ALGORITHMIC FAIRNESS, ALGORITHMIC DISCRIMINATION

THOMAS B. NACHBAR*

ABSTRACT

There has been an explosion of concern about the use of computers to make decisions affecting humans, from hiring to lending approvals to setting prison terms. Many have pointed out that using computer programs to make these decisions may result in the propagation of biases or otherwise lead to undesirable outcomes. Many have called for increased transparency and others have called for algorithms to be tuned to produce more racially balanced outcomes. Attention to the problem is likely to grow as computers make increasingly important and sophisticated decisions in our daily lives.

Drawing on both the computer science and legal literature on algorithmic fairness, this paper makes four major contributions to the debate over algorithmic discrimination. First, it provides a legal response to a recent flurry of work in computer science seeking to incorporate "fairness" in algorithmic decision-makers by demonstrating that legal rules generally apply in the form of side constraints, not fairness functions that can be optimized. Second, by looking at the problem through the lens of discrimination law, the paper recognizes that the problems posed by computational decision-makers closely resemble the historical, institutional discrimination that discrimination law has evolved to control, a response to the claim that this problem is truly novel because it involves computerized decision-making. Third, the paper responds to calls for transparency in computational decision-making by demonstrating how transparency is unnecessary to providing accountability and that discrimination law itself provides a model for how to deal with cases of unfair algorithmic discrimination, with or without transparency. Fourth, the paper addresses a problem that has divided the literature on the topic: how to correct for discriminatory results produced by algorithms. Rather than seeing the problem as a binary one, I offer a third way, one that disaggregates the process of correcting algorithmic decision-makers into two separate decisions: a decision to reject an old process and a separate decision to adopt a new one. Those two decisions are subject to different legal requirements, providing added flexibility to firms and agencies seeking to avoid the worst kinds of discriminatory outcomes.

---

* Professor of Law, University of Virginia School of Law. I would like to thank Rebecca Crootof, David Evans, Nikolas Guggenberger, Debbie Hellman, Aziz Huq, James T. Nachbar, Richard Primus, Fred Schauer and participants at the Yale Information Society Project conference on (Im)Perfect Enforcement for helpful comments and suggestions. I am also indebted to Thomas Barnett-Young, Aidan Coleman, and Kurt Swalander for excellent research assistance.

Examples of disparate outcomes generated by algorithms combined with the novelty of computational decision-making are prompting many to push for new regulations to require algorithmic fairness. But, in the end, current discrimination law provides most of the answers for the wide variety of fairness-related claims likely to arise in the context of computational decision-makers, regardless of the specific technology underlying them.

INTRODUCTION 511
I. FAIRNESS AND DISCRIMINATION IN COMPUTATIONAL DECISION-MAKING 516
A. Computational Decision-making 517
B. The Problem of Fairness 523
1. Fairness as an Essentially Contested Concept 523
2. Fairness as a Side Constraint 525
C. Legal Fairness Relevant to Algorithms: Discrimination Law 527
II. DISCRIMINATION LAW AND ALGORITHMIC DISCRIMINATION 528
A. Disparate Treatment, Disparate Impact, and Disparate Outcome 530
B. Outcome versus Justification in Discrimination Law 534
1. The Limited Role of Outcomes (or Impacts) in Discrimination Law 534
2. Discrimination Law as a Side Constraint 541
C. Translating Discrimination Law to Algorithmic Discrimination 542
III. IMPLICATIONS OF DISCRIMINATION LAW FOR ALGORITHMIC DECISION-MAKING AND IMPLICATIONS OF ALGORITHMIC DECISION-MAKING FOR DISCRIMINATION LAW 543
A. Keeping Side Constraints on the Side 544
B. Transparency, Accountability, and Liability 544
C. Dealing with Disparate Outcomes 548
1. The Practicalities of (Remedying) Discrimination in Algorithmic Decisionmakers 549
2. Disaggregating Algorithmic Affirmative Action 552
CONCLUSION 556

INTRODUCTION

In May of 2016, ProPublica, "an independent, nonprofit newsroom that produces investigative journalism with moral force,"[1] announced that "[t]here's software used across the country to predict future criminals. And it's biased against blacks."[2] The software, COMPAS, was being widely used in criminal justice systems throughout the country. The COMPAS software did not use the race of the offender in making recidivism predictions, but it did seem to provide racially disparate results anyway, demonstrating that racial disparity can exist in algorithms even if they do not explicitly make racial classifications. The possibility that racial bias could be mechanically systematized in the criminal justice system through computer software rightly prompted considerable alarm.

It did not take long, though, for doubts to arise about the nature of the "bias" reported on by ProPublica. That fall, a *Washington Post* story written by PhD students and professors in engineering and computer science pointed out that, mathematically, every method of classification is biased in some regard.[3] ProPublica had pointed out one form of bias: that blacks who did not reoffend were more likely to be classified as likely reoffenders than whites.[4] But COMPAS's maker, Northpoint, pointed out that, within each risk each category, blacks and whites were equally likely to reoffend, which Northpoint considered a better measure of bias.[5]

Both ProPublica and the *Washington Post* contributors, and I think probably Northpoint and the judges who use COMPAS, are all trying to make sure COMPAS provides "fair" results.[6] As Deborah Hellman rightly points out, the COMPAS controversy raises a conflict of forms of fairness that requires us to prioritize one form of fairness over another, which is a comparison that must take place on normative, not

---

1. *About Us*, PROPUBLICA, https://www.propublica.org/about [https://perma.cc/88HH-8FTQ] (last visited March 31, 2021).

2. Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/B9HF-PPS5].

3. Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.e7fdb765674c [https://perma.cc/C7SC-2NK3].

4. Angwin et al., *supra* note 2.

5. *See* Corbett-Davies et al., *supra* note 3.

6. Angwin et al., *supra* note 2 ("If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be *fairer* and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it's wrong in one direction, a dangerous criminal could go free. If it's wrong in another direction, it could result in someone *unfairly* receiving a harsher sentence or waiting longer for parole than is appropriate.") (emphasis added); Corbett-Davies et al., *supra* note 3 ("Here's the problem: it's actually impossible for a risk score to satisfy both *fairness* criteria at the same time.") (emphasis added).

mathematical terms.[7] COMPAS, with its combination of apparent racial bias and connection to an American criminal justice system already challenged for its racial disparities, has stimulated considerable attention, with over two hundred law review articles mentioning COMPAS to date. For lawyers writing about the perils of computational decision-making, COMPAS is the gift that keeps on giving.

But COMPAS is hardly an isolated example. ProPublica's story was done "as part of a larger examination of the powerful, largely hidden effect of algorithms in American life,"[8] and it comes at a time when society is confronting the implications of relying on computing technology for making decisions that affect the welfare and freedom of humans, as in the case of COMPAS. The Los Angeles Police Department recently abandoned software it used to identify crime "hot spots" and track violent criminals because the data provided to the software was itself subject to the bias of the police officers collecting it.[9] Some states are seeking to regulate certain kinds of computational decision-makers,[10] federal "algorithmic accountability" legislation has been introduced in Congress,[11] and the federal executive branch has issued reports on the potential for algorithmic discrimination.[12] Companies themselves are seeking regulation to prevent the most egregious forms of algorithmic discrimination.[13] Scholars, too, in both

7. Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811 (2020).

8. Angwin et al., *supra* note 2.

9. Mark Puente, *LAPD Moving Away Data-Driven Crime Programs Over Potential Racial Bias*, L.A. TIMES (Apr. 10, 2019), https://www.latimes.com/local/lanow/la-me-lapd-data-policing-20190410-story.html [https://perma.cc/9SJQ-HU9Y].

10. Michael J. Bologna, *'Hiring Robots' Restrictions Passed by Illinois Legislature*, BLOOMBERG L. (May 30, 2019), https://news.bloomberglaw.com/daily-labor-report/hiring-robots-restrictions-passed-by-illinois-legislature [https://perma.cc/8G4Y-FUGC].

11. Jon Fingas, *Senate Bill Would Make Tech Companies Test Algorithms for Bias*, ENGADGET (Apr. 10, 2019), https://www.engadget.com/2019/04/10/senate-algorithmic-accountability-act/ [https://perma.cc/WYH4-H9WW].

12. *See* EXEC. OFFICE OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS 5-6 (2016).

13. Brad Smith, *Facial recognition: It's Time for Action*, MICROSOFT: ON THE ISSUES (Dec. 6, 2018), https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/ [https://perma.cc/5MKV-F7HL].

the law and policy[14] and computer science[15] literature, are addressing the possibility that computational decision-makers will systematize unfair outcomes, either by repeating the prior un-fair decision-making processes they are designed to replicate or will develop biases on their own. The question underlying all of these inquiries is whether fairness is implicated if a computer program (even one that is designed to ignore protected categories such as race) produces disparate outcomes along protected lines, like race, sex, religion, or disability.

---

14. *See generally* VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2017); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 830 (2020); Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265 (2020); Danielle Keats Citron, *Technological Due Process*, 85 WASH. UNIV. L. REV. 1249, 1262 (2008); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4-14 (2014); Hellman, *supra* note 7; Kimberly A. Houser, *Can AI Solve the Diversity Problem in the Tech Industry? Mitigating Noise and Bias in Employment Decision-Making*, 22 STAN. TECH. L. REV. 290 (2019); Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611 (2020); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54 (2019); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 1 (2018); Joshua A. Kroll, et al., *Accountable Algorithms*, 165 UNIV. PA. L. REV. 633 (2017); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 665-66 (2017); Anya E.R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. (2020); Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 UNIV. PA. L. REV. 871, 909 (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017). Some work crosses the line between law and ethics. *See* Amitai Etzioni & Oren Etzioni, *Incorporating Ethics into Artificial Intelligence*, 21 J. ETHICS 403 (2017).

15. *See generally* Michael Veale, Max Van Kleek & Reuben Binns, *Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making*, PROC. OF THE 2018 CHI CONF. ON HUMAN FACTORS IN COMPUTER SYSTEMS 1 (2018) ("Calls for heightened consideration of fairness and accountability in algorithmically-informed public decisions—like taxation, justice, and child protection—are now commonplace."); *see*, *e.g.*, Micah Altman, Alexandra Wood & Effy Vayena, *A Harm-Reduction Framework for Algorithmic Fairness*, 16 IEEE SECURITY & PRIVACY 34 (2018); Reuben Binns, *What Can Political Philosophy Teach Us About Algorithmic Fairness?*, 16 IEEE SECURITY & PRIVACY 73 (2018); Cynthia Dwork et al., *Fairness Through Awareness*, PROC. OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONF. 214 (2012); Kate Donahue and Jon Kleinberg, *Fairness and Utilization in Allocating Resources with Uncertain Demand*, PROC. OF THE 2020 CONF. ON FAIRNESS ACCOUNTABILITY AND TRANSPARENCY 658 (2020); Min Kyung Lee, *Understanding Perception of Algorithmic Decisions: Fairness, Trust, and Emotion in Response to Algorithmic Management*, 5 BIG DATA & SOCIETY 1 (2018); Manish Raghavan et al., *Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practice*, PROC. OF THE 2020 CONF. ON FAIRNESS ACCOUNTABILITY AND TRANSPARENCY 469 (2020); Christian Sandvig, et al., *When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software*, 10 INTL. J. COMM. 4972 (2016); Nripsuta Saxena et al., *How Do Fairness Definitions Fare? Examining Public Attitudes Toward Algorithmic Definitions of Fairness*, PROC. OF THE 2019 AAAI/ACM CONF. ON ARTIFICIAL INTEL., ETHICS, AND SOC'Y 99 (2019).

Concerns about computational fairness have prompted a variety of approaches. Many have pointed out that computational decision-making often lacks transparency,[16] while others have analyzed the degree to which due process concerns should control how we think about computational decision-making.[17] Some have highlighted the connection between transparency and due process and the necessity of having good information about algorithmic decision-makers in order to challenge the process they afford.[18] Some have suggested that current discrimination law prohibits even unintentional but systematic bias by computational decision-makers,[19] while others have suggested we change discrimination law in order to deal with the problems raised by computational decision-making.[20] Some have suggested a variety of correctives to the problems of computational decision-making, including focusing less on law and more on how system developers should take care to design systems that do not further exacerbate unfairness.[21] Others have argued that, in order for computational decision-making to adequately account for racial disparities, they should be permitted to explicitly include consideration of race in their processes.[22] Still others have suggested that any approach to handling unfairness in computational decision-making will require a variety of approaches, some specific to technology, some to law, and some outside of both.[23]

Despite much academic scrutiny, little tangible progress has been made on how to respond to the possibility of rampant unfairness caused by computational decision-makers, partly because there is no widely held understanding of the fairness required,[24] and even if there

16. Pasquale, *supra* note 14, at 2; Bloch-Wehba *supra* note 14, at 1265; Chander, *supra* note 14, at 1039; Katyal, *supra* note 14, at 120; Raghavan, *supra* note 15, at 478; Anton Vedder & Laurens Naudts, *Accountability for the Use of Algorithms in a Big Data Environment*, 31 INT'L REV. OF L., COMP. AND TECH. 206, 214-15 (2017); Christian Zimmerman & Johana Cabinakova, *A Conceptualization of Accountability as a Privacy Principle*, in BIS 2015 WORKSHOPS 261, 266.

17. Huq, *supra* note 14, at 654.

18. Citron, *supra* note 14, at 1298-99; Citron & Pasquale *supra* note 14, at 32-33; Zimmerman & Cabinakova, *supra* note 16, at 266-67.

19. Kim, *supra* note 14, at 911.

20. Katyal, *supra* note 14, at 101.

21. *See* Kroll et al., *supra* note 14, at 662-71; Lehr & Ohm, *supra* note 14, at 715.

22. Altman et al., *supra* note 15, at 43; Bent, *supra* note 14, at 845; Hellman, *supra* note 7, at 846-62.

23. Barocas & Selbst, *supra* note 14, at 672 (suggesting a "a wholesale reexamination of the meanings of 'discrimination' and 'fairness.'" is necessary); *id*. at 729-32 (suggesting a variety of possible correctives, from increasing the burden on employers to using data to compare the effect of policies across employers to the status quo); Prince and Schwartz, *supra* note 14, at 1266 (offering a "menu of potential strategies"); Rich, *supra* note 14, at 929.

24. Saxena et al., *supra* note 15, at 99 ("While several definitions of fairness have recently been proposed in the computer science literature, there's a lack of agreement among researchers about which definition is the most appropriate."). There are the beginnings, however, of an effort to provide a common framework for discussing fairness across disciplines. *See* Deirdre K. Mulligan et al., *This Thing Called Fairness: Disciplinary*

were, such a requirement would be inconsistent with how law works to govern behavior.[25] Many suggest increased transparency for computational decision-makers, but the practicality of insisting on transparency varies widely with the technology underlying a particular system. Moreover, if the objective is to correct a problem created by shifting from human to computational decision-making, it's not clear what role transparency has, since much of human decision-making is hardly transparent. Discrimination law provides the answer, largely eschewing unattainable transparency in favor of accountability, and discrimination law itself contains a mechanism for forcing that accountability that can be applied to computational decision-makers in much the same way it has applied to human ones.[26] Some see the computerization of decision-making as an opportunity to consciously address existing inequality, but doing so ignores that it is the decision-making itself, not the disparate outcomes it produces, that is relevant to law. Altering decision-making processes to produce balanced outcomes is extremely problematic, at least if current law is any measure. I suggest a different approach, one that draws upon both the nature of computer systems development and features of discrimination law to better describe the steps by which systems are modified in response to sub-optimal outcomes, including racially disparate ones. The law applies differently to the different steps of modifying an algorithm. Disaggregating the process of modifying an algorithm into its component parts—separating the rejection of a poorly performing algorithm from the distinct process of developing its replacement—shows how discrimination law affords considerable flexibility to both firms and government agencies in modifying algorithms to produce less racially disparate outcomes.[27]

In Part I, I provide a brief overview of the problem, from the question of how algorithms (applied by both computers and humans) work to control decision-making to the specific technology for enabling computers to make decisions that affect the rights and welfare of human beings. After a discussion of the various forms of computational decision-making and how their technologies affect the fairness inquiry, I address the problem of fairness as applied to computational decision-making. As is clear from brief examination, there is no widely held normative or legal concept of computational "fairness;" the closest analogy the law can provide are various prohibitions on discrimination, the most common being those of employment discrimination and constitutional equal protection law. Those constraints exist as side constraints on other activities, and their

Confusion Realizing a Value in Technology, 3 PROC. ACM HUM.-COMPUTER INTERACTION, Art. 119.

