevens chides us for our 'supposed inability to differentiate between 'invidious' and 'benign' discrimination,' because it is in his view sufficient that 'people understand the difference between good intentions and bad.' But, as we have just explained, the point of strict scrutiny is to 'differentiate between' permissible and impermissible governmental use of race." (citations omitted)).

141. *Id.*

years has construed the anti-classification principle to prohibit racial classifications in all but the most exacting circumstances, arguably a key part of the strict scrutiny analysis, as laid out in *Anarand*, should be an examination of whether race is used to ameliorate or perpetuate discrimination.[142]

One potential solution that has been proposed to reconcile anti-classification and anti-subordination principles is anti-balkanization.[143] Anti-balkanization is the idea that the constitutionality of race-conscious classifications depends on the extent to which it promotes versus undermines social cohesion.[144] The challenge with this approach in the algorithmic context is that it would discourage greater transparency and formal auditing for algorithms as such measures would increase the salience of and public awareness around race-conscious algorithmic methods. Legislative efforts requiring the measurement and mitigation of algorithmic bias could prove divisive. It has yet to be seen how the public at large would view such efforts, and algorithmic bias is still a relatively new issue in the public consciousness, so it is highly uncertain whether the issue becomes racially polarizing. Moreover, whereas marginalization and alienation in the context of schools with limited minority representation has been well-documented for affirmative action cases, the threats of algorithmic bias to social cohesion are less visible given that algorithmic bias itself is harder to detect. Thus, how courts would balance the concerns of social marginalization and racial resentment in the context of algorithmic bias is highly uncertain, making it unclear whether "race moderate[]" judges[145] would actually support race-conscious algorithmic decision-making.

Instead of focusing on social cohesion or the visibility of protected categories as the relevant axis, causality is a more appropriate factor for evaluating whether the protected class variable is being used to ameliorate or perpetuate discrimination. As will be discussed further in Part VII, interrogating the causal effect of using protected class variables can allow judges and regulators to distinguish between uses of such variables that arguably make the algorithms more race-

---

142. *See id.*

143. *See* Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278, 1283 (2011)).

144. *See id.* at 1299.

145. *Id.* at 1282. Siegel defines "race moderate[]" judges as those "who allow and limit civil rights initiatives in order to preserve social cohesion." *Id.*

neutral by reversing existing discriminatory trends versus uses that amplify existing disparities.

Thus, although the importance and relevance of anti-subordination as a motivating force for anti-discrimination law is by no means a new observation,[146] the emergence of algorithmic bias highlights that this long-standing feud in anti-discrimination law will soon see a new battleground in the realm of automated decision-making. The mathematical nature of algorithmic bias puts in stark relief existing contradictions in jurisprudence.

IV. RESPONSIBILITY FOR HISTORICAL DISCRIMINATION

A key preliminary issue in anti-discrimination law is the scope of responsibility that various actors have for addressing discrimination. As discussed above, algorithmic bias is often attributed to bias in past decisions and/or systemic inequities that are then reflected in the data used to train the algorithm.[147] As this Part will discuss, identifying the cause of an allegedly discriminatory outcome, especially separating out the role of the allegedly discriminating entity from general societal discrimination, is vital from a legal perspective for establishing liability that would then justify taking race-conscious remedial action. While most technical methods proposed to mitigate algorithmic bias do not distinguish between different sources of bias, methods based in causal inference naturally lend themselves to making this distinction.

In the context of equal protection, the Court in *Regents of the University of California v. Bakke*[148] reasoned that historical discrimination was only a compelling state interest if there was evidence of that specific school discriminating.[149] But the Court determined that general societal discrimination was insufficient to create a compelling state interest.[150] This reasoning is echoed in other cases, where decision-makers are generally responsible only

---

146. *See* Balkin & Siegel, *supra* note 18, at 9–10; Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N.Y.U. L. REV. 1003, 1007 n.12 (1986).

147. *See supra* Part I.

148. 438 U.S. 265 (1978).

149. *See id.* at 307–10, 308 n.44 (Powell, J., concurring); *see also* Freeman v. Pitts, 503 U.S. 467, 494 (1992) ("Racial balance is not to be achieved for its own sake. It is to be pursued when racial imbalance has been caused by a constitutional violation.").

150. *See Bakke*, 438 U.S. at 295–99 (Powell, J., concurring).

for biased outcomes that are the direct result of their own decision-making. In *Wygant v. Jackson Board of Education*,[151] a case where the Court was assessing a school that used race-based preferences in determining which teachers to lay off, the Court concluded that "[s]ocietal discrimination, without more, is too amorphous a basis for imposing a racially classified remedy."[152] In order to undertake an affirmative action program, the school needed to "have sufficient evidence to justify the conclusion that there has been prior discrimination."[153] This issue also emerged in *Parents Involved*, where the Court found that, although the state has the "compelling interest of remedying the effects of past intentional discrimination" in the context of schools, such a justification did not apply in cases where schools "have not shown that they were ever segregated by law, and were not subject to court-ordered desegregation decrees."[154] The Court further found that, even in cases where there had been prior legal segregation and desegregation decrees, once the "harm that is traceable to segregation" had been remedied, "[a]ny continued use of race must be justified on some other basis."[155] The Court explicitly excluded "'racial imbalance in the schools, without more'" as a compelling interest satisfying strict scrutiny for race-conscious decision-making.[156]

Even outside of equal protection jurisprudence, a similar concept emerged in *Texas Department of Housing v. Inclusive Communities*, where the Court stated regarding disparate impact liability under the Fair Housing Act: "If a statistical discrepancy is caused by factors other than the defendant's policy, a plaintiff cannot establish a prima facie case, and there is no liability."[157] In *Ricci v. DeStefano*,[158] the Court found that "absent a strong basis in evidence of an impermissible disparate impact," taking action on the basis of or motivated by racial considerations would constitute disparate treatment.[159] As will be discussed further in Part VI.A, it is

151. 476 U.S. 267 (1986).
152. *Id.* at 276.
153. *Id.* at 277.
154. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) (citation omitted).
155. *Id.* at 721 (footnote omitted).
156. *Id.* (citations omitted) (quoting Milliken v. Bradley, 433 U.S. 267, 280 n.14 (1977)).
157. Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507, 2514 (2015).
158. 557 U.S. 557 (2009).
159. *Id.* at 585.

debatable to what extent *Ricci* applies to the algorithmic bias mitigation context, but the legal analysis pursued by the Court does suggest that a key question in determining whether disparate treatment is permissible is whether there would actually be disparate impact liability.[160]

On the face of it, the notion that remedial race-conscious decision-making cannot be justified by historical *societal* discrimination presents a legal obstacle to attempts to correct for algorithmic bias. In most cases, algorithmic bias does not arise from prior discrimination on the part of the institution developing or deploying the algorithm; it arises instead because the training data reflects many layers of historical biases.[161] This rule may not be straightforward to apply in practice, however.

One reason is that questions of responsibility are complicated when it comes to bias in algorithms due to the myriad parties involved in the creation of the data, model, and process behind an algorithmic decision. Suppose, for example, a jurisdiction with a history of racially biased policing contributes its data to a national dataset used to train a recidivism risk assessment tool. If that jurisdiction then uses the tool, can the jurisdiction take steps to mitigate the biases exhibited by the tool even though the majority of the data used to train the tool was not from that jurisdiction? On the one hand, the fact that the dataset is nationally representative would suggest that biases exhibited by the model are from societal bias. On the other hand, in this example, the jurisdiction directly contributed its own biased data to the creation of the dataset.

A second complication is that algorithmic decision-making not only reflects but also perpetuates the biases in past decisions. If a jurisdiction adopts a recidivism risk assessment tool trained on racially biased arrest data, that jurisdiction will likely make biased decisions going forward that further exacerbate existing societal inequities. For example, if higher proportions of black defendants were incorrectly labelled high risk by the tool and thus wrongfully detained, this would further entrench racial disparities in the criminal justice system. Thus, while it would be reasonable to say that users of such algorithmic decision-making tools should not be

160. *See infra* Part VI.A.

161. *See* Barocas & Selbst, *supra* note 23, at 680–81; Andrea Romei & Salvatore Ruggieri, *Discrimination Data Analysis: A Multi-Disciplinary Bibliography*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 109, 121 (Bart Custers et al. eds., 2013); Kristian Lum & William Isaac, *To Predict and Serve?*, SIGNIFICANCE, Oct. 2016, at 14, 16.

responsible for the biased historical decisions reflected *in their data*, they should be responsible for biased decisions made *by their tools*.

That said, if algorithm developers and users were allowed to mitigate societal biases reflected in their data, what kinds of bias should they be allowed to mitigate? Algorithmic fairness asks the question of what should be done to address algorithmic decision-making processes that might be based on biased historical decisions,[162] but most techniques for mitigating algorithmic bias do not distinguish between different historical sources of bias. One way to think about distinct sources of historical bias is the dichotomy between "measurement error" and "population disparities."[163] Measurement error refers to problems with the data collection process, such that the data do not accurately reflect reality. Population disparities, on the other hand, refer to real-world differences between populations, often due to systemic inequality. While in practice the lines between these two forms of bias blur—it can be difficult to determine when trends in data are based on flawed measurement—this dichotomy is a helpful framing given that correcting for measurement mistakes is less likely to be objectionable than correcting for population differences.

For example, there are several sources of measurement error that could make criminal justice data biased against minorities: biased police systems that are more likely to wrongfully arrest minorities, biased prosecutors who are more likely to levy charges against minority defendants, biased judges or juries who are more likely to convict minority defendants, and biased judges who are more likely to give minority defendants longer sentences.[164] These

---

162. Kleinberg et al., *supra* note 16, at 22 ("Because the data used to train these algorithms are themselves tinged with stereotypes and past discrimination, it is natural to worry that biases are being 'baked in.'"); *see also* David Madras et al., *Fairness Through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data*, *in* PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 349, 349–54 (2019).

163. *See also* Ben Green, *The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness*, *in* PROCEEDINGS OF THE 2020 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 594, 600–03 (2020) (discussing related concepts of "population inequity" and "human bias").

164. *See* Huq, *supra* note 27, at 1076–77 (discussing how face-value use of arrest data can reinforce historical patterns of policing); Rashida Richardson et al., *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15, 18 (2019) ("Dirty data—as we use the term here—also includes data generated from the arrest of innocent people who had evidence planted on them or were otherwise falsely accused, in addition to calls for service or incident reports that reflect false claims of criminal

factors all mean that criminal justice data does not necessarily measure what we care about—whether individuals actually committed a crime and the nature and severity of the crime—so efforts to tackle these biases and enable algorithmic decisions to better predict whether people will actually commit another crime (rather than be arrested again) should not be very objectionable.

Addressing population disparities can be more challenging. In the criminal justice context, mitigating population disparities would implicitly forgive or downplay the crimes of certain groups due to historical injustices. For example, cyclical poverty due to historical systematic disenfranchisement of certain demographic groups could contribute to higher rates of crimes committed by those groups.[165] Should anything be done to correct for algorithmic bias that stems from these trends? In the case of an individual who has committed theft, should poverty be considered as a mitigating factor given that the individual is more likely to have committed the crime out of necessity?[166] These questions of fairness and justice have been debated for much of human history,[167] so this Article will not claim

---

activity." (citation omitted)). *See generally* Mark W. Bennett, *The Implicit Racial Bias in Sentencing: The Next Frontier*, 126 YALE L.J.F. 391 (2017) (discussing implicit bias in sentencing); M. Marit Rehavi & Sonja B. Starr, *Racial Disparity in Federal Criminal Charging and Its Sentencing Consequences* 2–26 (Univ. Mich. L. & Econ. Empirical Legal Stud. Ctr., Working Paper No. 12-002, 2012) (discussing charges and sentencing).

165. *See* Philip Alston (Special Rapporteur on Extreme Poverty and Human Rights), *Rep. on His Mission to The United States of America*, at 7, U.N. Doc. A/HRC/38/33/Add.1 (May 4, 2018) (discussing the criminalization of poverty and indebtedness and its linkage to historical group disenfranchisement).

