Antidiscriminatory Algorithms

Stephanie Bornstein

University of Florida Levin College of Law
Legal Studies Research Paper Series Paper No. 19-6

ANTIDISCRIMINATORY ALGORITHMS

Stephanie Bornstein

ANTIDISCRIMINATORY ALGORITHMS

Stephanie Bornstein[*]

Can algorithms be used to advance equality goals in the workplace? A handful of legal scholars have raised concerns that the use of big data at work may lead to protected class discrimination that could fall outside the reach of current antidiscrimination law. Existing scholarship suggests that, because algorithms are "facially neutral," they pose no problem of unequal treatment. As a result, algorithmic discrimination cannot be challenged using a disparate treatment theory of liability under Title VII of the Civil Rights Act of 1964 (Title VII). Instead, it presents a problem of unequal outcomes, subject to challenge using Title VII's disparate impact framework only. Yet under current doctrine, scholars suggest, any disparate impact that results from an employer's use of algorithmic decision-making could be excused as a justifiable business practice. Given this catch-22, scholars propose either regulating the algorithms or reinterpreting the law.

This Article seeks to challenge current thinking on algorithmic discrimination. Both the "improve the algorithms" and the "improve the law" approaches focus solely on a clash between the anticlassification (formal equality) and antisubordination (substantive equality) goals of Title VII. But Title VII also serves an important antistereotyping goal: the principle that people should be treated not just equally across protected class groups but also individually, free from stereotypes associated with even one's own group. This Article is the first to propose that some algorithmic discrimination may be challenged as disparate treatment using Title VII's stereotype theory of liability. An antistereotyping approach offers guidance for improving hiring algorithms and the uses to which they are put, to ensure that algorithms are applied to counteract rather than reproduce bias in the workplace. Moreover, framing algorithmic discrimination as a problem of disparate treatment is essential for similar challenges outside of the employment context—for example, challenges to governmental use of algorithms in the criminal justice context raised under the Equal Protection Clause, which does not recognize disparate impact claims.

The current focus on ensuring that algorithms do not lead to new discrimination at work obscures that the technology was intended to do more: to improve upon human decision-making by suppressing biases to make the most efficient and least discriminatory decisions. Applying the existing doctrine of Title VII more robustly and incorporating a focus on its antistereotyping goal may help deliver on the promise of moving beyond mere nondiscrimination and toward actively antidiscriminatory algorithms.

---

* Associate Professor of Law, University of Florida Levin College of Law. For their helpful comments and questions on presentations or drafts of this Article, my sincere thanks to Scott Bauries, Jason Bent, Joseph Fishkin, Pauline Kim, Marcia McCormick, Katheryn Russell-Brown, Andrew Selbst, and Charles Sullivan. My thanks as well to the participants in the 2017 Colloquium on Scholarship in Labor & Employment Law, the 2018 Law & Society Association Conference Program on the Future of Workforce Management, and the 2018 SEALS Conference New Scholars Workshop. Thanks, too, to Dale Dowden and Kaley Jaslow for their excellent research assistance.

INTRODUCTION

In 2014, prior to several years during which they would hire thousands of workers, the leadership of online retail giant Amazon asked their machine-learning experts to develop an automated tool to help with hiring decisions.[1] By 2015, their programmers recognized that the tool was plagued by gender bias; by early 2017, they abandoned the effort.[2] The goal was to create an algorithm using artificial intelligence (AI) that could rank job candidates to automate hiring.[3] For the algorithm to learn what to value, the programmers trained it to find patterns in resumes submitted for technical jobs in the prior ten years, most of which—due to the demographics of who holds those jobs—came from male applicants.[4] As a result, the model "taught itself" to prefer male candidates, "penaliz[ing] resumes that included the word 'women's,' as in 'women's chess club captain,'" and "downgrad[ing] graduates of . . . all-women's colleges."[5] The programmers corrected for that particular problem, but ultimately shelved the project, concerned that there "was no guarantee that the machines would not devise other ways of sorting candidates that could prove discriminatory."[6]

Amazon is by no means unique in seeking technological solutions to its personnel needs. In 2018, LinkedIn conducted a survey of 9,000 hiring managers and recruiting professionals about current and future trends in workplace hiring.[7] Half of respondents identified that data analytics was "very" or "extremely important" to the future of hiring with nearly one-fifth reporting that they had "mostly" or "completely adopted" its use in their own practices to date.[8] Likewise, 35% said that AI would be "very" or "extremely important" to recruiting in the future, and nearly one in twelve had already adopted its use.[9] The survey confirms anecdotal evidence documenting the rise of data and AI in the workplace over the past decade.[10] It

1. *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, REUTERS (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

2. *Id.*

3. *Id.*

4. *Id.*

5. *Id.*

6. *Id.*

7. LINKEDIN TALENT SOLUTIONS, GLOBAL RECRUITING TRENDS 2018, https://business.linkedin.com/content/dam/me/business/en-us/talent-solutions/resources/pdfs/linkedin-global-recruiting-trends-2018-en-us.pdf (last visited Sept. 25, 2018).