25. *See infra* Section III.A.
26. *See infra* Section III.B.
27. *See infra* Section III.C.

position as side constraints dramatically limit their ability to serve as a design for any sort of system of computational fairness. In Part II, I describe how the law of discrimination applies to the kinds of decisions made by computational decision-makers. The disparate outcomes that are likely to become increasingly apparent in a world of computational decision-makers are generally legally irrelevant, a reminder that one cannot simply balance away problems of illegal discrimination, by algorithm or otherwise. Instead, virtually every anti-discrimination regime (statutory or constitutional) requires an inquiry into purpose, which is a problem not all that different for computational decision-makers than the traditional problems of inferring purpose from human-applied but institutionally devised policies. In Part III, I consider the implications of that legal analysis for computational decision-makers. Far from presenting unfathomable or insurmountable challenges, applying discrimination law to computational decision-makers demonstrates both the reach and the limits of discrimination law as a side constraint on productive activity. Recognizing the limits of transparency in human decision-making highlights the need to identify what it is that discrimination law demands, which is not transparency but rather accountability. That accountability is no harder to achieve with computational decision-makers than with human ones. Scholars are currently divided over whether it is permissible to alter algorithms in order to reduce disparity, some arguing it is and some arguing that doing so is itself prohibited discrimination.[28] I offer a third way, one that disaggregates the series of decisions that result in a new, altered algorithm and recognizes how the law applies differently to those different decisions. With that understanding, I provide a roadmap for how to modify algorithms to reduce disparities without running afoul of the law. The paper ends with a brief Conclusion.

I. FAIRNESS AND DISCRIMINATION IN COMPUTATIONAL DECISION-MAKING

As the allusion to the COMPAS case suggests, inquiries into the use of computational decision-making naturally gravitate toward questions of fairness,[29] which are claims that are usually supported by pointing to the unfair outcomes (like the disparate recidivism scores given by COMPAS) produced by computational decision-makers.[30]

28. *Compare* Kroll et al., *supra* note 14, at 694 (not permissible), *with* Altman et al., *supra* note 15, at 43, *and* Bent, *supra* note 14, at 845, *and* Hellman, *supra* note 7, at 846-62 (arguably permissible).

29. PASQUALE, *supra* note 14, at 9 ("The most obvious question is: Are these algorithmic applications fair?").

30. *See e.g.*, Angwin et al., *supra* note 2; Barocas & Selbst, *supra* note 14, at 674; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017); Hellman, *supra* note 7; Huq, *supra* note 14.

Although the emphasis on outcomes is understandable for systems designers, who are interested in solving an optimization problem, it is problematic as a matter of law, and one point of this paper is to shift the debate away from such arguments.

Before we can address the problem of computational fairness, though, it's helpful to define some terms. Deriving a definition of "computational decision-making" reveals that computational decision-makers have much more in common with human-centric decision-making systems than we might like to acknowledge. With an idea of computational decision-making in mind, we can consider the role of fairness in such systems. As a threshold matter, it seems unlikely that a comprehensive concept like fairness could be expressed in a suitably concrete form to allow it to be workably incorporated into computational decision-making.[31] But even if computer science could accommodate a coded form of fairness, it's not clear that either fairness itself or law could. The larger problem is that there is no such thing as "fairness," or at least there is no way to know and agree upon what fairness is, and law, being a social enterprise, requires such agreement as a precondition to its existence. Even if there were universal agreement on what fairness requires, the *law* is not generally interested in fairness. Rather, the law is far more specific, and is largely negative. The law prohibits particular forms of unfairness rather than affirmatively requiring anything resembling "fair" outcomes.

A. Computational Decision-making

One of the difficulties of thinking about fairness in computational decision-making is that the problem presents itself in such a wide variety of circumstances, both in the types of decisions being made and in the types of systems making those decisions. But the law applies differently based on context. The law takes a completely different approach, for instance, between governmental and private action,[32] and we can also expect differences based on the nature of the decision being made and the degree of discretion provided to computational decision-makers. A system for determining which seat I am assigned on a particular airline flight rightly deserves less scrutiny than one

31. *See generally* Rebecca Crootof, *Cyborg Justice*, 119 COLUM. L. REV. FORUM 233, 239 (2019) ("Nor can we translate our statutes and common law into easily applied rules. While it is tempting to imagine law as subject to algorithmic application, reality is far messier. AI can apply unambiguous rules, but even apparently simple laws are far from unambiguous.") (footnotes omitted); Lehr & Ohm, *supra* note 14 at 674; Nick Tarleton, *Coherent Extrapolated Volition: A Meta-Level Approach to Machine Ethics*, THE SINGULARITY INST., 1, 4 (2010) (describing "the profound challenges of describing these goals mathematically and creating a system that reliably implements them").

32. *Compare* Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 226 (1995) (constitutional strict scrutiny analysis for governmental race classifications) *with* McDonnell Douglas Corp. v. Green, 411 U.S. 792, 804 (1973) (burden-shifting framework for statutory race discrimination claims).

that determines whether I receive a job or a mortgage. Moreover, a system that executes a well-defined and documented set of "if" statements will be easier to audit than one that has designed itself in light of available data.[33] It is tempting given the headlines to focus on one kind of system, especially novel ones, such as artificial intelligence systems,[34] but in order to get a comprehensive understanding of the problem, it is necessary to break it down so that we can evaluate exactly what is at stake. Rather than start at the level of specific systems or types of systems, it's helpful to start with more abstract concepts that unite these systems and move toward a more detailed understanding. Those abstract concepts go beyond the world of computers and software.

As highlighted by the COMPAS controversy, the current conversation is superficially about the use of automated systems to make decisions that implicate the rights and welfare of humans. Those systems come in many forms, but before there can be an automated system, there must be some expression of what that system is supposed to do. Thus, at the most general, the form of decision-making implicated by systems like COMPAS is "algorithmic": "a step-by-step procedure for solving a problem or accomplishing some end,"[35] and I will refer to such decision-making processes generally as *algorithmic decision-making* and those entities that make decisions by algorithm *algorithmic decision-makers*.

Algorithmic decision-making can take place in a computer, but it need not. An employer's policy to hire only candidates with college degrees, or a mortgage lender's policy to give mortgages only to applicants with credit scores above 700, is equally algorithmic whether it is applied by a machine or a human.[36] But if my notional seven-hundred-point credit score cutoff is somehow problematic, then the problem is the algorithm, not the fact that it happens in a machine. If the problem is in the algorithms we use, then we should acknowledge that and address what's problematic about using *algorithms* rather than what's problematic about using *machines*.

One thing that might be problematic about using algorithms is the effect doing so has on the discretion of decision-makers; the types of

---

33. *See infra* the text accompanying notes 46-47.

34. *E.g.*, Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1831-32 (2019).

35. *Algorithm*, MERRIAM-WEBSTER, https://www.merriam-webster.com/dictionary/algorithm [https://perma.cc/ZWW6-9MBX] (last visited March 31, 2021). *See also* Vedder & Naudts, *supra* note 16, at 206 ("Algorithms have been defined as finite, abstract, effective, compound control structures, imperatively given and accomplishing a given purpose under given provisions, or even more concisely, as encoded procedures through which input data are being transformed into a usable, and therefore desired, output.") (internal citations omitted).

36. *See* Jennifer S. Light, *When Computers Were Women*, 40 TECH. & CULTURE 455, 458 (1999) (describing the changing relationship between human and computer approaches to ballistics calculations during World War II).

rules captured in algorithms have a huge effect on the relative discretion of actors in a decision-making system. Rules, including those captured in algorithms, effectively shift authority from the decision-maker in a specific case (such as a police officer or judge) to the actor who made the policy in the first place (such as a legislature).[37] At least they do so if they are captured correctly, and one source of concern over computational decision-makers is that something will go wrong in the translation from intent to operation. That is also true of human-centered systems—a point obvious to anyone who's been to a government office or a bank and witnessed a clerk applying a policy in a wrongheaded way. We generally hope that, if there is a human who makes the ultimate decision (such as a human airline pilot flying a plane equipped with autopilot, or a human "in the loop" of a weapons system[38]), that human recognizes the error in how the rules have been captured and applied and will intervene appropriately.

As the foregoing suggests, there is no single model for how decision-making is shared between humans and machines. Given that much of the present concern is over the automation of decision-making, it is helpful to think about the degree to which a particular decision-making system *delegates* decision-making authority among the various actors, computer, human, algorithmic, or otherwise. We generally perceive of systems as delegating decision-making authority from humans to computers, but that is a vast oversimplification, since humans delegate to each other (and to other animals – think of the role a drug-sniffing dog's "hit" plays in a police officer's decision to search your baggage[39]) in a variety of ways, and many uses of computers (such as confirming a loan applicant's home address via a Google search) are not really delegations at all.

Thus, before we even start talking about the various kind of automated decision-making systems out there, we have identified a major piece of the puzzle: the implications of humans delegating decision-making authority to automated systems. Delegating decision-making to computational decision-makers might be problematic,[40] but

37. FREDRICK SCHAUER, PLAYING BY THE RULES: A PHILOSOPHICAL EXAMINATION OF RULE-BASED DECISION-MAKING IN LAW AND IN LIFE 112-13 (1991).

38. *See* DEP'T OF DEFENSE, DIRECTIVE 3000.09, AUTONOMY IN WEAPON SYSTEMS (2012) ("Autonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force.").

39. *See* Florida v. Harris, 568 U.S. 237 (2013) (holding that certified drug-sniffing dog's alert can provide probable cause to search).

40. *See* State v. Loomis, 881 N.W.2d 759, 761-62 (regarding the delegation inherent in using COMPAS itself). Such delegations are prohibited by the E.U. General Data Protection Regulation. *See General Data Protection Regulation: Automated Individual Decision-Making, Including Profiling* 2016/679, art. 22, 2016 O.J. (L 119) 3 22. (protecting "the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or

so might delegating decision-making to humans using other algorithmic processes. Computerization of decision-making might make the problem worse in some way, but it's not superficially any different than other delegations through rules (algorithms) to systems that do not involve computers.

Although my inquiry is relevant to algorithmic decision-making generally, the focus is on the use of computers to exercise influence in decision-making, sometimes decisively. I call this *computational decision-making*, which I consider to be a sub-set of algorithmic decision-making. Those actors that engage in computational decision-making are *computational decision-makers*.

Computational decision-makers come in a variety of flavors, but there are three principle categories of computational decision-making methods, each with slightly different implications. First, there is *traditional programming*, which encodes explicit decision criteria in a series of sequentially executed statements.[41] Second, there are *rule-based systems*,[42] which also operate using explicit, pre-defined criteria, but do so through a collection of rules that are not necessarily executed in a particular order. Many human-executed algorithms look a lot like this, since humans can apply rules selectively in the way rule-based systems do. The existence of explicit rules make both traditional programs and rule-based systems relatively easy to analyze for determining what criteria they are using to make decisions.

When we go beyond traditional programming and rule-based systems, we encounter a third form of computational decision-making: *machine learning*,[43] which itself is a collection of methods[44] used to build computer systems that can perform specific functions without relying on explicit instructions, relying instead on inference and

---

her"). *See generally* Margot Kaminsky, *The Right to an Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 196-204 (2019).

41. *See* STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH, 236, 285-86 (3d ed. 2009) (discussing the traditional, "procedural approach" to programming).

42. Federico Cabitza, Marcello Sarini & B. Dal Seno, *D Jess - A Context-Sharing Middleware to Deploy Distributed Inference Systems in Pervasive Computing Domains*, ICPS '05. PROC. INT'L CONF. ON PERVASIVE SERV., 229, 231 (2005) (describing rule-based systems).

43. TOM M. MITCHELL, MACHINE LEARNING 1-2 (1997).

44. Machine leaning comes (in very general terms) in three different forms: *reinforcement learning*, *unsupervised learning*, and *supervised learning*. Computationally, reinforcement learning is perhaps the most interesting, since it provides the ability for a computer system to develop its own solutions to problems based on trial and error. RUSSELL & NORVIG, *supra* note 41, at 694-695. But machine learning currently has limited application to actual human-connected decision-making systems. Unsupervised learning groups items together based on their similarity, a process commonly used to provide information used to train supervised learning models but, because they identify similarity rather than the relationship between inputs and outputs, not decision-making systems themselves. Supervised learning models do exactly that: they learn from labeled historical data to predict outputs from inputs. KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 2 (2012).

pattern recognition, improving their capability through experience.[45]

The most relevant form of machine learning for present purposes is *supervised learning*. Supervised learning approaches use a learning algorithm to build a *model*, but the model itself is not algorithmic[46]–it captures associations but does not attempt to demonstrate the cause or nature of the association.[47] That is, supervised learning models identify a relationship between an input (credit score) or set of inputs (credit score, income, total debt, and ZIP code) and an output (the likelihood that someone will repay a loan) but without necessarily specifying the algorithm that produces that relationship. Supervised learning models have been at the center of discussions regarding the role of machine learning in computational decision-making.[48]

While all forms of algorithmic decision-making present concerns, and computational decision-making presents the additional issue of delegating decision-making to machines, machine learning changes the ballgame entirely, both because of what it makes possible and how it does so. Machine learning allows the computerization of problems previously too complex for traditional programming or rule-based systems to resolve reliably. For instance, although traditional programming worked for the development of master-level software for playing the game chess, machine learning techniques enabled the development of master-level software capable of for playing the far

---

45. MURPHY, *supra* note 44. Although many have written about "artificial intelligence," I generally eschew that term because it describes a capability rather than a method. *Compare* RUSSELL & NORVIG, *supra* note 41, at 14 (describing artificial intelligence as "the study of agents that receive percepts from the environment and perform actions") *with* IAN GOODFELLOW ET AL., DEEP LEARNING 96 (2016) ("Machine learning is essentially a form of applied statistics with increased emphasis on the use of computers to statistically estimate complicated functions and a decreased emphasis on proving confidence intervals around these functions."). As is clear below, how the law applies will depend on the specific methods used to develop computational decision-makers, and so my analysis is specific to those *methods* rather than to the *capability* for computers to engage in higher-order thinking. Because my analysis does cover the methods for developing artificial intelligence systems (as well as virtually any computational decision-maker), it is applicable to artificial intelligence generally.

46. And, thus, my inclusion of machine learning systems as "algorithmic decision-makers" is something of a fudge, since, while algorithms are used in their construction, machine learning systems make decisions based on their models, not on algorithms. Machine learning systems nevertheless present many of the problems of mechanical decision-making that algorithms do, even if the reasoning is captured in model, rather than algorithmic, form. For more about the non-algorithmic nature of machine learning models, see generally INFORMATICS EUR. & ASS'N FOR COMPUTING MACHINERY EUR. COUNCIL, WHEN COMPUTERS DECIDE: EUROPEAN RECOMMENDATIONS ON MACHINE-LEARNED AUTOMATED DECISION MAKING, 9 (2018), https://www.acm.org/binaries/content/assets/public-policy/ie-euacm-adm-report-2018.pdf [https://perma.cc/ULU8-9P96] [herinafter INFORMATICS EUR.].