166. While the necessity defense of poverty remains largely unexplored, a number of proposed strategies have developed over the years, for example, the Rotten Social Background/Severe Environmental Deprivation defense and the Coercion of Poverty defense. *See* Michele E. Gilman, *The Poverty Defense*, 47 U. RICH. L. REV. 495, 500–10 (2013).

167. *See* KRZYSZTOF J. PELC, MAKING AND BENDING INTERNATIONAL RULES: THE DESIGN OF EXCEPTIONS AND ESCAPE CLAUSES IN TRADE LAW 43–56 (2016) (providing an overview of the Western intellectual history behind "necessity knows no law"); s*ee also* BRYAN W. VAN NORDEN, VIRTUE ETHICS AND CONSEQUENTIALISM IN EARLY CHINESE PHILOSOPHY 228 (2007) (quoting Mengzi, a Confucian philosopher: "To lack a constant livelihood, yet to have a constant heart – only a scholar is capable of this. As for the people, if they lack a constant livelihood, it follows that they will lack a constant heart. And if one simply fails to have a constant heart, dissipation and evil will not be avoided. When they thereupon sink into crime, to go and punish them is to trap the people. When there are benevolent people in positions of authority, how is it possible to trap the people?" (citation omitted)).

to offer the solution to them but rather to point out that algorithmic decision-making forces these questions to the forefront.

Thus, there are particular challenges facing the need to distinguish between correcting for historical *societal* discrimination versus historical discrimination *by the entity* when seeking to take race-conscious measures to address discrimination. Separating out the roles of different actors and attributing the sources of algorithmic bias to specific actors can be difficult. Moreover, the fact that algorithms not only learn from historical biases but also perpetuate them suggests that the distinction between correcting for past wrongs and preventing future ones breaks down in the algorithmic context. Finally, existing algorithmic fairness techniques rarely distinguish between different sources of bias, and it can be difficult in practice to draw the line as to which biases to mitigate and how to do so. As will be discussed in Part VII, one of the major benefits of using a causal framework for understanding algorithmic discrimination is that causal inference can help distinguish between different historical sources of bias.

V. GOVERNMENT ENTITIES: LESSONS FROM AFFIRMATIVE ACTION JURISPRUDENCE

As government entities increasingly deploy algorithmic systems,[168] it is likely that the constitutionality of algorithmic bias mitigation techniques will face litigation. This Part will discuss equal protection cases that govern the constitutionality of race-conscious remedies. The next Part will discuss the application of disparate impact and disparate treatment doctrines to the question of how to remedy algorithmic bias. These doctrines are the most relevant for considering how efforts by private entities to mitigate algorithmic bias might be evaluated.

Given that anti-discrimination jurisprudence in the United States is highly suspicious of decision-making that takes into

168. *See, e.g.*, S.B. 10, 2017–2018 Cal. Leg. Reg. Sess. (Cal. 2018) (proposing the mandatory use of criminal risk assessment tools instead of cash bail in California); Chouldechova et al., *supra* note 2, at 3–13 (discussing an algorithmic system deployed to evaluate child abuse and neglect risk in Allegheny County); Jay Stanley, *Pitfalls of Artificial Intelligence Decision-Making Highlighted in Idaho ACLU Case*, ACLU BLOG (June 2, 2017, 1:30 PM), https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case (detailing a case against an Idaho Medicaid program that deployed an algorithmic system to determine benefits).

account protected class attributes,[169] this Part will focus on affirmative action jurisprudence as one of the few areas of case law where the Court has recognized that not taking race into account can lead to adverse disproportionate outcomes (i.e., lack of diversity), such that actively accounting for and adjusting decision-making in light of students' race can be permissible.[170] Prior literature has elaborated on the constitutionality of collecting demographic data as part of the Census and on racial targeting in police investigations.[171] While these cases are relevant support for the idea that government entities can be race-conscious, these efforts are not clearly tied to race-conscious efforts to actively mitigate bias. As other papers have focused on the government contracting line of affirmative action cases,[172] this Part will elaborate primarily on the higher education line of cases.

The history of affirmative action jurisprudence in the United States can lend insight into some of the constitutional challenges that algorithmic bias mitigation techniques might face. Indeed, perhaps nowhere is the tension between anti-classification and anti-subordination more obvious than in affirmative action doctrine. This is unsurprising given that affirmative action is fundamentally motivated by anti-subordination principles,[173] such that the Court's increasing adherence to anti-classification principles presents a threat to the continued constitutionality of affirmative action. Affirmative action requires consideration of the group identity of individuals, which contradicts the anti-classification principle's focus

---

169. *See, e.g.*, Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 709–11 (2007) (prohibiting the use of race in determining school placement in districts without a history of explicit segregation); J.E.B. v. Alabama *ex rel.* T.B., 511 U.S. 127, 146 (1994) (prohibiting the use of gender in peremptory jury selection challenges); United States v. Maples, 501 F.2d 985, 987 (4th Cir. 1974) (prohibiting the use of gender in sentencing).

170. *See, e.g.*, Fisher v. Univ. of Tex. (*Fisher II*), 136 S. Ct. 2205, 2214–15 (2016) (upholding the use of racial affirmative action in higher education under specific constraints); Grutter v. Bollinger, 539 U.S. 306, 343 (2003) (same); Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 320 (1978) (same).

171. *See, e.g.*, Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 857–58 (2020).

172. *See* Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, U. CHI. L. REV. ONLINE 135–37 (2020).

173. The history of affirmative action suggests that the original motivation (before *Bakke*) for such policies were to create racial targets to help dismantle historical hierarchies and ameliorate historical injustices. *See generally* Michael Selmi, *The Life of* Bakke*: An Affirmative Action Retrospective*, 87 GEO. L.J. 981 (1999) (reflecting on affirmative action law before and after *Bakke*).

on individual protections against discrimination; engaging in affirmative action directly involves racially conscious or motivated decision-making.

The Fourteenth Amendment's Equal Protection Clause has been interpreted by the Court to subject racial classifications to strict scrutiny,[174] sex classifications to intermediate scrutiny,[175] and most other classifications to rational basis review.[176] In the context of racial affirmative action, this means that the racial classification must be narrowly tailored to serve a compelling state interest.[177] This Part will be structured around the two key components of this standard: (1) what constitutes a compelling state interest for the use of protected class attributes in decision-making; and (2) what is needed for the use of such variables to be considered narrowly tailored.

A. Diversity as a Compelling State Interest

Throughout the history of affirmative action, there have been two competing compelling state interests: (1) rectifying historical injustices; and (2) diversity. The first rationale was a strong motivation in early jurisprudence,[178] but in the landmark case *Regents of the University of California v. Bakke*, the Court decided

---

174. "It is well established that when the government distributes burdens or benefits on the basis of individual racial classifications, that action is reviewed under strict scrutiny." *Parents Involved*, 551 U.S. at 720 (first citing Johnson v. California, 543 U.S. 499, 505–06 (2005); then citing *Grutter*, 539 U.S. at 326; and then citing Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 224 (1995)).

175. *See* Miss. Univ. for Women v. Hogan, 458 U.S. 718, 724 (1982); Craig v. Boren, 429 U.S. 190, 197 (1976).

176. See generally Katie R. Eyer, *The Canon of Rational Basis Review*, 93 NOTRE DAME L. REV. 1317 (2018) for a discussion on applications of rational basis tests in civil rights case law.

177. *Grutter*, 539 U.S. at 326 (citing *Adarand*, 515 U.S. at 227). It is important to clarify the scope of these affirmative action cases. Because they are based on the Fourteenth Amendment's Equal Protection Clause, they apply specifically to government actors rather than private actors. *See* Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass'n, 531 U.S. 288, 295 (2000) ("Our cases try to plot a line between state action subject to Fourteenth Amendment scrutiny and private conduct (however exceptionable) that is not." (citations omitted)).

178. Affirmative action policy originally applied to executive departments and agencies and federal contractors. The Revised Philadelphia Plan, for example, not only prohibited racial discrimination but also created hiring goals for African American employees. *See* Barnes et al., *supra* note 106, at 279. *See generally* Balkin & Siegel, *supra* note 18 (describing the history of anti-subordination and anti-classification principles).

that the compelling state interest that justified the use of race-conscious decision-making in higher education admissions was diversity rather than historical discrimination.[179] Although the Court in *Bakke* concluded that diversity was an acceptable compelling state interest for the use of affirmative action in higher education, clearly defining a legitimate diversity rationale can be challenging.[180]

An important driver of the Court's decision that the University of Texas' admissions system survived strict scrutiny in *Fisher II* was the fact that the diversity goals of the university were sufficiently measurable without constituting a quota.[181] Specifically, the university cited the following as goals: destruction of stereotypes, promotion of cross-racial understanding, preparation of students for a diverse workforce and society, and cultivation of leaders with legitimacy in eyes of citizenry.[182]

There is a tension, however, between how clear diversity goals are and how similar they are to a quota. In *Fisher I*, the Court expressed that:

> A university is not permitted to define diversity as "some specified percentage of a particular group merely because of its race or ethnic origin." "That would amount to outright racial balancing, which is patently unconstitutional." "Racial balancing is not transformed from 'patently unconstitutional' to a compelling state interest simply by relabeling it 'racial diversity.'"[183]

The Court in *Fisher II* further reiterated this point:

> As this Court's cases have made clear, however, the compelling interest that justifies consideration of race in college admissions is not an interest in

---

179. Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 311–12 (1978) (Powell, J., concurring).

180. *See id.* at 314–15 ("As the interest of diversity is compelling in the context of a university's admissions program, the question remains whether the program's racial classification is necessary to promote this interest.").

181. *See* Fisher v. Univ. of Tex. (*Fisher II*), 136 S. Ct. 2198, 2210–11 (2016).

182. *Id.* at 2211.

183. Fisher v. Univ. of Tex. (*Fisher I*), 570 U.S. 297, 311 (2013) (citations omitted).

> enrolling a certain number of minority students. Rather, a university may institute a race-conscious admissions program as a means of obtaining "the educational benefits that flow from student body diversity. . . ." Increasing minority enrollment may be instrumental to these educational benefits, but it is not, as petitioner seems to suggest, a goal that can or should be reduced to pure numbers. Indeed, since the University is prohibited from seeking a particular number or quota of minority students, it cannot be faulted for failing to specify the particular level of minority enrollment at which it believes the educational benefits of diversity will be obtained.[184]

On the one hand, having very clearly specified diversity objectives, especially if numerical, can be seen as a quota. On the other hand, a university cannot "assert[] an interest in the educational benefits of diversity writ large" because its goals "must be sufficiently measurable to permit judicial scrutiny of the policies adopted to reach them."[185] Walking this line between overly "elusory or amorphous" diversity goals and overly specific, quantifiable goals is challenging in the algorithmic context.[186] It is impossible for an algorithm to approach a goal without "reduc[ing] [it] to pure numbers."[187] Indeed, the algorithmic fairness literature has rarely considered what a diversity rationale would look like.[188] The closest

---

184. *Fisher II*, 136 S. Ct. at 2210 (citations omitted).

185. *Id.* at 2211.

186. *Id.*

187. *Id.* at 2210.

188. The Algorithmic Fairness literature instead often uses the framing given in Sorelle A. Friedler et al., *On the (Im)possibility of Fairness*, 2016 ARXIV 1, 2, basing their interventions on either a "What You See is What You Get" or "We're All Equal" worldview. *See, e.g.*, Mohsen Abbasi et al., *Fairness in Representation: Quantifying Stereotyping as a Representational Harm*, *in* PROCEEDINGS OF THE 19TH SIAM INTERNATIONAL CONFERENCE ON DATA MINING 801, 801–06 (2019); Lily Hu & Yiling Chen, *Fairness at Equilibrium in the Labor Market*, *in* PROCEEDINGS OF THE 2017 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING 1, 1–2 (2017); Michael Wick et al., *Unlocking Fairness: A Trade-off Revisited*, *in* PROCEEDINGS OF THE 33RD CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 2–3 (2019). A diversity rationale as conceived by the Court likely exists somewhere between these two extremes as it would not assume that all observed differences are necessarily caused by structural inequalities, nor would it naively assume that observed differences dictate future success or accomplishment. *See* Friedler et al., *supra*, at 14. It would instead draw from each of these while also

consideration has been the demographic parity definition of group fairness.[189] As that approach defines fairness as proportional outcomes across demographic groups, it is analogous to the most basic definition of diversity, but it is also equivalent to illegal racial balancing.[190]

The way the Court has navigated this conundrum outside the algorithmic context has been to permit quantitative data that informs a specific diversity goal as long as the diversity goal itself is not a fixed set of numbers. For example, a diversity goal could be quantified as the minimum number of minority students needed to ensure that minority students in the school do not feel tokenized[191]—in this case, the diversity goal is a not a fixed number of minority students but a function of how tokenized minority students feel.[192] Indeed, the Court in *Fisher II* cited that the University presented evidence that "minority students admitted under the [race-neutral] *Hopwood* regime experienced feelings of loneliness and isolation" and that this "anecdotal evidence [was], in turn, bolstered by further, more nuanced quantitative data."[193] Although the

---

adding a new optimization objective on top of antidiscrimination, namely emergent benefits from diverse group engagement.

189. *See* Hardt et al., *supra* note 16, at 1–2, 17.

190. *See* Grutter v. Bollinger, 539 U.S. 306, 330 (2003) (citations omitted).

191. *Id.* at 329 ("As part of its goal of 'assembling a class that is both exceptionally academically qualified and broadly diverse,' the Law School seeks to 'enroll a "critical mass" of minority students.'" (citation omitted)). The Law School defined "critical mass" as "numbers such that underrepresented minority students do not feel isolated or like spokespersons for their race." *Id.* at 319 (citation omitted). The Court accepted maintaining a "critical mass" as an appropriate motive for racial affirmative action and concluded that, "[t]he Law School's goal of attaining a critical mass of underrepresented minority students does not transform its program into a quota." *Id.* at 335–36.

192. In *Grutter*, the Court viewed the fact that the Law School did not quantify "critical mass" to be positive. *Id.* at 318–19 ("Like the other Law School witnesses, Lehman did not quantify critical mass in terms of numbers or percentages." (citation omitted)). The majority used the fact that "the number of underrepresented minority students who ultimately enroll in the Law School differs substantially from their representation in the applicant pool and varies considerably for each group from year to year" to dismiss Justice Rehnquist's concern that "the Law School's policy conceals an attempt to achieve racial balancing." *Id.* at 336 (citations omitted).

193. Fisher v. Univ. of Tex. (*Fisher II*), 136 S. Ct. 2198, 2212 (2016) (citation omitted) ("In 2002, 52 percent of undergraduate classes with at least five students had no African-American students enrolled in them, and 27 percent had only one African-American student. In other words, only 21 percent of undergraduate classes with five or more students in them had more than one African-American student enrolled. Twelve percent of these classes had no Hispanic students, as compared to 10 percent in 1996." (citations omitted)). Ironically the fact that this quantitative

University did not present evidence of what the minimum number of minority students would be in order to achieve a "critical mass,"[194] the Court found the evidence presented to be sufficient to conclude that the "critical mass" had not been achieved under the race-neutral admissions system.[195] Similarly, in *Parents Involved in Community Schools v. Seattle School District No. 1*, the Court found issue with the fact that:

> [t]he plans are tied to each district's specific racial demographics, rather than to any pedagogic concept of the level of diversity needed to obtain the asserted educational benefits. . . . The districts offer no evidence that the level of racial diversity necessary to achieve the asserted educational benefits happens to coincide with the racial demographics of the respective school districts . . . .[196]

Thus, the Court suggested that having demographic goals coincide with district demographics could be permissible if tied to specific educational benefits.[197]

Implicit in the Court's statements is the idea that the schools must establish that there is a causal relationship between the level of diversity they seek and the legitimate non-racial-balancing objectives they seek to achieve. One benefit of causal inference is that it would allow government algorithmic developers to frame the

---

data was considered relevant implies that the Court had some notion of the need for a minimum number of minority students in each class, which would imply a quota.

194. *Id.* at 2211 ("Second, petitioner argues that the University has no need to consider race because it had already 'achieved critical mass' by 2003 using the Top Ten Percent Plan and race-neutral holistic review. Petitioner is correct that a university bears a heavy burden in showing that it had not obtained the educational benefits of diversity before it turned to a race-conscious plan. The record reveals, however, that, at the time of petitioner's application, the University could not be faulted on this score." (citation omitted)).

195. *Id.* at 2212 ("Though a college must continually reassess its need for race-conscious review, here that assessment appears to have been done with care, and a reasonable determination was made that the University had not yet attained its goals.").

196. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 726–27 (2007). The Court also contrasted the finding by Jefferson County's expert that "having 'at least 20 percent' minority group representation for the group [is necessary] 'to be visible enough to make a difference,'" with the County's practice of seeking proportional representation for minority groups. *Id.* at 727–28.

197. *See id.* at 726–28.

question in terms of whether their legitimate objective could have been met if they had not taken a particular measure to mitigate bias. For example, if the legitimate objective were ensuring a critical mass of minority students were admitted using an admissions algorithm such that minorities did not feel tokenized, the key question would be whether minorities would have felt tokenized if active measures were not taken to mitigate bias in the algorithm (i.e., if there were no affirmative action).

B. Narrowly Tailored

Given that diversity—not historical discrimination—is the appropriate compelling state interest for affirmative action in higher education, the next relevant question is what kind of admissions system is considered to be narrowly tailored to serving this compelling interest. *Bakke* and subsequent affirmative action cases put significant limitations on the extent to which race can be considered in admissions decisions.

Bakke struck down the University of California Davis's policy which set aside sixteen seats in its medical school class for minority applicants.[198] The Court concluded that this was an impermissible racial quota because there was no competition between minority and non-minority candidates for those seats.[199] Notably, however, the Court commended Harvard's admissions system, which considered race as one of many factors.[200] Banning racial quotas while allowing race to be taken into account as a "plus factor" would suggest that a potential solution is a point system, whereby there are no specific minimum or maximum number of students from each demographic group is admitted, but whereby students of particular racial groups are given preference.

Indeed, the University of Michigan instituted such a system.[201] The University used a 150-point scale for applicants, where a score of 100 points was needed for admission, and minorities were given

198. Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 271, 275 (1978).

199. *See id.* at 289 (Powell, J., concurring).

200. *Id.* at 316–17 ("An illuminating example is found in the Harvard College program: . . . In such an admissions program, race or ethnic background may be deemed a 'plus' in a particular applicant's file, yet it does not insulate the individual from comparison with all other candidates for the available seats." (footnote omitted)).

201. *See* Gratz v. Bollinger, 539 U.S. 244, 255 (2003).

an additional 20 points.[202] This practice, however, was challenged in *Gratz v. Bollinger*, a case decided on the same day as *Grutter v. Bollinger*. *Grutter* assessed the University of Michigan Law School's policy, which did not use a point system but sought to achieve a critical mass of minority students. The law school justified its policy by stating that it was necessary to "ensure that these minority students do not feel isolated or like spokespersons for their race; to provide adequate opportunities for the type of interaction upon which the educational benefits of diversity depend; and to challenge all students to think critically and reexamine stereotypes."[203]

The *Gratz* Court stressed that students must be assessed individually,[204] and concluded that the undergraduate admissions system did not allow for sufficient consideration of individual circumstances because it "automatically distribute[d] 20 points, or one-fifth of the points needed to guarantee admission, to every single 'underrepresented minority' applicant solely because of race."[205] Nonetheless, the *Grutter* Court found that the Law School's admissions system was permissible as it used race as one of many factors evaluated on an individual basis.[206] That said, the Court in dicta speculated that, in twenty-five years, affirmative action should no longer be necessary and schools should move to color-blind policies.[207] The Court did not, however, find issue with the fact that the undergraduate admissions system gave students 20 points for athletic ability or socioeconomic disadvantage and 10 points for being a resident of Michigan,[208] suggesting that point systems for non-protected-class attributes did not violate individualized decision-making.

The challenge with applying these cases in the algorithmic context is that there is no meaningful distinction from a mathematical perspective between using race as one of many factors in a decision-making process versus allocating additional points to

---

202. *Id.* at 256–57.

203. Grutter v. Bollinger, 539 U.S. 306, 380 (2003) (Rehnquist, J., dissenting) (citation omitted); *see also supra* notes 191–97 and accompanying text (discussing the "critical mass" motivation in greater detail).

204. *Gratz*, 539 U.S. at 270–79.

205. *Id.* at 270.

206. *See Grutter*, 539 U.S. at 339–40.

207. *See id.* at 343. Ginsburg notably dissented on this point. *See id.* at 345–46 (Ginsburg, J., concurring).

208. *Gratz*, 539 U.S. at 294–95 (Souter, J., dissenting).

those of a certain race.[209] In fact, Justice Souter's dissent in *Gratz* cogently made this point:

> The very nature of a college's permissible practice of awarding value to racial diversity means that race must be considered in a way that increases some applicants' chances for admission. Since college admissions is not left entirely to inarticulate intuition, it is hard to see what is inappropriate in assigning some stated value to a relevant characteristic, whether it be reasoning ability, writing style, running speed, or minority race. Justice Powell's plus factors necessarily are assigned some values. The college simply does by a numbered scale what the law school accomplishes in its 'holistic review,' [a] distinction [that] does not imply that applicants to the undergraduate college are denied individualized consideration or a fair chance to compete on the basis of all of the various merits their applications may disclose.[210]

Although there are mathematical ways to use race within an algorithm that do not involve explicitly adding additional points based on race, it is unclear that these methods would change the legal analysis. For example, there could be separate algorithms for different races, or race could affect the student's score in multiplicative ways instead of additive ways. These modifications, however, would not address the Court's concern about the students not being evaluated on an individualized basis.

One way to reconcile these seemingly contradictory rules the Court has put forward is to consider the visibility of race-conscious

---

209. Once a student's probability of admission is reduced to a number, using race as a factor implies adding some number to this probability. At this point, doing so is indistinguishable from a point system where some number is added to the points used to evaluate admissions. That said, one mathematical way to distinguish a point system from "plus factors" is to construe "plus factors" as being probabilistic in and of themselves. As such, the same number would not automatically be added to admissions probability for minorities. It is unlikely, however, that the Court meant to suggest that schools employ a lottery for how many additional points each minority applicant would receive.

210. *Gratz*, 539 U.S. at 295 (Souter, J., dissenting) (citation omitted).

decision-making processes.[211] Although there is functionally little distinction between setting aside sixteen seats for minorities versus keeping racial composition in mind when making admissions decisions, the visibility of race is very different in these two scenarios. In the former, it is clear to everyone that minorities are being treated differently and are competing against each other rather than the applicant pool as a whole. In the latter, the underlying dynamic is the same—if an admissions officer knows at the end of the day the school wants roughly sixteen minority students, then he or she will choose the best sixteen or so minority students to admit, meaning that these students are effectively only competing against each other. To those outside the admissions office, however, the appearance is that the entire applicant pool competes against each other.

As further evidence of this, the Texas Ten Percent Plan, which automatically admitted any students in the top ten percent of their high school class into all state-funded universities,[212] has yet to be challenged. The Plan was designed by Texas in the wake of *Hopwood v. Texas*,[213] in which the Fifth Circuit ruled that the University of Texas School of Law could not use race as a factor in admissions.[214] Although *Hopwood* was later abrogated by the Court in *Grutter*,[215] the Ten Percent Plan remained in place. In fact, the Department of Justice under the Bush administration even encouraged the Court in *Grutter* to rule against Michigan's system and used Texas's system as an example of a race-neutral alternative.[216] Although the Ten Percent Plan was more obviously the reason why Fisher was not admitted,[217] it is notable that she still decided to challenge the plan

---

211. *See* Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1369–75 (2010) [hereinafter *The Future of Disparate Impact*]; Richard Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After Ricci and Inclusive Communities* 11–19 (Univ. of Mich. Pub. L. & Legal Theory Rsch. Paper No. 471, 2015) [hereinafter *Of Visible Race-Consciousness*].

212. Fisher v. Univ. of Tex. (*Fisher II*), 136 S. Ct. 2198, 2205 (2016).

213. 78 F.3d 932 (5th Cir. 1996).