47. RUSSELL & NORVIG, *supra* note 41, at 710. (laying out the code for "[a]n algorithm to select the model that has the lowest error rate on validation data by building models of increasing complexity, and choosing the one with best empirical error rate on validation data.").

48. Lehr & Ohm, *supra* note 14 at 676 (on the centrality of supervised learning to current debates over the social impact of machine learning).

more complex game of go.[49] That increased ability to handle complex problems is likely to lead to delegating increasingly sophisticated decisions, or more completely delegating decisions, to computational decision-makers. It may be possible for traditional programming to produce a credit score based on a fixed set of criteria, like income, current debt, and previous payment history, but the decision regarding whether to make a specific loan is left to a human based in part on that information. Machine learning might allow the aggregation of credit score along with a host of other criteria in order to give humans enough confidence to allow the computer to make the final lending decision. And, as many have pointed out,[50] while machine learning has the potential to vastly expand both the capability and autonomy of computational decision-makers, the use of models generated by learning algorithms—instead of decision algorithms written by humans—makes systems based on machine learning potentially less transparent than other systems.[51]

Although the different forms of computational decision-makers do matter, the problem at its core is fundamentally the same: how to address potentially snowballing unfairness caused by the pervasive use of increasingly sophisticated and potentially opaque computational decision-makers. Especially given the lack of clear lines between the various forms of programming,[52] or between various stages of decision-making (imagine a decision that is produced in-part by traditional programing and in-part by a machine learning system), it is not merely preferable but virtually essential to come up with a legal rule that treats all decision-makers the same. And, because the problem of algorithmic fairness—that is, the fairness of practices and policies that provide rules of decision—is not a new one, a solution that works for computational decision-makers should apply to algorithmic decision-making performed by humans as well.

---

49. David Silver & Dennis Hassabis, A*lphaGo: Mastering the Ancient Game of Go with Machine Learning*, GOOGLE: AI BLOG (2016) ("The search space in Go is vast -- more than a googol times larger than chess (a number greater than there are atoms in the universe!). As a result, traditional "brute force" AI methods -- which construct a search tree over all possible sequences of moves -- don't have a chance in Go."). Machine learning can also exceed traditional programming in domains where traditional programming *does* work. *See* David Silver et al., *Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm*, SCIENCE 5 (Dec. 7, 2018), https://science.sciencemag.org/content/362/6419/1140/tab-pdf [https://perma.cc/N64V-UYMM] ("AlphaZero is a generic reinforcement learning algorithm and search algorithm– originally devised for the game of Go – that achieved superior results [to traditional chess programs] within a few hours, searching 1/1000 as many positions, given no domain knowledge except the rules of chess.").

50. *See supra* the text accompanying notes 16-18.

51. INFORMATICS EUR., *supra* note 46, at 9. "While conventional computer applications may appear to behave similarly, they have an internal logic and are constructed out of abstractions that make an application's logic and behaviour comprehensible and reliably predictable to its software developers." Deeks, *supra* note 34, at 1833.

52. Silver & Hassabis, *supra* note 49, (explaining that AlphaGo's algorithm uses a combination of machine learning and traditional tree-search techniques).

B. The Problem of Fairness

With an understanding of what computational decision-making is and where it fits in the broader concept of decision-making by algorithm, procedure, or rule, we can now consider how fairness should operate in such systems. That turns out to be a very thorny problem because fairness is not a constraint that is frequently operationalized in systems for a variety of reasons, including that no one can agree what "fairness" is. Rather, "fairness" stands in for a host of ideas that are not only widely and morally disputed but apply to greater or lesser degrees in different contexts. Consequently, law does not consider fairness as an abstract concept—law takes a far more particularized and limited approach than the broader fairness inquiry.

1. Fairness as an Essentially Contested Concept

One initial problem with substantively focused inquiries in algorithmic or computational fairness is the absence of any universally, or even widely, held comprehensive understanding of "fairness." The intuitive appeal of the concept of fairness is virtually irresistible for those providing their own proposals for guiding conduct. Some have suggested particularly narrow conceptions, for instance fairness as reciprocity rather than as an absolute moral or political imperative.[53] Some conceive of fairness as requiring consideration of overall wellness[54] or other rules for allocating resources,[55] and so measuring this kind of fairness requires determining whether resources are actually allocated in fair ways.[56] For some, fairness is defined relative to perception, with a "fair" outcome being one that is *perceived* as "fair."[57] For some, fairness is simply the avoidance of bias,[58] although merely avoiding bias might be justified equally on instrumental grounds (avoiding the cost of inaccuracy[59]) as on moral

53. Matthew Rabin, *Incorporating Fairness into Game Theory and Economics*, 83 AM. ECON. REV., 1281, 1285 (1993).

54. *E.g.*, Altman et al., *supra* note 15, at 35.

55. *E.g.*, Donahue & Kleinberg, *supra* note 15, at 658; Matthew Joseph et al., *Fairness in Learning: Classic and Contextual Bandits*, PROC. OF THE 30TH INT'L CONF. ON NEURAL INFO. PROCESSING SYSTEMS 325, 325 (2016) ("Our fairness definition demands that, given a pool of applicants, a worse applicant is never favored over a better one . . . ."), https://dl.acm.org/doi/pdf/10.5555/3157096.3157133 [https://perma.cc/AY6E-RNCN]; Saxena et al., *supra* note 15, at 100.

56. *E.g.*, Altman et al., *supra* note 15, at 35; Nathan Kallus, Xiaojie Mao & Angela Zhou, *Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination*, PROC. OF THE 2020 CONF. ON FAIRNESS ACCOUNTABILITY AND TRANSPARENCY 110 (2020), https://doi.org/10.1145/3351095.3373154 [https://perma.cc/2LDC-84H6].

57. Lee, *supra* note 15, at 1.

58. Dwork et al., *supra* note 15, at 214; *but see e.g.*, Altman et al, *supra* note 15; *but see also* Kallus, Mao & Zhou (defining "fair affirmative action" in terms of statistical parity, which is an outcome measure). *See also* Veale, Van Kleek & Binns, *supra* note 15, at 6.

59. Binns, *supra* note 15, at 74; Yang Liu et al., *Calibrated Fairness in Bandits*, PROC. OF THE 2017 CONF. ON FAIRNESS ACCOUNTABILITY AND TRANSPARENCY 99, 102 (2017) (calibrating selection in order to best approximate the expected value of a randomized draw), https://arxiv.org/abs/1707.01875v1 [https://perma.cc/V6UW-LLS6].

or legal ones. Others are less precise, lumping outcomes and bias (and other problems) together.[60] Even limiting the definition to bias, though, does little to avoid the underlying moral question of what kinds of bias are and are not permissible.[61] But even if we could agree on what forms of bias were best avoided, we would have to agree on the baseline from which bias deviates, and there is little agreement on that.[62]

Indeed, even cursory examination reveals that not only is there is no single understanding of fairness[63] but that fairness is an "essentially contested concept"—a concept that is not merely highly contested but whose essence is itself disagreement; a concept whose very purpose is to give common name to what is understood by all to be an incompletely described phenomenon.[64] Fairness is not just difficult or impossible to define; the word "fairness" itself exists to absolve us of the need to define what it represents. It gives us a single word to name that which cannot be described, both as a matter of complexity (much like "Grand Canyon" describes a place that is itself too large and complex to be described in detail) and agreement (much like "beautiful" is understood to be an inherently relative term).

And here lies a fundamental difference between fairness and law. Unlike fairness, law cannot exist as an inherently contested concept because the foundation of law is agreement among society (at least generally) as to its content. Law—the judicially enforceable rules by which society operates—reflects a settlement of inherently unresolvable conflicts regarding fairness, or right, or justice.[65] As a shared social enterprise, law can only regulate that which can be agreed upon, and the content of law reflects that. Thus, it is little surprise that our legal systems have generally rejected legal mandates

---

60. *E.g.*, EXEC. OFFICE OF THE PRESIDENT, *supra* note 12, at 15 ("Promoting fairness, ethics, and mechanisms for mitigating discrimination in employment opportunity."); Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 REG. & GOVERNANCE 505, 516 (2017) (combining "key aspects of democracy, equality, fairness, and distributive justice"), https://doi.org/10.1111/rego.12158 5324/eip.v10i1.1961 [https://perma.cc/8DSC-MFG7].

61. *See* Kroll et al., *supra* note 14, at 678 ("[W]hat makes a rule unacceptably discriminatory against some person or group is a fundamental and contested question. We do not address that question here, much less claim to resolve it with computational precision."); Corbett-Davies et al., *supra* note 3 (pointing out that one form of racial fairness or bias at issue in the COMPAS is "mathematically guaranteed" given an alternative, equally plausible, definition of racial fairness or bias); Hellman, *supra* note 7 at 811.

62. Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES 175 (Tarleton Gillespie, Pablo Boczkowski & Kirstin Foot eds., 2012) ("To accuse an algorithm of bias implies that there exists an unbiased judgment of relevance available, to which the tool is failing to hew. Since no such measure is available, disputes over algorithmic evaluations have no solid ground to fall back on.").

63. Etzioni & Etzioni, *supra* note 14, at 406.

64. Walter B. Gallie, *Essentially Contested Concepts*, 56 PROC. OF THE ARISTOTELIAN SOC'Y 167, 169 (1956).

65. LARRY ALEXANDER & EMILY SHERWIN, THE RULE OF RULES: MORALITY, RULES, AND THE DILEMMAS OF LAW 170 (2001); SCHAUER, *supra* note 37, at 167-74; *see also* RONALD L. DWORKIN, LAW'S EMPIRE 114-15 (1986).

imposing fairness, which should be a clue in its own right. Even regimes with "fair" in the name, like "unfair competition," do not mandate *fair* conduct—instead they regulate negatively, prohibiting a specific set of acts deemed to be wrongful,[66] some of which invoke notions of fairness and others of which do not. We should give up on trying to codify notions of fairness in computational decision-makers. We will be lucky enough if we can codify notions of law,[67] which themselves are hardly without doubt or disagreement.[68]

But I will concede what I consider to be the easier point in order to make the harder one: Even if we could develop a conception of fairness and find a way to describe it in algorithmic form, we would not want to codify it in computational decision-makers.

2. Fairness as a Side Constraint

The reason we would not want a fairness machine even if we could create it is because fairness is not generally something to be optimized in society. Instead, fairness operates as what Robert Nozick called a "side constraint"—a rule that ignores the goals of the system on which it operates because it is in service of some other goal.[69] Two aspects of side constraints have particular relevance for analyzing how fairness generally works as a side constraint.

First, although fairness might be a goal, it is not very helpful to think of side constraints like fairness as representing discrete goals in their own right but rather as operating within or on other systems that have their own goals. Side constraints are second-order rules: rules that control the application of other rules.[70] No system exists to produce fairness in the abstract. For instance, the purpose of agriculture is to supply food. The agricultural industry should provide that food on some moderately fair basis, but it would be odd to say that the goal of the agriculture system is to increase fairness and that it does so through the production of food. We may want the agricultural system to produce food in a fair way, but the invention and continued

66. *See* Patel v. Zillow, Inc., 915 F.3d 446 (7th Cir. 2019) (Easterbrook, J.) (describing the comparative problems of imposing a requirement of fairness rather than outlawing misleading practices). The closest regime is copyright's "fair use" doctrine, but in that case, "fair use" is a conclusion. Fairness itself has no independent role in the doctrine's four-pronged analysis. *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

67. *See* sources cited *supra* note 31.

68. *See*, *e.g.*, Peter Westen, *The Empty Idea of Equality*, 95 HARV. L. REV. 537, 559 (1982) (on the indeterminacy of the concept of equality, a concept explicitly made law by the Equal Protection Clause of the Constitution).

69. ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 28-29 (1974) (describing side constraints). For Nozick, side constraints represent prohibitions on state action flowing from higher order individual right. *See id*. at 29. However, they are not so limited. *See* Fredrick Schauer, *The Annoying Constitution: Implications for the Allocation of Interpretive Authority*, 58 WM. & MARY L. REV. 1689, 1693-94 (2017) (adapting rights-based side constraints in service of higher order institutional design rather than individual rights).

70. Schauer, *supra* note 69, at 1693-94 (on the general applicability of second-order rules).

operation of the agricultural system is in service of producing food, not fairness. There is no system in which fairness itself is the principal goal, only systems with other principal goals on which fairness operates.

Second, side constraints, like law, preempt the operation of other systems[71]—a conflict between the operation of a system and a side constraint on that system is always resolved in favor of the side constraint. That means that side constraints must represent *higher* order principles than the systems upon which they operate in order to justify preemption of that lower order system.[72] If I produce food in violation of law, it is understood that the conflict can be resolved only one way: by my changing my food production to comply with the legal rule. The law will control my production of food, not vice versa.

A necessary consequence of these two aspects of side constraints is that they must operate not as values to be maximized in some way but rather as threshold requirements that, once satisfied, demand no further action. The alternative would be unworkable. Given their supremacy, if side constraints represented values to be maximized and not merely satisfied, then every system subject to the side constraint would effectively become a system designed to maximize the values underlying the side constraint. To take my agricultural example, if fairness has preemptive value (if we would invoke fairness as a mandate to prevent the unfair allocation of food) and also must be maximized, then every step taken in the agricultural system would be evaluated for its ability to produce fairness rather than for its ability to produce food (including the possibility that we could produce fairness in ways completely unrelated to agriculture). It is possible to think of the agricultural system in that way, but not without completely merging the concepts of fairness and agriculture, which would cause more confusion than clarity.

It cannot be any other way. If fairness were the value being maximized, it would become practically impossible to produce anything because scrupulous attention to fairness would interfere with other productive processes (in addition to the problem that it's impossible to have "enough" fairness). Fairness is not itself a maximand; once a threshold level of it is achieved, the fairness constraint is satisfied and the focus shifts to maximizing the goals of the systems on which fairness operates.

That is not to say that aspects of fairness cannot be programmed into an algorithmic decision-maker, at least if one observes the limitations of side constraints like fairness. For instance, most would agree that using race to make lending decisions is unfair (and

71. *Id*.
72. NOZICK *supra* note 69, at 29 (on rights trumping other utilitarian values).

illegal[73]), and it is possible to accommodate this aspect of fairness by excluding racial considerations from a lending algorithm. We can constrain the operation of lending algorithms in certain ways to satisfy the demands of fairness (at least the ones we can agree on) even if we can't program them to produce optimally fair outcomes without converting them into fairness algorithms instead of lending algorithms.

Law, which operates largely as a side constraint on other activity,[74] generally works the same way. Law places minimal demands on many different activities but, once law's minimal demands are satisfied, law becomes irrelevant and we are free to maximize the goal of the particular activity, not the values underlying the law.

C. Legal Fairness Relevant to Algorithms: Discrimination Law

Given the absence of "fairness law," inquiry into the role of fairness in law quickly becomes a search for examples of fairness-oriented legal regimes to see how they implement fairness.[75] It is helpful for examples to be ones that are relevant to computational decision-making. One could argue that contract law is based in fairness because it is fair to hold people to their promises, but that is no different a concern in computational decision-making than in human decision-making, since promises are individually entered into. Rather, the clearest case for considering how fairness might work in computational decision-making seems to come from concerns over discrimination on the basis of race, sex, religion, or other bases thought to be invalid. The widespread existence of laws preventing discrimination along those lines is a strong indication of our society's views on the invalidity of those characteristics, and so those laws represent the best source of ideas for how to implement a fairness-oriented regime (one in which the underlying substantive fairness question is widely agreed upon). As a practical matter, those laws also present the greatest legal risk to firms and governments worried that their algorithms might discriminate inappropriately.[76]

---

73. Equal Credit Opportunity Act, 15 U.S.C.A. § 1691.

74. It does not, for instance, in the case of power-conferring rules, such as the rules of contract law.

75. *But see* Huq, *supra* note 14, at 1102 ("In the dialogue between equal protection and algorithmic criminal justice, I suspect that constitutional law has much to learn and little to teach.").