214. *Fisher II*, 136 S. Ct. at 2205 (citing *Hopwood*, 78 F.3d at 934–35, 948).

215. *See* Grutter v. Bollinger, 539 U.S. 306, 325–43 (2003).

216. Brief for the United States as Amicus Curiae Supporting Petitioner at 8–9, *Bollinger*, 539 U.S. 306 (No. 02-241).

217. *Fisher II*, 136 S. Ct. at 2208–09 ("The component of the University's admissions policy that had the largest impact on petitioner's chances of admission was not the school's consideration of race under its holistic-review process but rather the Top Ten Percent Plan. Because petitioner did not graduate in the top 10 percent

that explicitly considers race. In fact, Fisher even argued that the Ten Percent Plan alone is sufficient for achieving diversity, suggesting that the Plan was not objectionable.[218] Although the Court in *Fisher II* suggested that the Plan might still be subject to strict scrutiny because it is racially motivated (even though it is facially racially neutral), in practice it has yet to be challenged.

A direct application of the visibility reading to the algorithmic context would thus suggest that any algorithmic decision-making process that is "conscious" of a protected class variable would only be permissible if done covertly. Just as admissions offices have been incentivized to be very opaque about how they actually consider race in admissions,[219] an algorithmic process that obfuscates its use of race might be permissible under existing anti-discrimination jurisprudence. On the one hand, lack of transparency can be easier in the algorithmic context because: (1) the algorithms and data used to train them are most often the proprietary intellectual property of private companies;[220] (2) expertise is needed to evaluate algorithms for potential biases;[221] and (3) the ability to leverage many weak proxies in big data makes it easier to conceal discriminatory patterns.[222] That said, encouraging algorithm developers to address algorithmic bias by being very opaque would be a highly problematic conclusion at a time where there are growing calls for algorithmic transparency and audits,[223] particularly in high stakes contexts like criminal justice, employment, and healthcare, where important decisions are being made about individuals' lives.

Thus, the Court's fraught jurisprudence on affirmative action in higher education has provided complex signals as to the acceptable scope of race-conscious decision-making. The most direct and obvious

---

of her high school class, she was categorically ineligible for more than three-fourths of the slots in the incoming freshman class.").

218. *Id.* at 2213 ("Petitioner's final suggestion is to uncap the Top Ten Percent Plan, and admit more—if not all—the University's students through a percentage plan.").

219. *See Of Visible Race-Consciousness*, *supra* note 211, at 11–12, 15–16.

220. *See* Burrell, *supra* note 49.

221. *Id.* at 4.

222. *See* sources cited *supra* note 78.

223. *See, e.g.*, Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143, 148–49, 167 (2018) (highlighting the growing "rallying cry" for transparency in "highly automated systems"); Nicholas Diakopoulos & Michael Koliska, *Algorithmic Transparency in the News Media*, 5 DIGIT. JOURNALISM 809, 811 (2017) (aiming, in part, "to inform both scholars and practitioners by raising awareness for algorithmic transparency").

applications of the Court's reasoning to the algorithmic context could lead to unintended consequences, incentivizing the adoption of biased and/or opaque algorithmic tools, where the use of protected attributes is obscured.

Arguably, the incoherencies of affirmative action jurisprudence in the higher education context stem from an excessive focus on the visibility of race as an admissions factor rather than the causal effect of the consideration of race. While the Court's comfort with quantitative evidence for what constitutes a "critical mass" of minority students to avoid tokenization suggests some emphasis on the causal effect of race-based admissions, the Court's overriding focus has been on form over function (e.g., whether a factor is considered qualitatively—as one of many factors—versus explicitly quantified in a point system).[224] Reorienting the legal analysis toward the key causal questions will thus help avoid incoherent applications of the law to the algorithmic context, where explicit quantification is inevitable.[225]

VI. PRIVATE SECTOR: APPLYING DISPARATE IMPACT AND DISPARATE TREATMENT DOCTRINES

Many algorithms being deployed in high stakes contexts are developed and used by private entities. Statutory law rather than constitutional law is the most relevant in this context. In particular, the jurisprudence in this space centers on the doctrines of disparate treatment and disparate impact, which were established by Title VII of the 1964 Civil Rights Act.[226] Disparate treatment covers intentional discrimination, when individuals are treated differently because of their protected class attribute.[227] Disparate impact, on the other hand, encompasses liability in the context of disproportionate outcomes across protected classes, where a showing of intentionality is not necessary.[228] The tension between race consciousness and race

224. *See, e.g.*, Fisher v. Univ. of Tex. (*Fisher II*), 136 S. Ct. 2198, 2212 (2016).

225. One of the fundamental tensions between affirmative action jurisprudence and the algorithmic context is the incentives the Court has created for decision-making systems that consider race but do not quantify it. This phobia of quantification is incompatible with algorithms. *See* Ho & Xiang, *supra* note 174, at 134.

226. *See* Civil Rights Act of 1964, 42 U.S.C. § 2000e-2 (2018).

227. *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, EMPLOYMENT TESTS AND SELECTION PROCEDURES (2007).

228. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431–32, 436 (1971) (holding that Title VII of the Civil Rights Act of 1964 disallows seemingly neutral

neutrality when attempting to mitigate bias can be seen in the tension between disparate impact and disparate treatment. Preventing and addressing disparate impact can raise concerns about disparate treatment. Exploring these doctrines can thus help shed light on when courts would permit the active remediation of potential disparate impact through race-conscious algorithmic bias mitigation methods.

In the ML algorithmic fairness literature, disparate treatment and disparate impact doctrines are commonly simplified, with disparate treatment analogized to using protected class variables in the algorithm,[229] and disparate impact simply being characterized as disproportionate outcomes across groups.[230] To concretize disparate impact, the Equal Employment Opportunity Commission's 80-20 Rule is often cited in this literature as the legal standard.[231] These analogies, however, ignore the complexities that actually arise when trying to apply these doctrines to the algorithmic context.

For disparate treatment, it is unclear what would constitute evidence of intentionality in the context of algorithms. Direct evidence would be rare, as it means that the defendant admitted that it was motivated by discrimination or the policy itself is overtly discriminatory.[232] In the algorithmic context, the lack of direct contact with human decisionmakers would make the former evidence unlikely. What would constitute an overtly discriminatory algorithm is also unclear given the many proposed definitions for algorithmic bias; the presence or absence of a protected class

---

employment practices if they result in discrimination on the basis of a protected class attribute even when the result is unintentional).

229. *See* Alice Xiang & Deb Raji, *On the Legal Compatibility of Fairness Definitions*, *in* PROCEEDINGS OF THE WORKSHOP ON HUMAM-CENTRIC MACHINE LEARNING AT THE 33RD CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 2–3 (2019); *see, e.g.*, Dwork et al., *supra* note 72; Lipton et al., *supra* note 16, at 3.

230. *See* Xiang & Raji, *supra* note 229, at 3; s*ee, e.g.*, Corbett-Davies & Goel, *supra* note 19, at 3; Feldman et al., *supra* note 16, at 259.

231. *See, e.g.*, Feldman et al., *supra* note 16, at 259; Muhammad Bilal Zafar et al., *Fairness Constraints: Mechanisms for Fair Classification*, *in* PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS (AISTATS) 1, 1–2 (2017).

232. Plaintiffs can also present circumstantial evidence, including "suspicious timing, ambiguous statements oral or written, behavior toward or comments directed at other employees in the protected group, and other bits and pieces from which an inference of discriminatory intent might be drawn." Troupe v. May Dep't Stores, 20 F.3d 734, 736 (7th Cir. 1994) (citations omitted).

attribute in an algorithm does not imply bias or lack thereof.[233] Alternatively, plaintiffs could show that certain employees were systematically treated differently based on their protected class, but the purpose of this evidence is to draw an inference of discriminatory intent. Because intentionality is key to disparate treatment analysis, it is unclear that the mere presence of a protected class attribute in the training data—especially if it were used to mitigate biased patterns in the data—would constitute disparate treatment.

In order to establish disparate impact liability, the prima facie showing of disproportionate outcomes across groups is only the first step of the burden-shifting framework. The defendant has the opportunity to show that there is a business necessity for the decision-making process.[234] For example, a requirement that firefighters be able to lift a certain amount of weight might lead disproportionately fewer women to become firefighters, but this requirement would not be considered discriminatory if the fire department could show that being able to lift such weight is necessary for the job. After the defendant establishes a business necessity defense, the plaintiff must show that there is a less-discriminatory alternative that would fulfill the business necessity.[235]

There are a number of challenges facing plaintiffs who seek to establish disparate treatment or disparate impact in the context of algorithmic decision-making. As discussed in the Introduction, the focus of this Article is algorithmic bias that does not result from discriminatory intent on the part of the algorithm's developer or user. In this context, the plaintiff would not be able to pursue the direct method of showing disparate treatment because there would be no evidence that the employer was motivated by discriminatory intent. That said, disparate treatment doctrine also features a burden-shifting framework: the *McDonnell Douglas* Framework.[236] After the plaintiff establishes a prima facie case,[237] the defendant

---

233. In addition, algorithms can make it easier to mask discriminatory intent. *See* Barocas & Selbst, *supra* note 23, at 692.

234. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

235. *See* Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(k) (2018).

236. *See* McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973).

237. *Id.* One way a plaintiff can establish a prima facie case by showing "(i) that he belongs to a racial minority; (ii) that he applied and was qualified for a job for which the employer was seeking applicants; (iii) that, despite his qualifications, he was rejected; and (iv) that, after his rejection, the position remained open and the employer continued to seek applicants from persons of complainant's qualifications." *Id.* (footnote omitted).

employer must articulate a legitimate, nondiscriminatory reason for the employment action. If they are able to do so, the plaintiff must demonstrate that the employee's reason was pretext for discrimination.[238]

Based on these burden-shifting frameworks, as long as an algorithm uses features that are not overtly discriminatory and is optimized for a reasonable business objective, then the employer would be able to satisfy its defenses under both doctrines. To illustrate this, take for example a hiring algorithm that is used to filter resumes for a job. The algorithm might have "years of relevant job experience" and "major in college" as inputs and be optimized to predict what the individual's first-year performance review would be, based on historical data from the employer. As the scandal around Amazon's employment algorithm showed, such algorithms can easily become biased due to historical biased decision-making.[239] Even though Amazon did not explicitly indicate to the algorithm that features related to gender were relevant, the algorithm learned to distinguish along gendered lines.[240]

Under disparate treatment, the employer could show that the plaintiff's overall score was lower than others, so there was a legitimate, nondiscriminatory reason not to hire them. For example, the employer might state that the plaintiff was not hired because they had an economics degree instead of a computer science degree, and based on the algorithm, those with computer science degrees had better first-year performance reviews. In practice, given that those with computer science degrees are overwhelmingly male, this might lead to female candidates being undervalued, but the employer could point to the algorithm's data-based optimization as evidence that those with computer science degrees performed better. Under disparate impact, the employer could show that the algorithm was optimized for a legitimate business objective because performance reviews reflect the quality of work of employees.

Moreover, the algorithmic context provides additional challenges for plaintiffs who seek to rebut an employer's defense. In the disparate treatment context, the plaintiff would struggle to show evidence that the employer's justification is a pretext for intentional discrimination. Unless the algorithm had features that were overtly related to a protected class attribute or obviously irrelevant to the

238. *Id.* at 804.
239. *See* discussion *supra* Part I.
240. *See* discussion *supra* Part I.

decision at hand, the plaintiff would not have evidence that other similarly situated individuals were treated differently.

In the disparate impact context, establishing the existence of a less-discriminatory alternative that achieves the same business objectives would be challenging. By their nature, algorithms are optimized to meet a specific objective. As a result, there is generally a trade-off between algorithmic fairness and accuracy.[241] The intuition behind this is that correcting for algorithmic bias reflects an assumption that the data used for training and testing the algorithm do not reflect the "ground truth"[242] or that the "ground truth" is reflective of past discrimination that should not be reproduced by the algorithm.[243] Accuracy, however, only shows how well the algorithm performs on the test dataset. As a result, modifying the algorithm to reduce algorithmic bias will, ceteris paribus, reduce the accuracy of the algorithm. In order to maintain the same level of accuracy while reducing the bias of the algorithm, the plaintiff would need to develop a fundamentally better algorithmic technique or find less biased data, either of which would be a tall order. If the original algorithmic technique used was truly state of the art, the former might be impossible.

As a result, disparate treatment and disparate impact doctrines do not neatly apply to the algorithmic context. There are particular challenges that plaintiffs will likely face to succeed in cases alleging discrimination by algorithms under these doctrines. These discrepancies are important when it comes to evaluating what can be done in the algorithmic context to mitigate potential disparate impact without triggering disparate treatment liability.

---

241. *See, e.g.*, Feldman et al., *supra* note 16, at 262; Hardt et al., *supra* note 16, at 18; Michael Kearns et al., *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness*, *in* PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 1, 4–5 (2018); Lipton et al., *supra* note 16, at 3, 13.

242. "Ground truth" is a statistics and ML term "that means checking the results of machine learning for accuracy against the real world." *Ground Truth*, TECHOPEDIA, https://www.techopedia.com/definition/32514/ground-truth (last visited Mar. 28, 2021). For example, data on recidivism seeks to measure whether an individual has committed another crime, so "ground truth" in this case would be whether the individual actually committed another crime. *See id.*

243. *See* Friedler et al., *supra* note 188, at 6.

A. Does Correcting for Disparate Impact Require Disparate Treatment? A Comment on Ricci v. DeStefano

Given the challenges discussed above with applying disparate treatment and disparate impact doctrines to the algorithmic context, a key question is what can be done to mitigate algorithmic bias that might lead to disparate impact without committing disparate treatment. One of the most notable cases that has been examined by the algorithmic fairness community on this question is *Ricci v. DeStefano*.[244] The context for this case was that the fire department in New Haven, CT (the "City"), created a new process for promotion that included both a written examination and interview.[245] Many firefighters invested significant time and energy in preparing for the exams.[246] After the results came back, however, only white and Hispanic firefighters would be promoted, so the City decided to invalidate the results and use a different promotion system instead.[247] The plaintiffs—white and Hispanic firefighters who would have been promoted—argued that this constituted disparate treatment while the City argued that it had thrown out the results due to concerns about being subjected to disparate impact liability.[248] The Court found that the City's actions constituted illegal disparate treatment because it could not show a "strong basis in evidence" that it would have been subjected to disparate impact liability.[249]

In the algorithmic fairness literature, some have interpreted the *Ricci* decision as suggesting that ML developers cannot conduct any bias mitigation after training or deploying their algorithm, the idea being that if you discover disparate impact after the model is trained or deployed, you cannot take any action to remedy it.[250] This reading of the case, however, ignores the fact that the Court was especially influenced by the fact that firefighters had spent a lot of time and money studying for the test, such that there were significant reliance interests at stake.[251] This, however, is generally not the case

---

244. 557 U.S. 557 (2009).
245. *Id.* at 564.
246. *Id.* at 583–84.
247. *Id.* at 566, 573–74.
248. *Id.* at 575.
249. *Id.* at 563, 587.
250. *See* Kroll et al., *supra* note 25, at 694–95.
251. *See* Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 198–99 (2017).

in the algorithmic context. Data collection is often a passive process—end users rarely expend significant time and energy with the goal of providing data to an algorithm developer.[252] It is also very common to modify an algorithm after it has been deployed, so doing so is unlikely to violate the expectations of users.

Moreover, the Court wrote extensively about how the original tests were designed with diversity in mind, such that the City's claims that the tests were not sufficiently scientific and would expose them to disparate impact liability, and were not credible.[253] This suggests that the Court was not opposed generally to methods to mitigate bias.[254] In fact, the Court did not find any problems with the fact that the City had deliberately assembled a racially balanced board to evaluate the testing process.[255] One way to understand this is through the importance of visibility, with the Court wanting to reduce the social salience of race.[256] The decision to discard the results was divisive because it created an identifiable set of victims while it was difficult to discern who was harmed by the race consciousness of the initial design. As discussed in Part V.B, however, applying a visibility reading to the algorithmic context can lead to unintended consequences as it would incentivize further opacity in the algorithmic design process at a time when there are growing calls for greater algorithmic transparency.[257]

Finally, there is the "institutional reading" of the Court's decision.[258] Under this reading, the Court in *Ricci* was especially motivated by the concern that the City had proactively proceeded with a remedy that was racially motivated rather than having a court impose a remedy.[259] The Court is comfortable with race-conscious remedies to disparate impact as long as the remedy is imposed under court order after a finding of disparate impact.[260] In

---

252. In fact, it is arguably often *too* passive of a process, with users often not realizing that their data is being collected and used to train algorithms. *See generally* BROOKE AUXIER ET AL., PEW RSCH. CTR., AMERICANS AND PRIVACY: CONCERNED, CONFUSED AND FEELING LACK OF CONTROL OVER THEIR PERSONAL INFORMATION (2019) (explaining how the surreptitious collection of personal data has affected the way Americans live their lives).

253. *Ricci*, 557 U.S. at 574–76.

254. *See* Kim, *supra* note 251, at 199–200.

255. *Ricci*, 557 U.S. at 585.

256. *The Future of Disparate Impact*, *supra* note 211, at 1346.

257. *See supra* Part V.B.

258. *The Future of Disparate Impact*, *supra* note 211, at 1364.

259. *See id.*

260. *See id.*

fact, in the South, courts gave disparate impact relief against employers with histories of overt racism and large-city police and fire departments.[261] This distinction reflects a concern that public employers cannot be trusted to deal with race appropriately.[262] Patronage and self-perpetuation made it very difficult for minorities to become firefighters and police officers despite the lack of formal discriminatory policies.[263] In the 80s and 90s, however, black and Latino voters became more important, so cities began working harder to integrate police and fire departments, often relying on disparate impact to provide the necessary cover to do so.[264] Thus, the City's incentives in *Ricci* were quite different than the typical incentives of a disparate impact case where the employer seeks to prove a business necessity.[265]

Applying the institutional view to the algorithmic context suggests there might be distinctions between how public and private entities are treated.[266] If the algorithm developer is a legislative or executive government entity that might be affected by political motivations, efforts to address algorithmic bias might be more suspect.[267] The fact that most algorithms are developed by private companies (including those used by public entities), however, suggests that under the institutional reading, the Court should be less concerned with race-conscious bias mitigation conducted in the algorithmic context.

Thus, *Ricci* should not be interpreted as broadly prohibiting the use of bias mitigation post-training or post-deployment of algorithms. That said, the analysis above implies that the reasoning that motivated the *Ricci* case makes it very ambiguous how the Court would view an effort to proactively mitigate disparate impact by an algorithm. As Justice Scalia noted in his concurrence in *Ricci*, "the war between disparate impact and equal protection will be waged sooner or later, and it behooves us to begin thinking about how—and on what terms—to make peace between them."[268] This war that Scalia predicted might come in the form of litigation regarding algorithmic bias. The necessity, in the algorithmic context,

---

261. *Id.* at 1365.
262. *See id.*
263. *Id.* at 1366.
264. *Id.*
265. *See id.* at 1368.
266. *See id.* at 1376 n.164.
267. *See id.* at 1366 n.132.
268. Ricci v. DeStefano, 557 U.S. 557, 595–96 (2009) (Scalia, J., concurring).

of making precise and mathematical the objectives the algorithm is optimized for and the metrics on which the algorithm is evaluated implies a degree of clarity typically not found in anti-discrimination cases. This in turn requires much more clarity around what exactly can be done to address disparate impact without committing disparate treatment.

VII. BENEFITS OF CAUSAL INFERENCE FOR ALGORITHMIC BIAS MITIGATION

One promising path forward for engaging in algorithmic bias mitigation while respecting legal precedent is to look to causality. Causality is a key concept both for legal liability and for identifying algorithmic bias. Illegal discrimination is often defined in statutes as making decisions "because of" a protected class attribute.[269] In fact, the importance of causality came up in the proposed HUD rule (somewhat ironically given that the proposed algorithmic safe harbor had nothing to do with causality).[270] In *Texas v. Inclusive Communities*, the Court held that there must be a robust causal connection between the decision-making process and protected class attribute in order to establish disparate impact liability.[271] The *Texas* ruling emphasized that causality is ultimately a crucial question for a determination of illegal discrimination.[272] Referring to the lower court's decision, the Court stated, "if the ICP cannot show a causal connection between the Department's policy and a disparate impact—for instance, because federal law substantially limits the Department's discretion—that should result in dismissal of this case."[273] The Court explained that a "robust causality requirement" is needed to provide "adequate safeguards at the prima facie stage."[274] Otherwise, "disparate-impact liability might cause race to be used and considered in a pervasive way and 'would almost inexorably lead' governmental or private entities to use 'numerical quotas,' and serious constitutional questions then could arise."[275]

---

269. *See, e.g.*, Age Discrimination in Employment Act of 1967, 29 U.S.C. § 623(a) (2018); Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2018).

270. Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507, 2523 (2015).

271. *Id.* at 2523–24.

272. *See id.*

273. *Id.* at 2524 (citation omitted).

274. *Id.* at 2523.

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

The Court in this decision likely did not have in mind the tensions faced by the algorithmic fairness community, but in focusing on the importance of causality, the Court suggested a solution that would reconcile these legal and technical approaches to fairness.[276]

From a technical perspective, the intuition for causal methods is to impute what the counterfactual outcomes would have been had the individuals' protected class attributes (or the perception of these attributes by decision-makers) been different at a particular point in time.[277] Then the developer can compare these imputed outcomes with the actual outcomes to see if there is evidence of bias. For example, in a hiring discrimination case, a key question would be whether the individual would have been hired had they been (perceived to be of)[278] a different race or sex.[279] Of course, the most difficult aspect of these methods is how to impute the counterfactual

---

276. That said, other dicta in the *Inclusive Communities* case creates complications for the algorithmic context: "It must be noted further that, even when courts do find liability under a disparate-impact theory, their remedial orders must be consistent with the Constitution. Remedial orders in disparate-impact cases should concentrate on the elimination of the offending practice that 'arbitrar[ily] . . . operate[s] invidiously to discriminate on the basis of rac[e].' If additional measures are adopted, courts should strive to design them to eliminate racial disparities through race-neutral means." *Id*. at 2524 (alteration in original) (citations omitted). The Court thus is still very uncomfortable with approving any remedies that are not race-neutral, particularly to the extent quotas are implicated.

277. The fundamental problem of causal inference is that, in the real world, we can only see one version of events. For example, if I have a headache, take an aspirin, and my headache goes away an hour later, I do not directly observe whether I would still have the headache if I had not taken an aspirin. As we can never observe both "potential outcomes"—what happened when I took the aspirin and also what would have happened if I had not—the gold standard in causal inference is the randomized controlled experiment. The intuition behind the randomized controlled experiment is that by randomly assigning people to treatment versus control groups, we can isolate the effect of the treatment from other potential factors. When assessing potential discrimination, it can be challenging in practice to isolate whether a particular decision was made based on the demographic attributes of the individual or other factors that might distinguish them.

278. As will be discussed further below, it is very difficult to define the counterfactual of what would have happened if someone were of a different race or sex. Instead, what is more relevant for the purposes of determining discrimination is what would have happened if someone were perceived by the alleged discriminator to have been of a different protected class.