76. *See generally* Roy Maurer, *AI-Based Hiring Concerns Academics, Regulators*, SHRM: TALENT ACQUISITION (Feb. 14, 2020), https://www.shrm.org/resourcesandtools/hr-topics/talent-acquisition/pages/ai-based-hiring-concerns-academics-regulators.aspx [https://perma.cc/CGW3-5WTF]; Robert Bartlett et al., *Consumer-Lending Discrimination in the FinTech Era* (Nat'l Bureau of Econ. Research, Working Paper No. 25943, 2019), https://www.nber.org/papers/w25943 [https://perma.cc/QG2U-J33R] (lending discrimination laws).

Thus, discrimination law is not only practically important to those employing computational decision-makers, it serves as a useful reference point for an operationalized system to combat what most consider to be inherently unfair practices, such as discrimination on the basis of race, or sex, or religion.[77] Laws prohibiting discrimination by race or sex in contexts such as employment or lending also reflect much of the outstanding legal[78] and computer science[79] literature on fairness in the context of computational decision-making, and so they seem a good place to start.

II. DISCRIMINATION LAW AND ALGORITHMIC DISCRIMINATION

In the United States, many forms of discrimination are legally prohibited. The Civil Rights Act of 1866, adopted in the wake of the Civil War, outlawed discrimination on the basis of race in the making and enforcing of contracts generally (along with other civil rights).[80] There are modern federal (and many state) laws outlawing discrimination in employment on the basis of race, color, religion, sex, national origin,[81] along with age,[82] the possession of certain genetic characteristics,[83] citizenship,[84] or membership in the reserve component of the U.S. armed forces or the National Guard,[85] among others. Discrimination in housing is prohibited along the lines of race,

77. *E.g.*, Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(a) (2018) ("It shall be an unlawful employment practice for an employer . . . to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin."); Equal Credit Opportunity Act, 15 U.S.C. § 169(a) (2018) ("It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction . . . on the basis of race, color, religion, national origin, sex or marital status, or age . . . ."); Fair Housing Act, 42 U.S.C. § 3604(b) (2018) ("[I]t shall be unlawful . . . [t]o discriminate against any person in the terms, conditions, or privileges of sale or rental of a dwelling, or in the provision of services or facilities in connection therewith, because of race, color, religion, sex, familial status, or national origin.").

78. *E.g.*, Barocasa & Selbst, *supra* note 14, at 685-86; Bent, *supra* note 14, at 804-05; Hellman, *supra* note 7; Kim, *supra* note 14, at 888-90; Kroll et al. , *supra* note 14, at 692-94.

79. Saxena et al., supra note 15, at 100 ("the three fairness definitions examined here agree that, conditioned on the task-specific metric, an attribute such as race should not be relevant to decision-making"); Veale, Van Kleek & Binns, *supra* note 15, at 6 ("Discrimination has taken centre-stage as the algorithmic issue that perhaps most concerns the media and the public."); Yeung, *supra* note 60, at 516.

80. Civil Rights Act of 1866, 42 U.S.C. § 1981(a) (2018) ("All persons within the jurisdiction of the United States shall have the same right in every State and Territory to make and enforce contracts, to sue, be parties, give evidence, and to the full and equal benefit of all laws and proceedings for the security of persons and property as is enjoyed by white citizens, and shall be subject to like punishment, pains, penalties, taxes, licenses, and exactions of every kind, and to no other.")

81. 42 U.S.C. § 2000e-2 (2018).

82. Age Discrimination in Employment Act, 29 U.S.C. § 623 (2018).

83. Genetic Information Nondiscrimination Act, 4238 U.S.C. § 2000ff-14311 (2018).

84. Immigration Reform and Control Act, 8 U.S.C. § 1324b (2018).

85. Uniformed Services Employment and Reemployment Rights Act, 38 U.S.C. § 4311 (2018).

color, religion, sex, handicap, familial status, or national origin,[86] and discrimination in granting credit is prohibited as to race, color, sex, religion, national origin, marital status, age (over eighteen), or the receipt of public assistance income.[87] Title IX of the Education Amendments of 1972 prohibits discrimination on the basis of sex in participation in educational programs funded by the federal government.[88] The United States Constitution prohibits certain forms of discrimination by government, with heightened scrutiny for discrimination on the basis of race[89] or sex,[90] prohibitions on discrimination intended to convey animus toward a particular group,[91] and a general, if rather weak, prohibition on irrational discrimination.[92]

Legal prohibitions on discrimination are similar in structure. They effectively make certain characteristics legally irrelevant, but they nevertheless represent a very limited form of fairness. First, they apply only to a very small sub-category of both conduct (employment and housing are covered, for instance, but non-economic relationships such as friendship and acquaintance are not) and protected classes (race, sex, disability, and religious affiliation are covered, but class, political affiliation, and non-disabling physical characteristics like left-handedness are not). The difficulty of applying broad anti-discrimination rules to widely varying activity might be one reason why legal discrimination prohibitions are so specific in both the traits and activities they cover.

Much more importantly, discrimination laws operate largely negatively—they *prohibit* certain forms of discrimination but they do not *require* consideration of particular characteristics regardless of their relevance. Some statutes encourage consideration of characteristics that are relevant as a way of demonstrating that one was not considering a prohibited characteristic, but neither Congress nor the Constitution[93] imposes much in the way of affirmative obligations to engage in "fair" activity. The closest might be the constitutional "rational basis" test, which requires that, when the state acts it must do so rationally,[94] but even that doctrine is applied largely negatively (as a prohibition on considering certain

86. Fair Housing Act, 42 U.S.C. § 3605(a) (2018).
87. Equal Credit Opportunity Act, 15 U.S.C. § 1691 (2018).
88. 20 U.S.C. § 1681(a) (2018).
89. Korematsu v. United States, 323 U.S. 214, 216 (1944).
90. Craig v. Boren, 429 U.S. 190, 197 (1976).
91. United States v. Windsor, 570 U.S. 744, 770 (2013) (based on sexual orientation); Romer v. Evans, 517 U.S. 620, 632-33 (1996) (same); City of Cleburne v. Cleburne Living Ctr., Inc., 473 U.S. 432, 464-65 (1985) (intellectually disabled); U.S. Dep't of Agric. v. Moreno, 413 U.S. 523, 534-35 (1973) ("hippies").
92. N.Y.C. Transit Auth. v. Beazer, 440 U.S. 568, 609-11 (1979).
93. Deshaney v. Winnebago Cty. Dep't of Soc. Servs., 489 U.S. 189, 195 (1989).
94. *See generally* Thomas B. Nachbar, *The Rationality of Rational Basis Review*, 102 VA. L. REV. 1627 (2016).

characteristics[95]), and virtually any reason will do; the Court has not required that regulators consider specific factors in their decision-making.

The negative operation of discrimination fits with its aspiration to fairness—like fairness itself, discrimination law operates as a side constraint on other activity. There is no legal mandate to engage in the act of "nondiscrimination" but rather to avoid discrimination while doing something else, like lending or hiring and firing employees. We can all agree that avoiding discrimination might represent a higher value than any particular employment or lending decision, but that is true of all side constraints—they represent higher order values. It is a critical feature of side constraints that while they govern other activity, they do not supplant the goal of the particular system (of lending or employment, for instance) being regulated. One way to avoid doing so is by operating negatively rather than affirmatively.

Among U.S. discrimination laws, Title VII and constitutional equal protection are the most commonly treated in the literature, and I will follow the literature by focusing on those two regimes. Title VII liability is frequently used as a model when applying and interpreting other discrimination laws,[96] and the Equal Protection Clause's singular protection against government discrimination (in which the relevant actor is not a business but rather is a democratically accountable governmental body) requires courts to address the core of what constitutes illegitimate discrimination. Understanding how these laws work goes a long way toward understanding how our society operationalizes fairness in law.

A. Disparate Treatment, Disparate Impact, and Disparate Outcome

Understanding how discrimination prohibitions would work in the context of computational decision-makers first requires developing a general understanding about how discrimination law works. In the U.S., statutory protections (in the employment context, as an example) can give rise to both "disparate-treatment" and "disparate-impact" theories of liability,[97] and both terms are somewhat misleading.

"Disparate treatment" liability is the more straightforward, but it includes both cases involving facially disparate treatment (such as explicit race classifications) and intentional, but facially neutral,

95. *See supra* the text accompanying notes 90-91.

96. *E.g.*, Smith v. City of Jackson, 544 U.S. 228, 240-42 (2005) (comparing reach of ADEA by analogizing to Title VII); Smith v. Metro. Sch. Dist. Perry Twp., 128 F.3d 1014, 1022-28 (7th Cir. 1997) (drawing from Title VII in a Title IX case).

97. *See generally* Ricci v. DeStefano, 557 U.S. 557, 577-78 (2009) (describing the two theories and the underlying statute, 42 U.S.C. § 2000e et seq.). Not every discrimination statute permits both theories. *See*, *e.g.*, Gen. Bldg. Contractors Ass'n, Inc. v. Pennsylvania, 458 U.S. 375, 388-91 (1982) (requiring intentional discrimination for a § 1981 claim).

discrimination.[98] In the absence of a facial classification, a disparate treatment case can be supported either by direct evidence of discriminatory intent or indirect evidence of intent established via the *McDonnell Douglas* burden-shifting framework, under which a plaintiff establishes 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. After the plaintiff establishes the *prima facie* case, an employer must provide a legitimate, nondiscriminatory reason to explain the practice that caused the plaintiff's exclusion, after which the plaintiff has an opportunity to demonstrate the proffered justification is pre-textual.[99] Thus, except in disparate treatment cases involving intentional discrimination for which there is direct evidence, employers are allowed to offer justifications for their discriminatory practices.

"Disparate treatment" is imprecise if it is taken to cover both facial and intentional discrimination, since whether facial discrimination itself is problematic is to some degree a matter of intent. That intentional invidious discrimination is illegal is a commonplace (in both statutory and constitutional discrimination law), but there is considerable argument over whether the intent to assist historically disfavored groups can save either facial[100] or intentional but facially neutral discrimination.[101] Consequently, we should eschew the term "disparate treatment" other than as a term of art[102] to refer to that particular form of discrimination law inquiry and, when discussing particular forms of discrimination, describe them using more precise terms like "facial" (or conversely "facially neutral") and "intentional" (or conversely "unintentional") as appropriate.

"Disparate impact" is an even more problematic term of art in discrimination law, since it misleadingly describes as a theory of liability what is actually a trigger for a deeper inquiry. An observable,

---

98. *Ricci*, 557 U.S. at 577-78.

99. *See* McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973). Not all discrimination statutes providing for intentional discrimination follow the *McDonnell Douglas* approach. *See*, *e.g.*, Carroll v. Del. River Port Auth., 843 F.3d 129, 133 (3d Cir. 2016) (declining to apply the *McDonnell Douglas* framework to USERRA).

100. *See* Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 245 (1995) (Stevens, J., dissenting) ("The consistency that the Court espouses would disregard the difference between a No Trespassing sign and a welcome mat.") (internal quotations omitted); George Rutherglen & Daniel R. Ortiz, *Affirmative Action Under the Constitution and Title VII: From Confusion to Convergence*, 35 UCLA L. REV. 467, 468-69 (1988).

101. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 788 (2007) (Kennedy, J., concurring in part and dissenting in part) ("In the administration of public schools . . . it is permissible to consider the racial makeup of schools and to adopt general policies to encourage a diverse student body, one aspect of which is its racial composition.").

102. Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1350 (2010).

disparate impact is only the beginning of the matter.[103] Employers are allowed to offer evidence to show that the disparate impact of a particular practice is the result of a valid, job-related reason for the classification leading to the observable disparate impact.[104] Other statutes generally follow suit, allowing some form of business-related justification defense.[105] One could easily imagine a true "disparate impact" approach that would make the demonstration of the disparate impact the end of the matter, which would elevate the protected classification above business efficiency, but "disparate impact" theories do not do so—disparate impact liability requires much more than just a showing of a disparate impact. That "much more" is an inquiry into the reasonableness of the employment practice having the disparate impact. As a result, even disparate impact cases involve a necessarily normative inquiry into the employer's reasons for the practice that produces the disparate impact,[106] an inquiry that has little to do with the impact itself or the employees affected by the practice. Indeed, much of discrimination law and scholarship is consumed with exactly what it is that employers must show after an initial showing of a disparate impact.[107] Even at the stage of disparate impact, though, there is room for confusion; the connection between the *observable* disparate impact and any particular employment practice can itself be difficult to demonstrate, since employees are subject to any number of employment practices or exogenous circumstances.[108]

---

103. Kleinberg et al. *supra* note 14, at 115.

104. *See generally* Ricci v. DeStefano, 557 U.S. 557, 578 (2009). There is no single articulation of the test for employer justification. *See* Linda Lye, *Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 348-53 (1998) (describing four different standards of scrutiny applied by lower courts in the years following the Civil Rights Act of 1991). In many cases, the inquiry is complicated by the burden-shifting nature of the inquiry, which alternately places responsibility for the justification and its refutation in defendant and plaintiff. *See*, *e.g.*, Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (shifting of burdens regarding alternative employment practices).

105. Americans with Disabilities Act (ADA), 42 U.S.C. § 12113(a) (2018) (business necessity); *id.*, § 12113(b) (health and safety requirements); Age Discrimination in Employment Act (ADEA), 29 U.S.C. § 623(f)(1) (2018) ("a bona fide occupational qualification reasonably necessary to the normal operation of the particular business, or where the differentiation is based on reasonable factors other than age"); W. Air Lines, Inc. v. Criswell, 472 U.S. 400, 414 (1985) (ADEA permits age discrimination if there is "a factual basis for believing, that all or substantially all [persons over the age qualifications] would be unable to perform safely and efficiently the duties of the job involved") (internal quotations omitted); ADEA, 29 U.S.C. § 623(f)(2) (2018) (established employee benefits or seniority plan).

106. Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 753 (2006).

107. *See generally* Lye, *supra* note 104; Andrew C. Spiropoulos, *Defining the Business Necessity Defense to the Disparate Impact Cause of Action: Finding the Golden Mean*, 74 N.C. L. REV. 1479 (1996).

108. Wal-Mart, Inc. v. Dukes, 564 U.S. 338, 367, 355-56 (2011) ("[M]ost managers in a corporation . . . would select sex-neutral, performance-based criteria for hiring . . . . Others may choose to reward various attributes that produce disparate impact . . . . And still other

Observability is key and is likely to increasingly be so as computational decision-making becomes increasingly prevalent. It is frequently impossible to identify the specific practice producing a specific disparate impact on a particular individual or group. Instead, what we are most likely to observe are far more general "disparate outcomes"—practical, real-life consequences as experienced by particular individuals or groups that cannot be conclusively connected to any particular practice. Blacks are nearly six times more likely to be incarcerated than whites in the United States.[109] The disparity in incarceration rates is a readily observable outcome that is produced by any number of effects, some of which might be the result of discrimination against blacks (intentional or unintentional, facial or facially neutral) that produces effects that disparately impact blacks—and some not. The degree to which each subsidiary effect is either intentional or unintentional, disparate or not, is a matter of debate, but the disparate outcome of disparate incarceration rates is not.

As should be obvious from even this cursory discussion, easily observed disparate outcomes have a major effect on our perception of the fairness of a particular system or practice. What is even more important, for present purposes, is that the introduction of computational decision-makers is likely to make them even more so.