279. *See Tex. Dep't of Hous. & Cmty. Affs.*, 135 S. Ct. at 2523–24. Note that while the Court in *Price Waterhouse v. Hopkins*, 490 U.S. 228 (1989) established a motivating factor standard for causation in Title VII cases, the Court in *Gross v. FBL Financial Services, Inc.*, 557 U.S. 167 (2009) held that but-for causation is required in the context of the ADEA. Gross v. FBL Fin. Servs., Inc., 557 U.S. 167, 176 (2009); Price Waterhouse v. Hopkins, 490 U.S. 228, 249 (1989).

outcome. The fundamental problem of causal inference is one of missing data.[280] For example, for a binary feature, like someone having taken or not taken aspirin for a headache, we only ever observe one of the possibilities at a given time.[281] To address this impossibility, the gold standard of causal inference is the randomized controlled trial. The intuition of this approach is that if we can randomly assign people to separate treatment and control groups (and ensure balance on relevant pre-treatment variables across groups), then the average difference in outcomes between the treatment and control groups can be attributed to the treatment effect. Of course, we cannot randomly assign protected class membership to individuals, so causal analysis around such attributes has focused on the perception of these attributes. For example, if you have pairs of resumes that are very similar but differ only on the basis of gender, you can estimate the effect of gender discrimination by observing the differences in call-back rates for the female versus male resumes. Where experimental methods are not possible, additional modeling assumptions and techniques are necessary in order to estimate what the counterfactual outcomes would have been if someone's protected attribute were perceived differently.

Causality permits a distinction between using protected class variables to make the protected class designation more versus less salient to the decision-making process. The goal of some methods in the algorithmic fairness literature is not racial balancing but rather removing discriminatory effects in the data.[282] In a causal framework, fairness is conceived of as the lack of a difference between the observed outcome and the counterfactual outcome where the (perception of the) individual's protected class attribute is changed.[283] This aligns with legal conceptions of fairness: if but for the individual's protected class, the decision would have been different, then the individual was illegally discriminated against.[284]

---

280. *See* DONALD B. RUBIN, BASIC CONCEPTS OF STATISTICAL INFERENCE FOR CAUSAL EFFECTS IN EXPERIMENTS AND OBSERVATIONAL STUDIES 92 (2005).

281. *See id.*

282. *See* Madras, *supra* note 162, at 354–57.

283. *See* Matt Kusner et al., *Counterfactual Fairness*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 3–4 (2017).

284. *See But-for test*, CORNELL L. SCH. WEX LEGAL INFO. INST., https://www.law.cornell.edu/wex/but-for_test (last visited Mar. 29, 2021). Note that there are also motivating factor and mixed motive standards in anti-discrimination law for evaluating the role of a protected class attribute, but but-for causality is generally seen as a more stringent causal standard.

In this sense, it's possible to argue that a causal method that seeks to cancel out discriminatory trends in the data is still race-neutral even if the protected class variable is used in the algorithm.

To illustrate this, let us say that a company is designing an algorithm to help determine who it should promote. The company has a large amount of historical data about the performance and attributes of past employees and which ones got promoted. It discovers, however, that due to biased decisions made by its management in the past, women were promoted at much lower rates than men, even after taking into account their performance reviews, qualifications, and experience. One possible causal method to address this would be to impute what the promotions for women would have looked like if they had been treated like men. This could, for example, involve building a model of the decision-making process used for men and then applying that model to the data for women.[285] At this point, the employer could use this counterfactual data (where women were promoted using the same standards as men) to train a less biased model. Although this method would involve modifying the past promotion data based on a protected attribute, it would only do so in service of preventing the algorithm from learning from past biases rather than having the algorithm use race-conscious decision-making going forward.

In addition, causal methods can help distinguish between different sources of bias. A key first step in causal inference is specifying a point in time for a hypothetical intervention in order to define a counterfactual. For example, in the context of hiring, in order to define a counterfactual for whether a woman who was not hired would have been hired if she were male, you need to specify precisely the timing and nature of the counterfactual. "Would she have been hired if she were born male?" is a very different question from, "would she have been hired if the company perceived her to be a man?" The former counterfactual is much harder to estimate because much of her life would have been different if she had been born male—maybe she would have had different work experience, maybe she would have had a different major in college, maybe she would not have gone to college at all. In fact, it can even be difficult to define what it would mean for her to be the same person yet have

---

285. For an example of a similar approach, see generally Alice Xiang & Donald B. Rubin, *Assessing the Potential Impact of a Nationwide Class-Based Affirmative Action System*, 30 STAT. SCI. 297 (2015) (simulating admissions and educational outcomes for minority students under a class-based affirmative action system).

a different sex; some might argue the male counterfactual of her is fundamentally a different person, such that the question is undefined.[286] The latter counterfactual, though still difficult, is comparatively easier to estimate. In this counterfactual, everything is the same about the woman except how her sex is perceived by the company. A famous resume study, for example, examined this type of question about perception through experimentation.[287] By sending out identical resumes with different names suggesting different applicant demographics (e.g., Jamal versus Emily), researchers were able to quantify the effect of being perceived as being female or minority versus male or white.[288]

This key step of causal inference—specifying the intervention—can help to distinguish between different kinds of bias and different sources of bias. As the example above illustrates, specifying an intervention around how the hiring company perceived an applicant's sex would isolate the effect of any bias on the part of the hiring company, whereas specifying an intervention around the applicant's sex itself would capture not only any biases on the part of the hiring company but also all other biases in society that would affect the life paths of an individual on the basis of their sex.

---

286. *See generally* D. James Greiner & Donald B. Rubin, *Causal Effects of Perceived Immutable Characteristics*, 93 REV. ECON. & STAT. 775 (2011) (discussing the challenges of evaluating the causal effect of an immutable characteristic like race or sex and proposing the solution of shifting the focus to the perception of the immutable trait).

287. *See generally* Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004) (studying race in the labor market by sending fictitious resumes—with African-American- or white-sounding names—to help-wanted ads in Boston and Chicago newspapers).

288. *See id.* at 991–93. Fair housing investigators use a conceptually similar experimental technique to gather evidence for housing discrimination cases. *See* Teresa C. Hunter & Gary L. Fischer, *Fair Housing Testing—Uncovering Discriminatory Practices*, 28 CREIGHTON L. REV. 1127, 1132–34 (1995). By having trained undercover testers of different races reach out to housing providers, they are able to determine how housing providers treat similar individuals differently based on their race. *Id.* at 1132–33. In this case, because the testers were paired such that each pair presented similar financial profiles and made identical requests for homes in the same areas, they were able to eliminate some major confounding factors and gain empirical evidence for whether person A would have been treated differently if they were the race of person B. *See generally Fair Housing Testing Program*, U.S. DEP'T OF JUST. (Mar. 26, 2021), https://www.justice.gov/crt/fair-housing-testing-program-1 (outlining the Department of Justice unit that brings suit to enforce the fair housing act).

Moreover, one of the major advantages of a causal inference framing is that it forces the individual doing the analysis to explicitly state the assumptions they are using. Because causal inference is trying to do something that is fundamentally impossible—comparing what happened in the real world with what would have happened in a counterfactual world with different conditions—causal inference relies on the validity of its assumptions.[289]

Clearly stating the assumptions of a causal inference analysis can provide courts with a starting point for determining the reasonableness of the inference. For example, if a jurisdiction were concerned that there was bias in the data it was using for training its recidivism risk assessment tools and thus wanted to make race-conscious adjustments to its algorithm to address this bias, the jurisdiction would need to clearly specify a causal intervention that would address the bias at issue. If the primary bias issue were disproportionately high rates of policing in minority neighborhoods, then the jurisdiction would need to build a model estimating what the arrest rates across neighborhoods would have looked like if there had not been such disproportionate rates of policing. A court could then evaluate whether correcting for past disproportionate rates of policing would be a compelling interest to justify race-conscious remedial action. Provided it is, the court could interrogate the assumptions used in the model to ensure that the counterfactual arrest rates are reasonable.

Finally, causal conceptions of algorithmic fairness can make it harder to conceal algorithmic bias. Studies have shown that non-causal approaches to providing explanations about how ML models are making decisions can be manipulated to conceal a reliance on protected class variables.[290] This finding is related to the Rashomon Effect (as discussed above in Part II.A), whereby models with

---

289. One of the most commonly used set of assumptions in causal inference is Stable Unit Treatment Value Assumption. *See* RUBIN, *supra* note 280, at 6. This implies that the potential outcomes for one unit should not be affected by the assignment of treatments to other units. *See id.* An example where this assumption might not hold would be if you were trying to evaluate the effect of aspirin on headaches, but whether an individual's headache disappeared depended in part on whether other individuals in the study got aspirin. This would potentially be the case if all the patients were in a room together and those with headaches complained loudly, giving others headaches. In this case, whether aspirin cures Patient A's headache depends in part on whether Patient B got aspirin.

290. *See generally* Dimanov et al., *supra* note 89 (identifying the ways in which particular algorithmic methods can conceal unfairness).

different weights on features can still perform similarly.[291] As a result, if algorithmic discrimination is conceived of as a strong reliance on protected class attributes or close proxies, it is possible to achieve biased outcomes while masking a model's reliance on these variables. If the model weights instead reflect underlying causal relationships, however, this type of manipulation is less feasible as the underlying causal relationships remain fixed.

A. Causal Inference in the Machine Learning Literature

This proposal to steer the focus of algorithmic discrimination toward causality ironically might be more radical from a ML perspective than a legal perspective. Courts already have some degree of experience in evaluating causal reasoning as current anti-discrimination cases frequently rely on statistical evidence of causality.[292] Although it remains an open question how well-equipped courts are to interpret this kind of expert witness testimony, causal methods would at least allow courts to evaluate algorithmic discrimination cases using the same framework as for human discrimination cases.

In contrast, ML, unlike most sciences, has historically focused on predictive accuracy instead of causal inference.[293] As a result, the field has paid relatively little attention to questions of whether the output of models reflects correlations or causal relationships in the underlying training data. With the recent push for fairer and more transparent algorithms in the popular discourse, however, causality is increasingly a topic of discussion in the ML community.[294] Indeed, many of the problems identified by the subfield of algorithmic fairness stem from a lack of causal interpretations of ML algorithms. For example, one of the challenges hampering efforts to develop techniques to explain "black box" algorithms is the fact that existing explainability techniques by and large do not provide causal

---

291. *See supra* note 91 and accompanying text.

292. *See* Thomas J. Campbell, *Regression Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 STAN. L. REV. 1299, 1299 (1984). *See generally* Ann Morning & Daniel Sabbagh, *From Sword to Plowshare: Using Race for Discrimination and Antidiscrimination in the United States*, 57 INT'L SOC. SCI. J. 57 (2005) (looking at use of statistics, but also making broader arguments about the racist history of data acquisition).

293. *See* Galit Shmueli, *To Explain or to Predict?*, 25 STAT. SCI. 289, 292 (2010).

294. *See generally* SOLON BAROCAS ET AL., *Causality*, *in* FAIRNESS IN MACHINE LEARNING, *supra* note 38 (discussing the relevance of causality for fair ML).

explanations.[295] As a result, they are vulnerable to spurious correlations.[296]

It is important to note, however, that using causality as a standard to reconcile efforts to address algorithmic bias while avoiding legally problematic racial balancing does not imply that the counterfactual fairness methods common in the current literature are the solution. While the counterfactual fairness literature focuses on what would have needed to change in the data in order to achieve a different decision by the algorithm,[297] the causality literature focuses on the effects of changes in reality, not simply in the data.[298]

The existing counterfactual fairness literature generally uses the structural causal model framework[299] under which each feature is represented as a node and causal relationships are represented by arrows, creating what is called a "directed acyclic graph" (DAG).[300] Under this framework, the first step is for the ML developer to draw the DAG summarizing how the variables relate to each other, with arrows denoting causal relationships.[301] In one line of methods, the next step would be to determine which causal pathways might be problematic from a fairness perspective.[302] For example, the developer might determine that a direct causal pathway between race and likelihood of arrest is problematic.[303] At that point, the developer would use methods to "trim" the graph, essentially cancelling out the role the problematic pathways would otherwise play.[304]

---

295. *See* Bhatt & Xiang et al., *supra* note 52, at 653.

296. *Id.* at 651.

297. *See* Kusner et al., *supra* note 283, at 3–6.

298. *See* Madras et al., *supra* note 162, at 349 ("To understand how past decisions may bias a dataset, we first must understand how sensitive attributes may have affected the generative process which created the dataset, including the (historical) decision makers' actions (treatments) and results (outcomes). Causal inference is well suited to this task: Because we are interested in decision-making rather than classification, we should be interested in the causal effects of actions rather than correlations. Causal inference has the added benefit of answering counterfactual queries: What would this outcome have been under another treatment? How would the outcome change if the sensitive attribute were changed, all else being equal?").