Computational decision-makers will produce much more consistent outcomes than human ones would, and they will do so inexpensively and therefore will likely be asked to do so more frequently.[110] The result will be that if a particular decision-making process has a disparate impact, we can expect many, many iterations of an identically disparate impact in rapid succession to produce readily observed disparate outcomes. The outrage sparked by a process that produces such disparate outcomes with such speed and consistency will be difficult to ignore. The COMPAS controversy provides an excellent example.[111] Because the same COMPAS software is used in many jurisdictions, it had the potential to produce biased outcomes that were both consistent and widespread, prompting much of the attention COMPAS has received.[112] Computational decision-makers can make consistent decisions at a scale that humans cannot match, which means they can also engage in discrimination on a grand scale.

---

managers may be guilty of intentional discrimination . . . . In such a company, demonstrating the invalidity of one manager's use of discretion will do nothing to demonstrate the invalidity of another's.").

109. John Gramlich, *The Gap Between the Number of Blacks and Whites in Prison is Shrinking*, PEW RES. CTR. (Apr. 30, 2019), https://www.pewresearch.org/fact-tank/2019/04/30/shrinking-gap-between-number-of-blacks-and-whites-in-prison/ [https://perma.cc/M2E3-URSC].

110. *See supra* the text accompanying notes 49-51.

111. *See supra* the text accompanying notes 1-7.

112. Angwin et al., *supra* note 2.

At the same time, discrimination cases readily lend themselves to concern about distributive consequences, especially when the victims of employment discrimination (to take an example) are likely to represent groups that have historically been discriminated against. Discrimination is not distributed randomly; it affects the same people repeatedly over time.[113]

B. Outcome versus Justification in Discrimination Law

Without minimizing the harms that accompany discrimination, it is important to recognize that discriminatory outcomes, and indeed discriminatory impacts, serve a very limited role in modern discrimination law. Although "disparate impact" is identified as a distinct theory (at least in employment discrimination law), it is only the beginning of the inquiry. Disparate impact claims ultimately depend not on the impact on the employee but rather on the employer's reasons for engaging in the challenged practice. That it does so is not because of the theory of discrimination underlying discrimination law—that we don't care about discrimination so long as employers can demonstrate the benefit of doing so. Rather, it is because of discrimination law's necessary nature as a side constraint on other productive behavior.

1. The Limited Role of Outcomes (or Impacts) in Discrimination Law

When an employee experiences a discriminatory outcome—or a group of employees observe a series of discriminatory outcomes—it is still a long way to a successful employment discrimination claim. There is the initial matter that it is essential for plaintiffs to distinguish generalized discriminatory outcomes from discriminatory impacts felt by them. Even in employment, wildly disparate outcomes might result from a combination of the employer's practices and social circumstances exogenous to the employment relationship. If an employer requires a college degree for a particular position, blacks will suffer a negative disparate impact as a result of that practice because they, as a population, are less likely to have graduated from college than whites,[114] for reasons unrelated to their work or the employer who has adopted the practice.[115]

---

113. *See generally* Lincoln Quillian et al., *Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time*, 114 PROC. NAT. ACAD. SCI. no 41 (Oct. 2017), https://doi.org/10.1073/pnas.1706255114 [https://perma.cc/LFX4-VRYL] (describing continuing trends of discrimination specifically toward blacks and Latinos over time).

114. Camille L. Ryan & Kurt Bauman, U.S. CENSUS BUREAU, *Educational Attainment in the United States: 2015*, 1 (2016) ("Educational attainment varied by . . . race.").

115. Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971) (citing educational differences between blacks and whites as the source of the disparity while assuming—in the face of

Even when plaintiffs do establish the necessary connection between a defendant's conduct and a disparate impact on them, the inquiry moves on to the defendant's justification for the practice, an inquiry that is largely insensitive to the disparity or magnitude of the impact on the employee. In the employment context, all that is necessary for the employer to rebut a disparate impact claim is to show some relationship between the practice and improved job performance. What degree of relationship is itself the subject of debate—the Supreme Court has been inconsistent in its articulation of the standard. In the seminal disparate impact case, *Griggs v. Duke Power*, the Court offered several different articulations of the necessary relationship, calling upon both "necessity" and a looser standard that the practice merely be "related" to job performance;[116] it was unnecessary in *Griggs* to distinguish between the two because the employer had failed to demonstrate any relationship between the practice and performance.[117] When the Court suggested "necessity" would raise an impossible standard in *Wards Cove Packing Co. v. Atonio*,[118] Congress got involved to clarify that business necessity was indeed part of the showing,[119] but in so doing it simply resurrected the earlier *Griggs* standard.[120]

Other disparate impact discrimination statutes work similarly, albeit with potentially different standards. The Fair Housing Act's disparate impact theory bans practices that produce

---

considerable past evidence of discrimination—that the defendant in the case was not discriminating intentionally through its educational requirements).

116. *Id.* at 431 ("The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."). *See also id.* at 432 ("Congress has placed on the employer the burden of showing that any given requirement must have a manifest relationship to the employment in question."); *id.* (noting that Title VII outlaws practices with a racially disparate impact that "are unrelated to measuring job capability").

117. *Id.* ("On the record before us, neither the high school completion requirement nor the general intelligence test is shown to bear a demonstrable relationship to successful performance of the jobs for which it was used.").

118.

> The touchstone of this inquiry is a reasoned review of the employer's justification for his use of the challenged practice. A mere insubstantial justification in this regard will not suffice, because such a low standard of review would permit discrimination to be practiced through the use of spurious, seemingly neutral employment practices. At the same time, though, there is no requirement that the challenged practice be "essential" or "indispensable" to the employer's business for it to pass muster: this degree of scrutiny would be almost impossible for most employers to meet, and would result in a host of evils we have identified above.

Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659 (1989).

119. Civil Rights Act of 1991, 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2018) ("[T]he respondent fails to demonstrate that the challenged practice is job related for the position in question and consistent with business necessity.").

120. Susan S. Grover, *The Business Necessity Defense*, 30 GA. L. REV. 387, 392-93 (1999). As Prof. Grover notes, because the pre-*Wards Cove* standard was itself inconsistent, the Civil Rights Act of 1991 did little to clarify the degree of relationship the defendant had to show to win a Title VII disparate impact case. *Id.* at 393. *See also* Lye, *supra* note 98, at 335.

"'disproportionately adverse effect on minorities' and are otherwise unjustified by a legitimate rationale."[121] The Age Discrimination in Employment Act's disparate impact theory permits employer decisions to have disparate impacts without incurring liability if the decision is based on "reasonable factors other than age,"[122] which the Court has interpreted as controlled by *Wards Cove*,[123] a lower standard than in Title VII disparate impact cases.[124]

Unsurprisingly, outcome (or impact) also has limited roles in both statutory and constitutional theories of liability that require *intentional* discrimination. Like disparate impact cases, statutory disparate treatment cases similarly search for the employer's reason for adopting the discriminatory practice and similarly (generally) allow even facially discriminatory practices if the employer offers a valid justification for them. Under the *McDonnell Douglas* burden-shifting framework, once the plaintiff establishes a *prima facie* case, an employer can rebut by showing a legitimate, nondiscriminatory reason to explain the plaintiff's treatment.[125] The same is true of constitutional race discrimination evaluated under the Equal Protection Clause of the Fourteenth Amendment.[126] In *Washington v. Davis*, the Court held that only intentional discrimination would trigger heightened scrutiny. Although the Court left a place for disparate impact as an indication of impermissible

121. Texas Dept. of Hous. and Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 524-25 (2015) (quoting Ricci v. DeStefano, 557 U.S. 557, 577 (2009)).

122. 29 U.S.C. § 623(0(1) (2018).

123. Smith v. City of Jackson, 544 U.S. 228, 240 (2005).

124. *See* Selmi, *supra* note 106, at 748.

125. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973). ("The burden then must shift to the employer to articulate some legitimate, nondiscriminatory reason for the employee's rejection."). Following that, the plaintiff has the opportunity to show that the proffered nondiscriminatory reason is merely pretextual.

The nature of the justification varies (in disparate treatment cases for sex, religion, or national origin, express discrimination must be justified by the higher "bona fide occupational qualification" or "BFOQ"). *See* 42 U.S.C. § 2000e-2(e) (defining "bona fide occupational qualification" as "reasonably necessary to the normal operation of that particular business or enterprise"). Sometimes justifications are prohibited (for instance, the BFOQ defense is not available for overt race discrimination). Courts have generally read the BFOQ standard narrowly. In the context of sex discrimination, the Supreme Court has stated the BFOQ "provides only the narrowest of exceptions to the general rule requiring equality of employment opportunities." Dothard v. Rawlinson, 433 U.S. 321, 333 (1977). It's not clear the BFOQ requirement does any work for religion given the exemption for religious qualifications for religious organizations and given the other exceptions for discrimination on the basis of religion in § 702. *See* GEORGE A. RUTHERGLEN & JOHN J. DONOHUE III, EMPLOYMENT DISCRIMINATION: LAW AND THEORY 505-06 (2019).

126. U.S. CONST. amend. XIV, § 1 ("No State shall . . . deny to any person within its jurisdiction the equal protection of the laws."). Although the Fourteenth Amendment is applicable only to the states, the Supreme Court has identified an identical protection in the Due Process Clause of the Fifth Amendment, which is applicable to the federal government. *See* Bolling v. Sharpe, 347 U.S. 497, 498-99 (1954).

intent, impact itself is not enough to trigger strict scrutiny, much less an equal protection violation in its own right.[127]

Thus, with the exception of intentional, overt discrimination (which itself is considered a both a form of *per se* disparate treatment in the employment context[128] and its own constitutional violation[129]), both types of cases conduct a similar inquiry[130] in somewhat different form: a search for the *reason* for the underlying the relevant practice. There is some difference in how good the reason has to be depending on the nature of the discrimination, but once that reason is discovered, it is the reason, not the impact, that determines the outcome of the case.

Notably, the law does not require a comparison of the magnitude of the discriminatory impact and the benefit conferred upon the entity engaged in the practice. The Americans with Disabilities Act requires employers to make "reasonable accommodation" for a covered individual's disability,[131] but the reasonableness of the accommodation is determined entirely with reference to factors relating to the employer and the cost of the accommodation, not the impact on the covered individual.[132] Similarly, the "business necessity" standard restored by the 1991 Civil Rights Act asks "whether there are other ways for the employer to achieve its goals that do not result in a

---

127. Washington v. Davis, 426 U.S. 229, 242 (1976) ("Disproportionate impact is not irrelevant, but it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution. Standing alone, it does not trigger the rule that racial classifications are to be subjected to the strictest scrutiny and are justifiable only by the weightiest of considerations.") (citations omitted).

128. *See* Primus, *supra* note 102, at 1351

129. *See* Loving v. Virginia, 388 U.S. 1, 11 (1967).

130. *See* Selmi, *supra* note 106, at 723-24 (noting how similar the Court's approach in disparate impact cases is to disparate treatment pretext cases). Indeed, the development of disparate impact theory itself could be seen as a correction for an inadequately searching approach to disparate treatment. *See generally* George Rutherglen, *Disparate Impact Under Title VII: An Objective Theory of Discrimination*, 73 VA. L. REV. 1297 (1987).

131. *See* 42 U.S.C. § 12112(b)(5)(A) (including "not making reasonable accommodations" in the definition of the term "discriminate against a qualified individual with a disability").

132. *See* 42 U.S.C. § 12111(10)(B):

> In determining whether an accommodation would impose an undue hardship on a covered entity, factors to be considered include--
> (i) the nature and cost of the accommodation needed under this chapter;
> (ii) the overall financial resources of the facility or facilities involved in the provision of the reasonable accommodation; the number of persons employed at such facility; the effect on expenses and resources, or the impact otherwise of such accommodation upon the operation of the facility;
> (iii) the overall financial resources of the covered entity; the overall size of the business of a covered entity with respect to the number of its employees; the number, type, and location of its facilities; and
> (iv) the type of operation or operations of the covered entity, including the composition, structure, and functions of the workforce of such entity; the geographic separateness, administrative, or fiscal relationship of the facility or facilities in question to the covered entity.

disparate impact on a protected class,"[133] a least-discriminatory means test.[134] But even under that test, the employer's goals remain paramount.[135] And none of the tests resemble anything like a balancing of the costs of the disparate impact on employees with the benefits of the discrimination for employers. To the law, the magnitude of the impact on the plaintiff is irrelevant once the impact itself is established.[136] At that point, the inquiry shifts entirely to the defendant's reason for the challenged practice, and the disparate outcomes and impacts that likely prompted the litigation become legally irrelevant.

The problems of focusing on disparity of outcomes is exemplified by the conflict between forms of discrimination in the 2009 case of *Ricci v. DeStefano*.[137] In *Ricci*, the New Haven Fire Department chose to invalidate a set of employment test results when they produced a racially disparate result.[138] White firefighters sued, arguing that the city's race-conscious decision to avoid a disparate impact was itself a Title VII disparate treatment violation. It is not clear that one can take much from *Ricci*, which is a fact-bound case decided based on a difference between proof standards applicable to Title VII claims—it is hardly a bellwether for discrimination cases.[139] But one thing the Court made clear was that correcting for racially imbalanced outcomes

133. Smith v. City of Jackson, 544 U.S. 228, 243 (2005) (interpreting 42 U.S.C. 2000e-2(k)(1)).

134. Ricci v. DeStefano, 557 U.S. 557, 578 (2009).

135. Grover, *supra* note 120, at 425-26; Gary A. Moore & Michael K. Braswell, *"Quotas" and the Codification of the Disparate Impact Theory: What Did Griggs Really Say and Not Say?*, 55 ALB. L. REV. 459, 492 (1991) (acknowledging that a court's inquiry in a disparate impact case does not involve a balancing harm to a plaintiff against the business necessity). There are some outlier cases that suggest such balancing. *Cf.* Nash v. Consol. City of Jacksonville, 895 F. Supp. 1536, 1545 (M.D. Fla. 1995) (requiring a business justification to be sufficiently compelling such that it overrides the harm done to racial minorities through its disparate impact). Notably, although the *Nash* court articulated a balancing test, it did not itself apply a balancing test. The court accepted the City's use of a written promotion examination as a "business necessity" without commenting on either the quantum of benefit from doing so or the quantum of harm to the plaintiff from the test, much less the balance between the two. *See id.* at 1544-49 (analyzing the reliability of the test but not the magnitude of its benefits). *See* Lye, *supra* note 98, at 349-50 (describing the difficulty of making sense out of the *Nash* court's analysis). *See also* Pamela L. Perry, *Two Faces of Disparate Impact Discrimination*, 59 FORDHAM L. REV. 523, 583-85 (1991) (describing the differences between the least restrictive means and harm-balancing approaches to disparate impact cases). Given the incommensurability of business benefit and discriminatory harm, it's hard to imagine how a court could sensibly balance the two. *See generally* Cass R. Sunstein, *Incommensurability and Valuation in Law*, 92 MICH. L. REV. 779, 795-96 (1994).

136. *Cf.* Dwork et al., *supra* note 15, at 221 (suggesting a mathematical solution that would factor in the lost business efficiency and balance it against harm to the discriminated-against individual—or group—resulting in a "bicriteria optimization problem, with a wide range of options").

137. Ricci v. DeStefano, 557 U.S. 557 (2009).

138. The test was only one part of a selection process, but it produced a pool of promotion candidates that excluded the possibility of promoting any black candidates. *See* Ricci v. DeStefano, 554 E Supp. 2d 142, 145 (D. Conn. 2006); Primus, *supra* note 102, at 1348.

139. *See* George Rutherglen, *Ricci v. DeStefano: Affirmative Action and the Lessons of Adversity*, 2009 SUP. CT. REV. 83 (2009).

does not qualify for any kind of deference and is likely to lead to liability under Title VII.[140]

Discrimination law avoids baking race discrimination into antidiscrimination regimes by limiting the role of disparate impact to that of a potential trigger for a deeper inquiry into the defendant's reasons for adopting the policy resulting in the discriminatory impact. Recognition of the role of the defendant's reasons in discrimination law has two important implications. First, it shows how close in practice disparate impact theories are to disparate treatment theories and why "disparate impact" is a misnomer as a theory of liability—it only describes the first stage of the inquiry. Even in disparate impact cases, employers are allowed to offer justifications for their practices—a process that largely resembles the *McDonnell Douglas* framework for intentional discrimination cases.[141] That is why disparate impact has largely failed to deliver on its promise to vastly expand the reach of anti-discrimination protections. As Michael Selmi explains, "The expectation that these claims would be easier to establish than intentional discrimination claims rests entirely on the first part of the theory regarding the prima facie case of discrimination, but ignores the business necessity prong, which has always proved the greater hurdle."[142]

Second, acknowledging the role of justification in disparate impact cases similarly demonstrates the deceptively narrow gap between statutory and constitutional anti-discrimination rules. Although frequently attacked,[143] the rule of *Washington v. Davis* requiring that discrimination must be intentional in order to prompt heightened constitutional scrutiny does not place constitutional discrimination law on a fundamentally different footing than the strictest U.S. statutory discrimination law; it is easy to overstate just how much the intentional discrimination rule of *Washington v. Davis* narrowed constitutional equal protection review.[144] After disparate impact became part of Title VII, the Court moved to provide greater deference to employers in disparate impact cases until Congress stepped in with the Civil Rights Act of 1991.[145] Had *Washington v. Davis* come out the other way and produced a constitutional disparate impact theory, it is hard to imagine that the Court would not have incorporated a similar

---

140. *See* Primus, *supra* note 102, at 1364 (on his "institutional" reading of *Ricci*: that "courts may order race-conscious remedies for disparate impact problems, but public employers may not"); *id.* at 1363 (on his "general" reading of *Ricci* as a prohibition on disparate treatment to correct disparate impact). See also Kleinberg et al., *supra* note 14, at 124.

141. Selmi, *supra* note 106, at 749.

142. *Id.*

143. *See*, *e.g.*, David A. Strauss, *Discriminatory Intent and the Taming of* Brown, 56 U. CHI. L. REV. 935, 1000-14 (1989); Reva Siegel, *Why Equal Protection No Longer Protects: The Evolving Forms of Status-Enforcing State Action*, 49 STAN. L. REV. 1111, 1145 (1997).

144. Selmi, *supra* note 106, at 753-54.

145. Grover, *supra* note 120, at 391-92.

degree of deference to regulators' justifications under equal protection review that it had applied to employers' justifications under Title VII. And in the case of a constitutional rule, congressional correction like that contained in the Civil Rights Act of 1991 would not have been available.

In the end, disparate treatment and disparate impact statutory cases, and constitutional and statutory discrimination cases are generally consistent and do seem to be at least partially unified by a theory of discrimination, at least if one looks at the entire inquiry and not just at the requirements for a *prima facie* case. Both systems largely ignore hugely discriminatory impacts so long as those impacts are produced by practices related to increasing business or regulatory efficiency.

The reason why discrimination law overlooks so many harms to focus only certain kinds of discrimination has less to do with discrimination law itself than it does the broader consequences of adopting rules that focus exclusively on impact. As the Supreme Court explained in the context of constitutional discrimination:

> A rule that a statute designed to serve neutral ends is nevertheless invalid, absent compelling justification, if in practice it benefits or burdens one race more than another would be far-reaching and would raise serious questions about, and perhaps invalidate, a whole range of tax, welfare, public service, regulatory, and licensing statutes that may be more burdensome to the poor and to the average black than to the more affluent white.[146]

And in the context of federal employment discrimination law, "[t]here are societal as well as personal interests on both sides of this equation. The broad, overriding interest, shared by employer, employee, and consumer, is efficient and trustworthy workmanship assured through fair and racially neutral employment and personnel decisions."[147]

The Court confronted the same problem under Title VII and similarly rejected broad Title VII liability for similar reasons. When, in *Wards Cove Packing Co. v Antonio*,[148] the Court considered the possibility that the "consistent with business necessity" requirement of Title VII[149] would require actual *necessity*, it rejected the standard as unworkable for reasons resembling those in *Washington v. Davis*.[150]

---

146. Washington v. Davis, 426 U.S. 229, 248 (1976).

147. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973).

148. Wards Cove Packing Co. v. Antonio, 490 U.S. 642 (1989).

149. 42 U.S.C. §2000e-2(k)(1)(A)(i).

150. 490 U.S. at 659 ("[T]here is no requirement that the challenged practice be 'essential' or 'indispensable' to the employer's business for it to pass muster: this degree of scrutiny would be almost impossible for most employers to meet, and would result in a host of evils we have identified above.") Congress reversed *Wards Cove* with the 1991 Civil Rights Act, but that simply reverted to the standard applicable in *Griggs*—it did not result in the

2. Discrimination Law as a Side Constraint

Thus, the limits on discrimination law, which arguably allow a huge amount of discrimination to go un-addressed, are the result of discrimination law's potential effects on *other* aspects of productive activity. Like fairness itself, anti-discrimination must operate as a side constraint on other systems lest the absence of discrimination (and specifically discriminatory impact) become the principal goal of all productive activity. Modern discrimination law implements the value of anti-discrimination like a side constraint by establishing a threshold prohibition on discrimination that must be satisfied in all cases (such as the prohibition on intentional discrimination). But beyond that threshold point, anti-discrimination mandates give way to gains in productive activity.

The possibility that broad anti-discrimination mandates might derail other aspects of productive decision-making is easily forgotten. Discriminatory outcomes are readily observable and highly salient—coming as they do in the form of lost job or a life in poverty—making them ready fodder in debates over injustice and unfairness. It is hard to argue against broad discrimination rules when one considers the harms that flow from discrimination.

But it was exactly these second- and third-order effects of broad antidiscrimination rules that have driven the Supreme Court to limit the reach of discrimination law in both the constitutional and statutory contexts. In *Washington v. Davis*, when ruling that intent was a necessary component to triggering heightened scrutiny of a facially neutral restriction, Justice White pointed to the possibility that an effects-based discrimination regime might require other areas of law to be re-ordered in order to produce racially balanced outcomes. For instance, if blacks are historically paid less than whites (or discriminated against in credit terms), minimum wage laws might have the effect of shifting jobs from blacks to whites (or usury laws might prevent blacks from borrowing at all). A rule requiring heightened scrutiny for such laws would likely lead to the invalidation of any number of legal regimes whose purpose is unrelated to discrimination,[151] since virtually no regime predicated on economic

codification of the actual necessity requirement the Court had held up as impossible in *Wards Cove*. *See supra* text accompanying note 120.

151. *See Washington*, 426 U.S. at 248, n. 14. As an example, Justice White pointed to an argument advanced for an effects test:

> The fourteenth amendment should protect blacks from government discrimination whether it is intentional or unintentional, and whether it is the result of economic or any other category of legislation.
> What is proposed is a new standard for judicial review. Courts should not defer to legislative judgments and priorities when the enactments that embody them have racially discriminatory impacts. Any law that has this result must be supported by a compelling government interest.

effects alone would likely survive the strict scrutiny applied to race discrimination.[152]

The Court used intent to impose that limit in *Washington v. Davis*, but even as the Court was acknowledging in *Griggs* that Title VII's disparate impact theory went beyond intentional discrimination, it relied upon similar considerations of effects beyond discrimination law to limit the reach of unintentional discrimination: "Congress has not commanded that the less qualified be preferred over the better qualified simply because of minority origins. Far from disparaging job qualifications as such, Congress has made such qualifications the controlling factor."[153] When considering the alternative intentional theory of discrimination in *McDonnell Douglas*, the Court relied upon exactly the same, external, limits to discrimination law.[154] One consistency among discrimination law, statutory or constitutional, intentional or unintentional, is that limits of anti-discrimination largely come from outside discrimination law, not within it.

C. Translating Discrimination Law to Algorithmic Discrimination

The application of discrimination law to algorithmic discrimination presents a host of both challenges and opportunities. As mentioned above, the systematization and reduction in cost of complex decision-making permitted by computerizing it is likely to lead to an explosion of outcomes, many of them likely to be disparate along historically important categories, such as race and sex.[155] Those outcomes are only a starting place for discrimination law, which requires a deeper inquiry into the justification for practices that lead to disparate outcomes.

As should now be obvious, determining the justification—or indeed the specific decision causing a particular effect—of any particular computational decision is not easy, or at least is a wildly variable task whose difficulty and efficacy is highly dependent on the particular form of computational decision-maker.[156] In the case of computational decision-making produced through machine learning, it may be impossible for humans to determine what factors went into a

---

William Silverman, *Equal Protection, Economic Legislation, and Racial Discrimination*, 25 VAND. L. REV. 1183, 1203 (1972), *cited in Washington*, 426 U.S. at 248, n. 14.

152. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701 (2007); Adarand Constructors, Inc. v. Peña, 515 U.S. 200 (1995). One commentator cited by Justice White argued that the requirement that doctors be licensed itself should be struck because of the disparate effect it had on the ability of blacks to access health care and should instead be replaced with a system of "certification" that would reduce the barriers that blacks would have to lower quality, but better than no, health care. Silverman, *supra* note 151, at 1200-01.

153. Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971).

154. *See supra* the text accompanying note 147.

155. *See supra* the text accompanying notes 49-51.

156. *See* Kroll, et al., *supra* note 14, at 643-52; Lehr & Ohm, *supra* note 14, at 705-10.

particular decision.[157] But even with regard to traditionally programmed and rule-based decision-makers, the answers may be unsatisfying.

It is not a question *whether* discrimination law will be applied to computational decision-making; it will be. The question is *how* discrimination law will have to adapt to computational decision-making and how computational decision-making will have to adapt to discrimination law.

III. IMPLICATIONS OF DISCRIMINATION LAW FOR ALGORITHMIC DECISION-MAKING AND IMPLICATIONS OF ALGORITHMIC DECISION-MAKING FOR DISCRIMINATION LAW

As the previous section suggests, although "algorithmic fairness" may be a hopeless aspiration, discrimination law has implications for the delegation of increasingly sophisticated decision-making to computational decision-makers. The opportunity to codify both discriminatory (or anti-discriminatory) preferences in the same code that performs business functions is an invitation to enshrine preferences regarding discriminatory outcomes (or the lack thereof) in computational decision-makers. That way lies peril, but the possibility does present a valuable opportunity to reconsider discrimination law as applied not only to computational but also human decision-makers. Computational decision-makers do present some fundamental differences from humans, leading many to call for increased transparency for computational decision-making.[158] But transparency is not a particularly important feature of discrimination law, even as applied to human decision-makers. What discrimination law demands is not transparency but rather accountability, and it already contains a mechanism for providing that accountability, one that is readily applicable to computational decision-makers. One way computational decision-making differs from human decision-making is that the flexibility and regularity of computational decision-makers provide opportunities for correcting decision-making processes that systematically produce discriminatory outcomes. Intentional tuning of algorithms to produce racially balanced outcomes is illegal, but that does not mean that those who employ computational decision-makers are bound to discriminatory outcomes produced by those systems. Rather than high-minded policy, the law requires an extremely practical understanding of how such processes are modified, and sensitivity to the series of decisions that lead to the modification of a system to eliminate discriminatory disparities.

157. Lehr & Ohm, *supra* note 14, at 707.

158. *See supra* the text accompanying notes 16-18.

A. Keeping Side Constraints on the Side

Perhaps the most profound implication of algorithmic decision-making for discrimination law is the opportunity algorithmic discrimination presents for thinking about discrimination law. The capacity of algorithmic decision-makers (especially rule-based ones) to rationalize exactly how they discriminate in both beneficial and invidious ways[159] requires us to confront exactly what it means to discriminate and what forms of discrimination our society is going to prohibit. That is the process that policymakers, computer scientists, and lawyers are going through right now by writing (and occasionally reading) papers like this one. The ability to program racial parity in computational decision-makers forces us to confront the degree to which anti-discrimination rules must operate as side constraints rather than principal goals lest the values of anti-discrimination convert all productive activity into a massive policy of intentional discrimination to avoid discriminatory outcomes. What *Ricci* makes clear is that forgetting that discrimination is a side constraint will lead to intentional discrimination, and that is true whether or not one believes such race-conscious interventions are legitimate.[160]

The kind of mechanized, but hidden (inside a computer) forms of discriminatory-treatment-as-correction enabled by computational decision-makers raises the specter of never-ending systematized discrimination. After all, once corrective disparate treatment is programmed into our computational decision-makers, inertia may become a strong force for continuing the policy, and there may be no readily recognizable moment at which to remove it.[161] Inexpensive and mechanized disparate treatment will reduce the costs (both economic and social) of optimizing for racial parity, and so we need to pay even more attention to second-order reasons for not doing so.

B. Transparency, Accountability, and Liability

One largely uncontroversial proposal for dealing with potential discrimination by computational decision-makers is to increase their transparency.[162] It might be easier to do in the case of rule-based decision-makers, when each practice is captured in code and effects can be disaggregated and measured mathematically.[163] In other cases,

---

159. Kleinberg et al., *supra* note 14, at 119-20.

160. *See supra* the text accompanying notes 137-140.

161. *Cf.* Grutter v. Bollinger, 539 U.S. 306, 343 (2003) ("It has been 25 years since Justice Powell first approved the use of race to further an interest in student body diversity in the context of public higher education. . . . We expect that 25 years from now, the use of racial preferences will no longer be necessary to further the interest approved today.").

162. PASQUALE, *supra* note 14, at 4; Bloch-Wehba *supra* note 14, at 1265; Citron, *supra* note 14, at 1308; Raghavan, *supra* note 15, at 478.

163. *See supra* text accompanying notes 16-18. *See* Houser, *supra* note 14, at 294; Huq, *supra* note 14, at 618; Kleinberg et al., *supra* note 14, at 114; Ben Wagner, *Algorithmic*

transparency might not be enough, but most if not all[164] agree there should be more of it for algorithmic decision-makers. Discrimination law, though, shows that transparency might be not only unnecessary but even not useful to preventing discrimination.

Whether more transparency is good depends on what the objective is. Discrimination law generally seeks justification. Thus, it's helpful to distinguish between transparency—the ability to view the working of a system—and accountability, an explanation for why the system is operating as it does.[165] Zimmerman and Cabinakova helpfully distinguish between transparency and accountability, with transparency as an enabler (among others) to providing accountability.[166] Transparency might be necessary to providing the explanation inherent in accountability,[167] but they are not the same thing, especially when design rules are contained in code and implemented by machine.

Discrimination law provides a structure for how to evaluate particular decisions, but it does so through a process that interrogates human decision-makers for their justifications. The question is how to adapt those practices to algorithmic decision-makers. Paying attention to how discrimination law works provides insight into how to deal with similar issues as they arise in computational decision-makers.

As a comparative matter, it is not clear whether algorithmic decision-makers could be any more opaque than their human counterparts.[168] When one considers the process of human decision-making, transparency is in short supply. Unless one is telepathic, it is impossible to view the decision-making process contained in the "black box" that is another person's brain. Both the *McDonnell Douglas* burden-shifting framework and the *Griggs* approach to disparate impact are designed to reveal whether there *exists* a justifiable reason for the practice, but the existence of a justification does not tell you

---

Regulation and the Global Default: Shifting Norms in Internet Technology, 1 ETIKK I PRAKSIS 5 (2016).

164. Here Richard Primus's concerns over "visibility" might actually argue in favor of less transparency if the disparate treatment is in service of reducing disparate impacts. *See* Primus, *supra* note 201, at 318 ("On a visibility reading of the caselaw, then, equal protection limits disparate-impact remedies to those that minimize the visibility of their own race-consciousness—including, perhaps crucially, by avoiding the imposition of concrete costs on determinate and innocent third parties.").