299. *See* JUDEA PEARL, CAUSALITY: MODELS, REASONING, AND INFERENCE 35 (2d ed. 2009).

300. *See id.* at 44.

301. *Id.*

302. *See* Kusner et al., *supra* note 283, at 5.

303. *See id.* at 4–5.

304. *See id.* at 12, 15.

One challenge to this approach is that constructing the DAG and determining which "pathways" are problematic versus acceptable requires significant domain expertise and subjective judgment. In addition to knowing how each variable is causally related to the other variables, the model developer needs to be aware of potential confounding variables not present in the dataset.[305] Confounding variables are variables that influence both an input variable and the outcome variable such that they distort the relationship between these variables, resulting in spurious correlations.[306] Drawing the DAG is thus very difficult, and the DAG itself is highly contestable.

What constitutes a problematic pathway is also highly debatable. The existing literature generally considers a causal pathway between a protected class variable and the outcome variable to be problematic as long as there is not a "resolving variable" between the two.[307] A resolving variable is one that is considered perfectly reasonable and relevant for the decision at hand. What a resolving variable is, however, is a matter of judgment. For example, on the one hand, some might consider test scores to be a resolving variable in the context of school admissions if they believe that test scores are an accurate metric for ability. On the other hand, the social science literature indicating that test scores can be biased against minorities and immigrants, in the sense of less accurately reflecting their abilities,[308] would suggest that test scores perhaps should not be considered a resolving variable.[309] It is hard to say how courts would interpret these methods, but it is likely they would be skeptical of an

---

305. *See* PEARL, *supra* note 299, at 78.

306. *See id.*

307. *See* Niki Kilbertus et al., A*voiding Discrimination Through Causal Reasoning*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 3 (2018); *see also* Razieh Nabi & Ilya Shpitser, *Fair Inference on Outcomes*, *in* PROCEEDINGS OF THE 32ND ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE (AAAI) CONFERENCE ON ARTIFICIAL INTELLIGENCE 1931, 1931–39 (2018) (proposing a counterfactual fairness method that removes direct causal influences, not mitigated by resolving variables); Lu Zhang et al., *A Causal Framework for Discovering and Removing Direct and Indirect Discrimination*, *in* PROCEEDINGS OF THE 26TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE 3929, 3933 (2017) (proposing an approach to remove path-specific discrimination).

308. *See* Richard Delgado, *Standardized Testing as Discrimination: A Reply to Dan Subotnik*, 9 U. MASS. L. REV. 98, 103–04 (2014); Hutchinson & Mitchell, *supra* note 41 (surveying history of fairness in standardized testing).

309. In other words, this would imply the existence of a causal pathway between race/immigrant status and test score and between test score and school admission should be considered problematic.

algorithm developer's attempts to categorize problematic versus unproblematic pathways related to protected class attributes because the degree and nature of a causal relationship that constitutes illegal discrimination is ultimately a legal question.[310] Thus, existing counterfactual fairness methods should be distinguished from causal methods as a whole given that they do not necessarily have the desirable properties of the latter and might raise additional legal complications.

As this Section has summarized, there are many benefits to causal approaches to algorithmic bias mitigation. By putting the focus squarely on eliminating the causal relationships between the protected class variable and the algorithmic decisions, causal methods would arguably yield decision-making processes that are neutral to the protected class attribute. In doing so, they could side-step many of the legal concerns associated with methods that approximate balancing across protected classes. That said, pushing for causal methods for measuring and mitigating algorithmic bias is only the first step. Causality is very challenging to show without conducting experiments.[311] In most algorithmic contexts, observational rather than experimental data is used, so causal inference will require additional assumptions.[312] Courts in turn

---

310. *See generally* Hutchinson & Mitchell, *supra* note 41 (recounting the history of legal challenges related to educational testing).

311. The potential outcomes framework of causal inference originated from concepts key to randomized experiments, and its application to observational data centers on modeling the assignment mechanism between "treatment" and "control" groups. *See* GUIDO W. IMBENS & DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES: AN INTRODUCTION 23–30 (2015); *see also* Madras et al., *supra* note 162, at 349 ("[T]he presence of *hidden confounders*—unobserved factors that affect both the historical choice of treatment and the outcome—often prohibits the exact inference of causal effects. Additionally, understanding effects at the individual level can be especially complex, particularly if the outcome is non-linear in the data and treatments.").

312. *See, e.g.*, Madras et al., *supra* note 162, at 353 ("Given some treatment T and outcome Y, the classic 'no hidden confounders' assumption asserts that the set of observed variables O blocks all backdoor paths from T to Y."). The starkness of these assumptions in the context of using causal methods to measure discrimination has been criticized, with some arguing for the importance of "thick[er]" conceptions of the meaning of protected class attributes. Issa Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination*, 113 NW. U. L. REV. 1163, 1163 (2019); *see* Lily Hu & Issa Kohler-Hausmann, *What's Sex Got to Do with Fair Machine Learning?*, *in* PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 2 (2019).

should interrogate the assumptions underlying the counterfactuals to ensure that they are reasonable.[313]

B. Shortcomings of Other Approaches

This Section discusses other potential approaches that have been proposed for using protected class variables to mitigate algorithmic bias. The first method, disparate learning processes, has been proposed in the literature as a potential solution for avoiding legal complications with the use of protected class variables, but the approach primarily addresses the issue through obscuring how protected class variables are used to develop the algorithm. The second set of methods, group fairness methods, are likely to raise equal protection concerns, as they are similar to racial balancing. The third set of methods, individual fairness methods, were inspired by a desire to avoid the potential legal issues with group fairness methods. Similar to disparate learning processes, however, they achieve legal compatibility through obscuring the extent to which protected class attributes are considered in the bias mitigation process, albeit through different means. As a result, none of these approaches are advisable from a transparency perspective and will likely still suffer from incompatibilities with anti-discrimination law.

1. Disparate Learning Processes

One proposal to address the potential legal impediments facing algorithmic bias mitigation is to simply use the protected class variables at training time but not at deployment time.[314] The rationale is that this would prevent the algorithm from actually using the protected class variable as a feature, so if the algorithm itself and not its training process were examined, it would appear to not consider the protected class variable.[315] While this approach might save some algorithms in practice, there are both technical and legal reasons to be wary of this approach. First, from a technical perspective, such approaches can be suboptimal both in terms of accuracy and fairness.[316] Second, these approaches do not address

313. *See, e.g.*, Kohler-Hausmann, *supra* note 312, at 1165–66.

314. *See* Alekh Agarwal et al., *A Reductions Approach to Fair Classification*, *in* PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 1, 1–3 (2018); Harned & Wallach, *supra* note 26, at 617, 635.

315. *See* Harned & Wallach, *supra* note 26, at 639–40.

316. *See* Lipton et al., *supra* note 16, at 4.

the ways in which including the protected class variable can improve both accuracy and fairness.[317] Third, from a legal perspective, if courts did interpret the presence of protected class variables as being evidence of algorithmic discrimination, for consistency, they might similarly consider the presence of such variables in the training data as evidence of discrimination. As discussed above in Part II.A, from a technical perspective, the presence or absence of protected class variables in an algorithm provides little evidence regarding its potential biases.[318] Thus, if courts bluntly viewed the presence of such variables as suspicious, it is unclear why they would draw a distinction between whether the variables appear in the training data or model itself.

2. Group Fairness Methods

The most common types of methods proposed by the ML community to address algorithmic bias are known as "group fairness" methods, but these methods would likely be fraught from a legal perspective as they are similar to racial balancing. Group fairness calls for proportionality across a specific metric of interest, such as outcomes, false positive rates, or false negative rates for different demographic groups.[319] The advantages of this approach include that it is very easy to quantify and is relatively intuitive and transparent.[320] For example, when evaluating the fairness of an employment promotion algorithm, it is natural to ask what proportion of women versus men the algorithm recommended for promotion. The baseline assumption of these metrics is that in a "fair" world, the false positive and false negative rates would be equalized across demographic groups, and/or outcomes would be proportional to the group's representation in the general population.[321]

The literature in this area thus seeks to adjust algorithms in order to conform to one of these notions of fairness. For example, methods have been developed to eliminate all correlation between

---

317. *See* discussion *supra* Part II.A. See also Lipton et al., *supra* note 16, at 3–5 for proof on optimality of treatment disparity.

318. *See* discussion *supra* Part II.A.

319. *See, e.g.*, Cynthia Dwork et al., *Decoupled Classifiers for Fair and Efficient Machine Learning*, *in* PROCEEDINGS OF THE 1ST CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 2 (2017).

320. *See generally* Ho & Xiang, *supra* note 172.

321. *See generally id.*

data used in training the model and protected class variables.[322] This in turn guarantees that any predictions using the data will be proportional across the groups, also known as "demographic parity" or "statistical parity" in the ML literature.[323] If this method were applied to a dataset for training a risk assessment tool to predict re-arrest, the same percentage of black defendants would be predicted to recidivate as white defendants. This would be very similar to a quota system because the number of black versus white defendants in high versus low-risk categories would be predetermined. As discussed previously, the Court in *Bakke* struck down the use of racial quotas in higher education on equal protection grounds.[324]

Other methods in this literature seek to equalize false positive rates, false negative rates, or both.[325] Equalizing false positive or negative rates across demographic groups can be less controversial than equalizing outcomes as it addresses how well the algorithm performs in correctly predicting outcomes.[326] In fact, one proposed algorithmic bias definition is "equal opportunity," which requires that those in the "advantaged" group have an equal opportunity to obtain the beneficial treatment, regardless of their protected class.[327] In the recidivism context, this would mean that among those who do not recidivate (the "advantaged group"), the probability of the algorithm classifying them as "low-risk" (the beneficial treatment) does not differ by the defendant's race.

While equalizing false positive and/or false negative rates across demographic groups would not constitute use of a quota, doing so might still raise legal flags as improper racial balancing. After all, from a mathematical perspective, differing false positive or false negative rates across demographic groups is related to differing baselines across the groups. In fact, there is an impossibility theorem that states that it is impossible to achieve error rate balance—equalized false positive rates and false negative rates across demographic groups[328]—while maintaining calibration as long

322. *See* Lum & Johndrow, *supra* note 93, at 1–2.

323. *See* Corbett-Davies & Goel, *supra* note 19 (defining statistical parity); Hardt et al., *supra* note 16, at 1–6 (defining and explaining demographic parity).

324. Regents of the Univ. of Cal. v. Bakke, 438 U.S. 265, 271 (1978); *id.* at 319 (Powell, J., concurring).

325. Hardt et al., *supra* note 16, at 8.

326. *See generally id.* at 2 (explaining how this method aligns fairness with the central goal of creating more accurate predictors).

327. *See id.* at 4.

328. *See id.* at 8. "Error rate balance" is also referred to as "disparate mistreatment." *See* Zafar et al., *supra* note 16, at 1.

as there are differing baselines across the demographic groups.[329] Calibration implies that a given score has the same interpretation across demographic groups.[330] A 70% risk score for a white defendant would thus imply the same thing (70% probability of re-arrest) as a 70% risk score for a black defendant.[331]

To illustrate the impossibility theorem in the context of recidivism risk assessment tools, the false positive rate is the proportion of individuals who will not recidivate whom the algorithm incorrectly predicts will recidivate, and vice versa for the false negative rate.[332] The impossibility theorem shows that as long as there are different baselines between demographic groups and the model is not perfectly predictive, it is impossible to have a model that achieves both of these metrics.[333] Recidivism data, for example, usually show higher rates of re-arrest for black defendants than white defendants.[334] As a result, if an algorithm is well-calibrated, the false positive and false negative rates will not be equal.

Thus, group fairness methods are very similar in principle to racial balancing, which will likely create legal complications with the use of these methods. Fundamentally, group fairness methods operate from a premise that in a fair world, metrics of interest would be equal or proportional across demographic groups. Moreover, the biases they address stem directly from differing baselines between demographic groups in the data used to train the algorithm.

3. Intersectionality

Another critique of group fairness approaches is that they struggle with accounting for intersectionality, such that methods might satisfy group-level notions of fairness while violating individual-level notions of fairness.[335] This critique is also a major

---

329. *See* Pleiss et al., *supra* note 62.