165. Vedder et al., *supra* note 16, at 206 ("Accountability is the ability to provide good reasons in order to explain and to justify actions, decisions and policies for a (hypothetical) forum of persons or organisations."); Maranke Wieringa, *What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability*, PROC. OF THE 2020 CONF. ON FAIRNESS ACCOUNTABILITY AND TRANSPARENCY 1, 4 (2020) (equating "accountability" with "explanation," which she considers to be something greater than mere transparency); Yeung, *supra* note 60, at 516-17.

166. Zimmerman & Cabinakova, *supra* note 16, at 263-64.

167. Vedder et al., *supra* note 16, at 214-15.

168. Houser, *supra* note 14, at 293; Huq, *supra* note 14, at 640-46; Kleinberg et al., *supra* note 14, at 42; Chander, *supra* note 14, at 1030.

whether the decision-maker relied on it in making their decision. The tests are designed with the possibility of pretext in mind, but that doesn't mean there isn't a lot of room for hidden discrimination to take place before it rises to the level of (detectable) pretext.

Thus, what anti-discrimination regimes truly seek is not transparency—which for human decision-makers is unattainable—but, rather, the explanation that constitutes accountability. Rules like *McDonnell Douglas* require employers to offer up a rationale, and then it is up to us to decide whether to accept it, both as a matter of its legitimacy (whether it is a valid requirement) or credibility (whether it is actually held and caused the relevant decision).[169] We might choose to believe the human decision-makers or not, but we know who to believe and on what basis they are asking us to believe them.

In the case of computational decision-makers, concepts like whether we should "believe" their stated justifications simply do not apply, which means that even perfectly transparent decision-makers might not provide the requisite accountability. The one thing not revealed even by a perfectly transparent machine is its purpose. Unlike humans, computers are not capable of answering open-ended questions like "Why?" Consequently, accountability looks somewhat different for computational decision-makers. Sometimes the justifications will not even be stated (as might the case for machine learning systems), and for rule-based decision-makers, the code will offer transparency as to what the program is doing but not what the purpose was behind it. Even assuming there are no facially racist variables in the algorithms, in many cases, non-facially racial criteria might have arisen as proxies for race.[170] It will likely be impossible to say what the purpose is of many computational decisions. The most widely available information will be in the form of outcomes; depending on the nature of the decision-maker, that might be the best information we have.[171]

The danger is obvious: Without the ability to call upon purpose in answering observable racially disparate outcomes, the danger is that we will instead design systems in order to avoid those racially disparate outcomes and in so doing realize the fear expressed by Justice White in *Washington v. Davis*. And whether you agree or

169. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802, 804 (1973) (explaining that after the plaintiff makes a prima facie case "[t]he burden then must shift to the employer to articulate some legitimate, nondiscriminatory reason for the employee's rejection" but the plaintiff must then "be afforded a fair opportunity to show that [the employer's] stated reason for respondent's rejection was in fact pretext.").

170. Kleinberg et al., *supra* note 14, at 9;. Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 313-17 (2020) (explaining whether something should be considered a proxy for race or simply correlates with race is its own question). *See generally* Jung et al., Omitted and Included Variable Bias in Tests for Disparate Impact 2 (August 30, 2019) (unpublished manuscript), https://arxiv.org/abs/1809.05651 [https://perma.cc/8Q8V-GDSK].

171. *See supra* the text accompanying notes 111-112.

disagree with *Washington v. Davis*'s requirement of purposeful discrimination in order to trigger heightened scrutiny, we can all agree that elevating the importance of race in the design of computational decision-makers is problematic. Doing so threatens to enshrine in code race's role in organizing our society.

If regulating on the basis of outcome alone is a mistake and justification is not readily attainable from computational decision-makers, the question is how to get accountability. One way might be to insist that certain decisions must be made by humans,[172] who can then be held to answer for their actions—to require that certain decisions be made only by someone who can answer the open-ended "Why?" and then allow us to either believe them or not.

But if we are going to allow computational decision-makers to either make their own decisions or, like drug-sniffing dogs,[173] so completely influence human decision-making as to essentially supplant human discretion, how will we provide that accountability? It may not be so hard. Holding humans to account for the discriminatory decisions made by their computational counterparts actually looks a lot like the system we have now, which means that the best answer might be that the current system does not require much modification at all. The way discrimination law does so is by relying on the potential for *liability* to provide an incentive for decision-makers to be able to explain themselves.[174]

The delegation of decision-making to computational decision-makers enabled by advanced computing approaches is in practice not very different than the *Griggs*-era delegation to human-driven human resource "systems" through employer policies and practices.[175] As applied to computational decision-makers, the burden-shifting of current discrimination law will provide potential defendants the incentive to manage their computational decision-makers in a way that allows them (the humans) to explain the "Why?" behind those decisions. Otherwise, a disparate impact sufficient to establish a *prima facie* case that is not responded to with some permissible justification will lead to liability. Just as employers have an incentive

---

172. This is the approach taken by the European Union General Data Protection Regulation. *See General Data Protection Regulation: Automated Individual Decision-making, Including Profiling* 2016/679, art. 22, para. 1 2016 O.J. (L 119) 46 1 ("The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.").

173. At least one scholar has analogized algorithms to drug-sniffing dogs, who provide most of the information a police officer would use in that context in deciding whether to search. *See* Rich, *supra* note 14, at 918-20 (likening "automated suspicion algorithms" as providing an input to a human decision-maker akin to that of a drug-sniffing dog); *see also supra* text accompanying note 39.

174. *Cf.* Kleinberg et al., *supra* note 14, at 35-36 (discussing the various stages of discrimination litigation and how they might apply to algorithmic decision-makers).

175. Griggs v. Duke Power Co., 401 U.S. 424, 427 (1971).

to document their hiring and promotion practices (that is, their algorithmic delegations embodied in employer practices) today, they will have an incentive to capture and explain their practices tomorrow. So, although it is possible to develop a principled approach to reason-giving in algorithmic decision-makers,[176] as a practical matter, it's not necessary given current law. The only necessary modification (or really, adaptation) to the law would be to prohibit employers from defending against discrimination suits on the basis that a machine rather than a human made the relevant decision.[177] So long as there is no such defense, the potential for liability will provide employers an incentive (an optimal one if the amount of damages is correctly calibrated) to provide explanations for the discriminatory conduct of computational decision-makers. Transparency might be unnecessary in most cases (such as those in which outputs do not raise an inference of discrimination) and might be unnecessarily expensive to provide even in the cases in which it would be desirable. If properly calibrated to the harm caused by otherwise hidden decisions, liability can buy the accountability we desire, with or without transparency.

C. Dealing with Disparate Outcomes

There is no question whether some computational decision-makers will produce racially disparate outcomes—they will. The question is what the human owners of those systems should do about it. Tuning computational decision-makers to produce racially optimized outcomes is legally problematic,[178] but it seems equally wrong that our hands would be tied to accept what we all acknowledge to be poor outcomes produced by systems we control, delegation or not. There is a way out of the conundrum, one that requires attention to exactly how such corrections would be made.

---

176. *See* Huq, *supra* note 14, at 662-63.

177. On the theory that machines cannot form discriminatory intent: There is considerable debate in the computer science literature whether computers can form simple "intent" to carry out an action, but no serious argument that they can develop their own motivation for conduct, which is what discriminatory intent is. *See generally*, STEPHEN OMOHUNDRO, THE BASIC AI DRIVES, (2008); ELIEZER YUDKOWSKY, COMPLEX VALUE SYSTEMS IN FRIENDLY AI, 388, 389-90 (2011) ("It is not as if there is a ghost-in-the-machine, with its own built-in goals and desires (the way that biological humans are constructed by natural selection to have built-in goals and desires) which is handed the code as a set of commands, and which can look over the code and find ways to circumvent the code if it fails to conform to the ghost-in-the-machine's desires. The AI is the code."). This might be the most valuable aspect of disparate impact liability: to require defendants to provide some justification lest the existence of the discriminatory impact itself establish liability. The outcome may not be different than in disparate treatment cases, but the freeing the plaintiff from having to frame their complaint in terms of intent of any kind is particularly suited to a world with algorithmic decision-makers.

178. *Compare* Kroll et al., *supra* note 14, at 694, *and* Barocas & Selbst, *supra* note 14, at 726 (likely prohibited) *with* Kim, *supra* note 14, at 925-26, *and* Hellman, *supra* note 7, at 862-64 (not prohibited). *See also* Yang and Dobbie, *supra* note 170, at 4-5 (suggesting the use of race to de-bias data and arguing that doing so would "uphold the primary principles underlying the Equal Protection doctrine").

1. The Practicalities of (Remedying) Discrimination in Algorithmic Decision-makers

It is a commonplace that we do not want algorithms to be racist; the real question is what that means in terms of law and policy as applied to algorithmic decision-makers and the humans who (help) create and operate them. Although cases like *Ricci* suggest at least some tension between avoiding disparate impact and engaging in disparate treatment—the possibility that attempts to avoid or correct a disparate impact will themselves constitute disparate treatment[179]—it cannot be the case that decision-makers (human or computational) are prohibited from attempting to avoid invidious discrimination in their decision-making. The existence of Title VII and the categories of heightened scrutiny in constitutional equal protection doctrine are indications that it is permissible to elevate concerns about certain kinds of discrimination over others—disparate treatment means that we have to treat all people equally, but it does not mean we have to treat all forms of discrimination equally. Race (and sex, religion, and other prohibited categories) does matter. But if avoiding discrimination is a permissible (indeed a laudable) goal, certainly something can be done to program computational decision-makers to avoid discriminatory outcomes without running afoul of anti-discrimination rules.

For instance, consciously using data that does not reflect the effects of past discrimination seems unproblematic.[180] No one has a legitimate interest in including the effects of past discrimination in one's decision-making models and every reason to avoid it. As such, selecting data that is free of racial bias (like any bias) would seem to further both the interests in productivity privileged by discrimination law and the normative interests underlying discrimination law itself. That method, however, is also of extremely limited use. That understanding is as old as disparate impact liability itself. In *Griggs*, the Court pointed out that the disparity in performance on the tests used by Duke Power was the result of disparities in the educational opportunities afforded to blacks in segregated schools,[181] something having little to do with Duke Power's business or operations.

The approach of identifying discrimination in data before it is used to build a model also ignores an excellent source of information about

179. Primus, *supra* note 102 at 1350 (on the tension between disparate treatment and disparate impact).

180. *See* Chander, *supra* note 14, at 1044 ("An affirmative action approach would seek to ensure that the data used to train an algorithm are evaluated for being embedded with viral discrimination."); Sandvig et. al, *supra* note 15, at 4979; Yeung, *supra* note 60, at 516. It may, however, be impossible given how deeply entrenched inequality is in today's data, giving rise to the suggestion that algorithmic decision-making based on historical data may be fundamentally flawed and should be jettisoned entirely. *See* Mayson, *supra* note 14, at 2277.

181. Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971).

the discriminatory data used to train a model or validate a procedure: the outcomes produced by the model itself.[182] New Haven didn't discover the racially disparate outcomes of its tests until it received the results.[183] The discrimination inherent in data used to train a model is likely to be discovered by operation of the model, in which case the user of the model will be in the situation of tuning the algorithm to provide (comparatively) racially balanced outcomes, raising exactly the paradox posed by *Ricci*.[184]

On the opposite extreme, some have proposed explicitly taking race into account within algorithmic decision-making either to detect[185] racially imbalanced outcomes or specifically to include for correction of certain forms of racial bias.[186] The former may be possible, but the latter seems to be largely foreclosed in the statutory context by Title VII's disparate treatment prohibition under *Ricci*[187] and in the constitutional context by a long line of cases running from *J.A. Croson* to *Adarand Constructors* to *Parents Involved in Community Schools*, which requires the application of strict scrutiny to race classifications regardless of their underlying purpose.[188]

The real question, then, is the degree to which courts will allow the racially conscious use of facially neutral characteristics in the design and operation of computational decision-makers. This is where the nature of computational, rather than human, decision-making makes such an important difference in the application of discrimination law it at least two important and related ways.

First, because we can assume that most organizations will not intentionally build racist algorithms, it is likely that most algorithmic decision-makers will be designed *ab initio* to serve a particular business purpose without much thought to race. Consequently, it is

---

182. For instance, in the same paragraph in which Prof. Chander suggests using good data to train algorithms, he goes on to say, "Such an approach would require companies to anticipate how their algorithms are likely to operate in the real world and to review those operations for discriminatory results," which would require the kind of reconsideration and revision at issue in *Ricci*. Without checking and evaluating outcomes, it's practically impossible to know the degree to which the data set one started with is itself the product of past discrimination. Chander, *supra* note 14, at 1044.

183. Ricci v. DeStefano, 557 U.S. 557, 564-66 (2009).

184. *See* Chander, *supra* note 14, at 1041 ("The counterintuitive result of affirmative action is that the decisionmaker must take race and gender into account in order to ensure the fairness of the result.").

185. Kim, *supra* note 14, at 880.

186. Hellman, *supra* note 7, at 834-40; Kleinberg et al., *supra* note 14, at 127.

187. *See* Erin Kelly & Frank Dobbin, *How Affirmative Action Became Diversity Management: Employer Response to Antidiscrimination Law, 1961 to 1996*, 41 AM. BEHAV. SCIENTIST 960, 971-73 (1998) (describing the evolution of "affirmative action" programs to "diversity management" outreach-oriented programs following *Adarand Constructors*).

188. City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493-94 (1989); Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 227 (1995); Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 720 (2007) ("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.").

most likely to be in the *modification* (to resolve unintended disparate outcomes) of algorithmic decision-makers that racial considerations are likely to figure. The iterative nature of computer systems development[189] particularly lends itself to the use of facially neutral classifications to further racially conscious ends. The need to explicitly state decision criteria, especially in rule-based systems, means that any classification used by the system must be explicit in order to operate, and so prohibited facially racial considerations will be readily detected. Moreover, the ease of modifying and testing computational decision-making systems provides ample opportunity for tuning to produce what are considered by programmers to be optimal (racially balanced) outcomes.

Second, assuming that human programmers are savvy enough not to program race classifications, the primary source of information regarding actionable discrimination will not be in the decision-making algorithms themselves but in the human interaction with the computational decision-making system—specifically (re)programming rule-based systems or the human supervision of machine learning algorithms.

As a result, an adjudicator seeking information about a potential *Ricci* situation is likely to have both a baseline from which the system deviated and information about the changes made to bring about that deviation. The availability of both that baseline and the changes made to the system are what has led to so much conversation about whether it will be legal to make changes in the name of resolving racial disparities in algorithmic decision-making. Unlike in human-centric systems where such tuning might happen implicitly, such as through the use of subjective "plus" factors,[190] race conscious but facially neutral modifications to bring about racial balancing of outcomes will be explicit,[191] observable, and even quantifiable (by comparing to the earlier baseline).

The law regarding the use of facially neutral classifications to achieve racially related ends is not as crystal clear as that on the use of race classifications themselves, but it is well-developed. Because racial purposes are themselves generally prohibited, many cases regarding the relationship between facially neutral ends and racially related ends are about using one to uncover the other. That is the entire basis of the *McDonnell Douglas* framework and its search for pretext in facially neutral employment practices,[192] and it was much of

189. Fred Miller et al., *Iterative Development Life Cycle (IDLC): A Management Process for Large Scale Intelligent System Development*, *in* THIRD INT'L CONF. ON TOOLS FOR ARTIFICIAL INTELLIGENCE - TAI 91 521 (1991).

190. Grutter v. Bollinger, 539 U.S. 306, 383-86 (2003) (Rehnquist, C.J., dissenting) (describing how the "plus factor" approach to race has been used by admissions officials to produce constitutionally prohibited racially proportionate admissions decisions).

191. Kleinberg et al., *supra* note 14, at 119-20.

192. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 804 (1973).

what was at issue in *Ricci* itself, in which the City's reason for rejecting the racially disparate test results was a major subject of the litigation.[193] The same is true in the constitutional context, albeit with a less-developed framework. In *Gomillion v. Lightfoot*,[194] the Court rejected a facially race-neutral city boundary change because the oddly shaped result correlated so closely with a race-based motivation.[195] The process of repeated changes to decision-making algorithms to produce racially optimized outcomes would reveal a great deal of information about the racially motivated objectives of the programmers. If the intent of (facially neutral) modifications were to confer a benefit upon a particular racial group, it would quickly become subject to the more searching inquiry the Court described in *Village of Arlington Heights v. Metropolitan Housing Development Corporation*,[196] and, if the racially motivated purpose were revealed, it would be subject to strict scrutiny under *Adarand Constructors*,[197] regardless of the way in which the benefits flowed.

Thus, it would seem at first blush that the legal prohibitions on discrimination combined with the practical realities of algorithmic decision-making would present considerable difficulties for the practice of compensating for racially disparate outcomes by algorithmic decision-makers. But that ignores the realities of how software, or really any system, is modified.

2. Disaggregating Algorithmic Affirmative Action

Key to understanding how humans can and cannot respond to racially disparate outcomes is recognition that such responses involve two separate decisions: the negative decision to reject the previous method and the affirmative decision to adopt a new one. Those two decisions, although perhaps sharing a common motivation, are very different in both their form and in their impact and consequently should be (and are) treated differently under discrimination law.

The decision to reject a particular decision-making process should receive considerable deference under discrimination law and should rarely lead to liability. Such decisions are, as an initial matter, facially neutral, since they affect everyone subject to the decision-making

193. *See* Primus, *supra* note 102, at 1361.

194. 364 U.S. 339 (1960).

195. *Id*. at 341 ("If these allegations upon a trial remained uncontradicted or unqualified, the conclusion would be irresistible, tantamount for all practical purposes to a mathematical demonstration, that the legislation is solely concerned with segregating white and colored voters by fencing Negro citizens out of town so as to deprive them of their pre-existing municipal vote.").

196. Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 266-70 & n. 21 (1977). *See* Shaw v. Reno, 509 U.S. 630, 644 (1993) (applying *Arlington Heights* to a facially neutral voting restriction like the one in *Gomillion*, 364 U.S. 339).

197. Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 226 (1995).

process.[198] A decision to reject a process wholesale does not itself classify on the basis of prohibited characteristics.

Perhaps more importantly, the decision to reject a particular decision-making process has a limited cognizable discriminatory impact, since plaintiffs are unlikely to have a right to any particular decision-making process. That was not the case in *Ricci*, which pertained to New Haven's rejection of the results of its previous promotion process. The Court found the firefighters to have a cognizable injury[199] in the form of their substantial reliance on the testing procedure, which had both involved considerable study and had been codified in their union's collective bargaining agreement.[200] But in *Ricci* not only was the promotions process particularly well-codified, it had proceeded essentially to the point of appointment before the City had rejected the test results. Even that would not have been enough for Justice Ginsburg, who argued in dissent that the plaintiffs could claim neither a vested right to a promotion nor that others had been promoted ahead of them, since no promotions had taken place.[201] The firefighters in *Ricci* had a tenuous claim to cognizable injury based on the City's rejection of the promotion process for the current round of promotions. It is inconceivable that, absent extraordinary circumstances like a collective bargaining agreement requiring it, the

198. Ricci v. DeStefano, 557 U.S. 557, 619-20 (2009) (Ginsburg, J., dissenting).

199. The City agued below that the firefighters had no standing to bring a constitutional equal protection claim, but they do not appear to have argued the plaintiffs lacked injury or standing under Title VII. *See* Ricci v. DeStefano, 554 F. Supp. 2d 142, 160-61 (D. Conn. 2006).

200. *Ricci*, 557 U.S. at 593 ("The injury arises in part from the high, and justified, expectations of the candidates who had participated in the testing process on the terms the City had established for the promotional process. Many of the candidates had studied for months, at considerable personal and financial expense, and thus the injury caused by the City's reliance on raw racial statistics at the end of the process was all the more severe.").

201. *Id*. at 608 (Ginsburg, J., dissenting). The salience of the firefighters claim led Richard Primus to argue for a "visible victims" reading of *Ricci*—that the Court was moved to find a conflict between the disparate impact and disparate treatment theories under Title VII because of the presence of visible victims of the discrimination, and in a case with less visible victims, the Court might see the relationship between the provisions differently. Primus, *supra* note 102, at 1369-70. Professor Primus drew the connection to Justice Kennedy's opinion in *Parents Involved* and its emphasis on race-neutral means, but I think that confuses means with injury. That is, if a plaintiff can establish that a race-neutral practice led to their firing (or disenrollment from a school), they are certain to satisfy any standing or injury requirement, although the standard of review might be lower than for a racial classification. Professor Primus later moved away from an emphasis on victims (and the potential interpretation as relating to standing or injury) to explain that visibility is more about the relative rhetorical attractiveness of the plaintiff's narrative regarding the putative discrimination than it is about the circumstances or characteristics of the plaintiff himself—that is, away from "victims" and toward "visibility" generally. *See* Richard Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact after* Ricci *and* Inclusive Communities, *in* TITLE VII OF THE CIVIL RIGHTS ACT AFTER 50 YEARS: PROCEEDINGS OF THE NEW YORK UNIVERSITY 67TH ANNUAL CONFERENCE ON LABOR 296 n.4 (2015) ("Visible victims are important because innocent and identifiable victims lend themselves to catchy narratives of injustice that raise the visibility of the practices that victimize them. But the importance of victims is in this way derivative—as a step toward the thing that ultimately matters, which is visibility.").

firefighters would have had standing to object to the City's refusal to use the same test in the *next* round of promotions.[202]

Law, including equal protection law, views the decision—even a discrete and demonstrable decision—*not* to do something differently than the decision to do it. In *Palmer v. Thompson*, decided the same year as *Griggs*, the Court upheld a city's decision to close its municipal pool in the face of a desegregation order (a manifestly race-based decision) by focusing not on the city's intent but rather on the lack of any affirmative duty for the city to operate the pool in the first place.[203]

It's not clear exactly what standard, if any, a court would apply to a decision to reject a process because it produced racially imbalanced results. Justice Kennedy's controlling view in *Parents Involved* would have held such a facially race-neutral policy (such as the choice to not apply a particular procedure to any applications) could be *justified* by race-conscious considerations,[204] which *a fortiori* would make it permissible. Even if the Court does not carry forward Justice Kennedy's then-controlling view, it is clear that a rejection of past practice (even a racially motivated rejection) is likely to receive more permissive review than the distinct decision to adopt a particular procedure with the intent to produce a particular set of racially balanced outcomes.

---

202. *Cf.* Kim, *supra* note 14, at 199. Professor Kim posits that future employees would have no right to the continued existence of the previous hiring policy. That is true, but that does not answer the question of what comes next. Professor Kim explains, "After *Ricci*, then, employers are permitted to audit automated decision processes and change them prospectively in order to eliminate identified biases." *Id.* at 200. The problem is in her use of the word "change," which conflates the rejection and the adoption of the new policy. If the new policy is adopted *in order to produce a specific racial outcome*, then those prospective future employees will have standing to challenge the new policy for its disparate treatment when they are rejected for employment under it. *See infra* pp. 557-58.

203. Palmer v. Thompson, 403 U.S. 217, 220-21 (1971). *Cf.* Griffin v. Cty. Sch. Bd. of Prince Edward Cty., 377 U.S. 218 (1964). In *Griffin*, the Court had held unconstitutional a school district's decision to close public schools in the face of a desegregation order. But in *Griffin*, the "private" schools that opened after the closing of the public schools served only whites and received considerable support from the state and county, a point the Court found in *Palmer* distinguished it from that case. *See Palmer*, 403 U.S. at 221-22. After *Palmer*, it's apparent that it was the state's involvement in the operation of the private schools in *Griffin*, not the decision to close the public ones, that constituted the equal protection violation. *See* James M. DeLise, *Racial Impermissibility Under the Equal Protection Clause from* Strauder v. West Virginia *to* Ricci v. DeStefano, 17 RUTGERS RACE & L. REV. 179, 188 (2016).

204. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 788-89 (2007) (Kennedy, J., concurring in part and dissenting in part) ("If school authorities are concerned that the student-body compositions of certain schools interfere with the objective of offering an equal educational opportunity to all of their students, they are free to devise race-conscious measures to address the problem in a general way and without treating each student in different fashion solely on the basis of a systematic, individual typing by race."). Justice Breyer's dissent on this point would have allowed even race-based classifications if intended to reduce racial disparities. *Id.* at 823 (Breyer, J., dissenting) ("A longstanding and unbroken line of legal authority tells us that the Equal Protection Clause permits local school boards to use race-conscious criteria to achieve positive race-related goals, even when the Constitution does not compel it.").

The standard applicable to that second decision—to adopt a new practice—would be treated like any other decision under both statutory or constitutional discrimination law. If it included a race classification, it would receive searching scrutiny and virtual per se illegality if it classifies on the basis of race under Title VII. But even a facially neutral replacement could be considered disparate treatment depending on the intent underlying it (under *McDonnell Douglas*), including the intent to avoid racially disparate outcomes—that much is clear from *Ricci* itself. Consequently, potential defendants seeking to replace practices they previously rejected for their disparate impacts must do so *without* regard to the likelihood that the replacement practice will produce more balanced outcomes. Rather, they must start from scratch (albeit with the benefit of experience) and attempt to put in place a process that best serves the productive interest the practice serves. The employer might hope that the practice will produce racially balanced outcomes (as I am guessing most employers already hope whenever they adopt a new practice), but hope is different from intent. If the practice is adopted with the intent of producing particular racial balancing, it is illegal just like any other practice would be.

Even if the second decision (to adopt a new practice) will still receive normal levels of scrutiny, the permissive review applicable to decisions to stop using old algorithms allows comparatively more latitude for firms and government agencies to consider race in the separate decision to reject an old algorithm than does considering the two decisions as one.

Two thoughts on where all this leads:

First, my proposal might sound like asking employers to repeatedly, but blindly, attempt to achieve racially balanced outcomes, and it is subject to the objection that such an approach is unlikely the most efficient way to produce those outcomes. But that is no different than the system that employers are living under right now with regard to practices applied in a less systematic manner by human decision-makers. Employers observe and correct their hiring practices constantly, and they are expected to do so without discriminatory intent, although they are encouraged by the existence of disparate impact liability to produce racially balanced ones. What I am suggesting is really no change at all but for the fact that with algorithmic decision-makers, the practice must be explicitly included in the decision-making process, as described above.

Second, this approach of race-blind recalibration raises the question of how many times an employer can reject practices before the facially neutral rejection effectively becomes an intentional, disparate treatment claim—the point at which a series of race-neutral rejections combine into a meta-practice of trying different processes at random until the employer finds one that produces the employer's preferred

racial balancing, at which time the employer elects settle on that one. There is no answer to that question, although both *McDonnell Douglas* and the constitutional equal protection cases like *Gomillion* and *Arlington Heights* suggest approaches to providing one.[205] My point is only that disaggregating the rejection and subsequent adoption decisions both provides the best understanding of attempts to respond to disparate outcomes produced by algorithmic decision-makers and allows some space (in the decision to reject but not to adopt a replacement) for potential defendants to operate. Exactly how much space they have remains to be seen.

CONCLUSION

As businesses and governments delegate increasing—and increasingly sophisticated—decisions to computational decision-makers, disparate outcomes produced by computational decision-makers will garner even more attention in the media and among scholars. The temptation will be to respond to specific instances of unfairness with correctives designed to eliminate it. But that is hardly a new phenomenon. As the scale and automation of business and government have grown, attention to disparate outcomes frequently serves to prompt reform efforts. But those reform efforts do not generally attempt to instill "fairness" or anything like it in business or government. Law is quick to abandon as unworkable anything like a fairness test for business practices. That is so both because fairness has to be a side constraint on other activity but also because fairness is a concept that should be applied in every context but cannot be applied without context. The law is designed with that concern in mind, which is why discrimination law exists as a set of specific, largely negative mandates that accommodate productive activity by evaluating practices based not on the disparate outcomes they produce but on the business efficiency that is their goal. Readily apparent outcomes matter a lot to reporters and media outlets whose business is to catch readers' attention, but they mean comparatively less in the law of discrimination for reasons having little to do with why discrimination is wrong.

Although the demands of transparency might seem less demanding than those of fairness, it is no more attainable than fairness itself, or at least it has not been so for as long as humans have been making decisions. Compared to other ways in which machines operate differently from humans, computational decision-makers are not really any less transparent than the humans that have previously applied algorithms—business policies and practices—to people and their problems. The problem of systematized rather than individual discrimination is hardly a novel one to discrimination law, from the

205. *See supra* text accompanying notes 192-196.

educational requirements in *Griggs* to present day. Discrimination law has developed a variety of tools for dealing with institutionalized discrimination, and there is little reason to believe those tools cannot be applied to policies applied by computers rather than their far-more-inscrutable human counterparts. Legal restrictions on human discretion like discrimination law have always been about liability for a thought process we can only observe at remove, with the explanation of accountability less accessible and the visibility of transparency even less so. Employers cannot open up the minds of their managers—they cannot provide true transparency—and so it is no surprise that the legal system essentially ignores transparency as a value in discrimination law. Accountability, as motivated by the potential for liability, has always been what the law demands. Providing that incentive for accountability rather than requiring unattainable transparency, should be the goal in regulating computational decision-makers.

Even in the absence of potential liability, businesses will rightly want to modify practices that have a discriminatory impact on their customers and employees, but how they do so will necessarily change as more decisions are delegated to computational decision-makers. In an era of human decision-making, a hint or lament over lunch might be enough to push a decision-maker to adopt an approach that reduces discriminatory impact. In an era of computational decision-making, the change will have to be explicit. That will require businesses to confront exactly what it is they are trying to do. As with discrimination generally, though, the right question for a lender (for example) is not "What kind of racial balance do we want to see in our loans?" but rather "How do we make the best loans possible?" If lenders suspect race, which itself is irrelevant to the quality of a loan, is playing a role in lenders' decision-making, they are justified in rejecting that process and developing one that better serves their business objectives. That doesn't change if the lender's objection to employing a racially discriminatory process is moral rather than economic; rejection of institutionalized discrimination, intentional or not, is not only morally right but legally permissible. But what comes next is a different matter. Rejecting institutional race discrimination is uncontroversial; explicitly adopting race-conscious, or even race-based, decision-making is not. Fortunately, it's not necessary to do so in order to serve either the business or moral interest against race, sex, religious, age, or other forms of invidious discrimination.

We may very well lose something as we delegate more and more decision-making to computational decision-makers. Human decision-making, for all its faults, necessarily humanizes decision-making for both the objects of decisions and the decision-makers themselves. It is harder to turn a blind eye to disparate outcomes when we are the ones day after day refusing job applicants or denying loan applications.

Delegation to computational decision-makers may require us to confront many of the things that have been implicit and unspoken in our own human decision-making. In some cases that's going to be good, and in some cases it's going to be bad. Given current social disparities, the law prevents the kind of disparate treatment necessary to produce racially balanced outcomes. We cannot (and should not) program computational decision-makers to take account of characteristics such as race in order produce racially balanced outcomes. But so long as the power to make decisions is delegated, rather than ceded, to computational decision-makers, they make decisions in our name and for our benefit. We do not have to resign ourselves to accepting algorithms that propagate today's disparities, and that is no different whether those decision are being made by computers or humans.