330. *See* Corbett-Davies & Goel, *supra* note 19, at 6.

331. *See id.*

332. *See* Pleiss et al., *supra* note 62, at 1 (explaining the false-positive error of the Angwin ProPublica study). *See generally* Angwin, *supra* note 1 (explaining the bias in recidivism prediction software).

333. *See* Kleinberg et al., *supra* note 61, at 8.

334. *See, e.g.*, Matt Clarke, *Long-Term Recidivism Studies Show High Arrest Rates*, PRISON LEGAL NEWS (May 3, 2019), https://www.prisonlegalnews.org/news/2019/may/3/long-term-recidivism-studies-show-high-arrest-rates/.

335. *See generally* Kearns et al., *supra* note 243 (explaining the shortcomings of common statistical fairness definitions and proposing more adaptable alternative definitions).

motivation for "individual fairness" techniques, which I discuss in the next Subsection.

For a simple illustrative example of this issue, assume that everyone has a 50% chance of recidivism, and we have equal numbers of white men, white women, black men, and black women in the population. If an algorithm automatically detained all white women and black men, and automatically released all black women and white men, then for any group fairness metric along the dimensions of race or sex, the algorithm would appear to be fair even though it is actually treating certain subgroups very unfairly based on their race and gender. The false positive rate for black defendants would be 50%, for example, because half of those detained (i.e., the black *men*) would be incorrectly classified, and the false positive rate for white defendants would also be 50% because half of those detained (i.e., the white *women*) would be incorrectly classified. The same logic applies when comparing false positive rates across sex.

A solution to this problem within the group fairness framing is to define each group as an intersectional subgroup—black man, black woman, white man, white woman—and compare the relevant group fairness metric among these subgroups. This would involve ensuring that false positive rates are constant across each of these four groups. This approach becomes more difficult, however, as the number of intersected features increases. With ten binary demographic features, for example, performance would have to be equalized across 1,024 (2^{10}) subgroups, adding significant constraints to the model.[336]

This issue is related to the more general challenge in algorithmic fairness of how to account for intersectionality. The fact that traits are not simply additive in their effects on individuals' lived experience and instead intersect in ways that create unique dynamics implies the need for variables that capture not only individual attributes but also their intersection. In the context of anti-discrimination law, Kimberle Crenshaw documented challenges that black women have faced in succeeding in cases where they were discriminated against, not on the basis of their race or sex alone but

336. *See id.* at 3 n.2. In addition, attempting to equalize performance across so many subgroups would create data sparsity problems, as it is unlikely that each of the 1,024 subgroups would be sufficiently well-represented in the data for the model to learn about the unique dynamics for each subgroup.

on the basis of their intersected identity.[337] As one example, Crenshaw cited *DeGraffenreid v. General Motors Assembly Division*,[338] where black women sued General Motors for not hiring any black women prior to 1964 and laying off all of the black women it hired after 1970 in a seniority-based layoff.[339] The court granted summary judgment to the defendant given that the plaintiffs could not show that women generally or black people generally were being discriminated against at General Motors;[340] the company had, for example, hired white women prior to 1964.[341] The court concluded that:

> The legislative history surrounding Title VII does not indicate that the goal of the statute was to create a new classification of 'black women' who would have greater standing than, for example, a black male. The prospect of the creation of new classes of protected minorities, governed only by the mathematical principles of permutation and combination, clearly raises the prospect of opening the hackneyed Pandora's box.[342]

Just as the court in *DeGraffenreid* essentially rendered the experiences of black women invisible in the eyes of anti-discrimination law,[343] the same issue can plague efforts to correct for algorithmic bias that do not take intersectionality into account. Similar to the court's concern that permuting and combining protected class attributes would open a Pandora's box, there are

---

337. Kimberle Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine*, *Feminist Theory and Antiracist Politics*, 1989 U. CHI. LEGAL F. 139, 140–50.

338. 413 F. Supp. 142 (E.D. Mo. 1976).

339. Crenshaw, *supra* note 337, at 141.

340. *Id.* (citing *DeGraffenreid*, 413 F. Supp. at 143).

341. *Id.* at 142 (citing *DeGraffenreid*, 413 F. Supp. at 144).

342. *DeGraffenreid*, 413 F. Supp. at 145.

343. Crenshaw further emphasizes the ways in which a lack of intersectional consideration marginalizes black women. *See* Crenshaw, *supra* note 337, at 150 ("*DeGraffenreid*, *Moore* and *Travenol* are doctrinal manifestations of a common political and theoretical approach to discrimination which operates to marginalize Black women. Unable to grasp the importance of Black women's intersectional experiences, not only courts, but feminist and civil rights thinkers as well have treated Black women in ways that deny both the unique compoundedness of their situation and the centrality of their experiences to the larger classes of women and Blacks.").

technical challenges to considering all the unique ways in which protected class variables intersect and interact.[344] Perhaps as a result, this issue remains underexplored in the algorithmic fairness literature,[345] which creates a risk that algorithmic decision-making and efforts to audit such decisions for bias might be similarly blind to intersectional experiences.

4. Individual Fairness

Individual fairness has also been proposed as a way to account for intersectionality and achieve fairer algorithmic decision-making without engaging in legally prohibited racial balancing. Individual fairness is inspired by the idea that similar individuals should be treated similarly.[346] Given a similarity metric, computer scientists have designed methods to ensure that the difference in the algorithm's predictions for individuals is limited by the scaled difference between the individuals' similarity metric.[347] For example, if there is an algorithm that allocates students into AP versus honors versus regular versus remedial classes, a possible similarity metric would be the weighted sum of these students' GPA and PSAT scores. An individual fairness constraint would say that students with similar total scores should not have radically different outcomes: if you have two students with very similar scores, one should not be placed in the AP class while the other is placed in the remedial class. The advantage of this method over group fairness is that the similarity metric can include as many features as the model developer would like, and it provides more flexibility in how the features reflect similarity.

The challenge with this method, however, is determining what the similarity metric should be. The similarity metric will reflect whatever biases exist in the data used to construct the metric. In the example above, the similarity metric would reflect any biases in GPA

---

344. *See* Foulds et al., *supra* note 63, at 2.

345. See *id.* at 1 for a preliminary operationalization of intersectional fairness.

346. *See* Dwork et al., *supra* note 72.

347. *See id.* at 1; *see also* Matthew Joseph et al., *Fairness in Learning: Classic and Contextual Bandits*, *in* PROCEEDINGS OF THE 30TH ANNUAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 3 (2016) (operationalization of individual fairness in bandit problems); Christopher Jung et al., *An Algorithmic Framework for Fairness Elicitation*, *in* PROCEEDINGS OF THE 2ND ANNUAL SYMPOSIUM ON FOUNDATIONS OF RESPONSIBLE COMPUTING 1, 1–3 (2020) (operationalization of individual fairness through value elicitation).

and PSAT scores. If we were worried that the PSAT is not an accurate reflection of the ability of minority students, then the approach above would not address these bias concerns.

As a result, many of the proposals for how individual fairness could be implemented involve creating new datasets. For example, in one implementation of this method, human "judges" (in practice, these might be workers on Amazon's Mechanical Turk or college students recruited in a study) must decide if individuals should be treated similarly or not based on a set of features they are shown about each individual.[348]

It is limiting, however, that the human judges are only shown the data available to train the algorithm. For example, in order to generate a similarity metric for a risk assessment algorithm, a human judge might be shown the defendants' number of prior convictions, whether the offenses were violent, and a few other statistics.[349] Because one of the advantages of human judges over machine judges is that human judges can consider less quantifiable data, such as the nuances of a defendant's circumstances or prior convictions, this method does not leverage that advantage as the human judges would only have access to a few quantified pieces of information about each individual.

Moreover, the default for any algorithm is to treat individuals with similar predicted probabilities of recidivism similarly, so it is not clear that this approach would actually yield fairer outcomes than not using any bias mitigation method at all. As the human judges in this context are limited to data that would otherwise be available to train an algorithm, the only advantage of individual fairness is that the human judges should have a better understanding of how to weigh and contextualize different features.[350] In terms of bias, this would mean, for example, that the human judge should have an understanding that certain neighborhoods are subjected to higher rates of policing, such that an individual living in a heavily-policed neighborhood with five prior arrests might have only committed three crimes on average whereas an individual living in a less-heavily-policed neighborhood with five prior arrests is more likely to have actually committed five crimes.

---

348. *See* Jung et al., *supra* note 347, at 6–7.

349. *See id.* at 18–19.

350. *See id.* at 18 ("[S]ubjects could choose to ignore demographic factors or criminal histories entirely if they liked, or a subject who believes that minorities are more vulnerable to overpolicing could discount their criminal histories relative to Caucasians in their pairwise elicitations.").

The problem, however, is that the judge would only be able to make this more informed determination if the judge had access to sensitive data about the individuals. Showing the human judges the individuals' neighborhood, race, sex, age, etc., however, would raise concerns that the decisions are being made because of protected class attributes. Moreover, implicitly the hope of this method is that the human judges will be able to correct for the biases in criminal justice data by rebalancing the data in their heads. That not only would constitute a form of racial balancing or affirmative action, but it also would make this method roughly equivalent to simply using a group fairness method (e.g., striving for proportional outcomes across groups).

Even if these concerns were assuaged, there is no guarantee that human judges would use these protected class attributes in a way that would achieve fairer outcomes. After all, the fundamental problem with existing biased algorithms is that they were trained on data from biased historical human decisions.[351] Finding "unbiased" human judges has always been a fundamental challenge. Thus, this approach would not necessarily address the legal barrier to using protected class variables in decision-making and would not necessarily lead to less biased decision-making.

CONCLUSION

Algorithmic bias is now widely recognized as a serious concern—the AI principles of major technology companies virtually all include fairness as a key principle[352]—but methods to mitigate it are rarely used in practice.[353] This Article has addressed both the technical and related legal challenges that hamper the application of bias mitigation techniques. For now, cases about mitigating algorithmic bias have yet to be litigated, so there is still ambiguity in terms of how courts will apply existing anti-discrimination law. Given the growing pervasiveness of ML, however, as well as the growth in public awareness around algorithmic bias, it is likely these issues will be litigated in the near future. When they are, the mathematical

---

351. *See* Zafar et al., *supra* note 16, at 1.

352. *See, e.g.*, *Artificial Intelligence at Google: Our Principles*, GOOGLE AI, https://ai.google/principles/ (last visited Feb. 8, 2021) (emphasizing the need to "[a]void creating or reinforcing unfair bias"); *Microsoft AI Principles*, MICROSOFT, https://www.microsoft.com/en-us/ai/responsible-ai (last visited Feb. 8, 2021) ("Fairness[:] AI systems should treat all people fairly.").

353. Andrus et al., *supra* note 13, at 9.

nature of algorithmic decision-making might force the Court to draw more precise contours around anti-classification and anti-subordination doctrines.

Ironically, some of the most obvious applications of existing law to the algorithmic context would enable the proliferation of biased algorithms while rendering illegal efforts to mitigate bias. The conflation of the presence of protected class variables with the presence of bias in an algorithm or its training data is a key example of this: in fact, removing protected class variables or close proxies does not eliminate bias but precludes most techniques that seek to counteract it.

Causal inference methods provide one promising solution for reconciling legal and technical approaches to mitigating algorithmic bias given that they can help delineate between efforts to make protected class attributes more versus less salient for decision-making. Focusing on causal relationships also alleviates some of the concerns around bias from proxy variables. Causality is already a key concept in anti-discrimination law, so adopting causal frameworks for thinking about algorithmic fairness can be one path toward greater legal and technical compatibility.

There is no simple solution to algorithmic bias from the standpoint of either lawmakers or ML practitioners. One cannot simultaneously readdress historical biases reflected in data without taking into account the protected class designations along which people have been discriminated against. The sobering reality is that continued public discourse around algorithmic bias will likely raise broader, more intractable questions about fairness in society, similar to those that have made affirmative action so polarizing. Further dialogue between the legal and ML communities is thus necessary to provide the tools and frameworks to deploy algorithms in a fair and responsible manner.