8. *Id.* at 4.

9. *Id.*

10. *See, e.g.*, Ken Gaebler, *The Future of Hiring: Human Resources, Without the Humans*, ATLANTIC (Feb. 3, 2012), https://www.theatlantic.com/business/archive/2012/02/the-future-of-hiring-human-resources-without-the-humans/252518; Aki Ito, *Hiring in the Age of Big Data*, BLOOMBERG

also comes as no surprise: as with countless other facets of life (advertising, banking, voting, tax auditing, medicine, and criminal justice, to name a few), algorithms and data analytics are aiding or replacing decisions once made entirely by humans.[11]

The rise of big data at work has sparked concerns about privacy and procedural fairness[12] and, more recently, discrimination.[13] Employers are now using algorithms to make hiring and other workplace decisions quickly and automatically. If the underlying data on which an algorithm relies is

(Oct. 24, 2013), https://www.bloomberg.com/news/articles/2013-10-24/new-way-to-assess-job-applicants-online-games-and-quizzes; Nathan R. Kuncel, Deniz S. Ones & David M. Klieger, *In Hiring, Algorithms Beat Instinct*, HARV. BUS. REV. (May 2014), https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct; Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, N.Y. TIMES (June 28, 2015), https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html.

11. For a definition of "algorithm," see Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 674 n.10 (2016) (citing SOLON BAROCAS ET AL., DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER (2014), http://www.datacivilrights.org/pubs/2014-1030/technology.pdf) ("An 'algorithm' is a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions. Algorithms play a role in both automating the discovery of useful patterns in datasets and automating decision making that relies on these discoveries."). For definitions of "artificial intelligence" (AI) and "machine learning," see Bernard Marr, *What Is the Difference Between Artificial Intelligence and Machine Learning?*, FORBES (Dec. 6, 2016, 2:24 AM), https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/3/#157219c52bfc ("Artificial Intelligence is the broader concept of machines being able to carry out tasks in a way that we would consider 'smart'. . . . Machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves.").

12. A sizeable body of literature now addresses concerns about privacy and procedural fairness in the collection and use of employee data—topics that are beyond the scope of this Article. *See generally, e.g.*, Ifeoma Ajunwa, *Algorithms at Work: Productivity Monitoring Platforms and Wearable Technology as the New Data-Centric Research Agenda for Employment and Labor Law*, 63 ST. LOUIS U. L.J. (forthcoming 2019); Ifeoma Ajunwa, *Genetic Testing Meets Big Data: Tort and Contract Law Issues*, 75 OHIO ST. L.J. 1225 (2014); Ifeoma Ajunwa, Kate Crawford & Jason Schultz, *Limitless Worker Surveillance*, 105 CALIF. L. REV. 735 (2017); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18 (2017); Pauline T. Kim & Erika Hanson, *People Analytics and the Regulation of Information Under the Fair Credit Reporting Act*, 61 ST. LOUIS U. L.J. 17 (2016); Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393 (2014); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. (forthcoming 2018). *But see* Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, 66 STAN. L. REV. ONLINE 35 (2013).

13. *See generally, e.g.*, Ifeoma Ajunwa, *Age Discrimination by Platforms*, 40 BERKELEY J. EMP. & LAB. L. (forthcoming 2019); Barocas & Selbst, *supra* note 11; Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189 (2017) [hereinafter Kim, *Auditing Algorithms*]; Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017) [hereinafter Kim, *Data-Driven*]; Pauline Kim & Sharion Scott, *Discrimination in Online Employment Recruiting*, 63 ST. LOUIS U. L.J. (forthcoming 2019); Charles A. Sullivan, *Employing AI* (Feb. 18, 2018) (Seton Hall Public Law Research Paper), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125738; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017).

itself biased, incomplete, or discriminatory, the decisions it makes have the potential to reproduce inequality on a massive scale.[14]

Yet while data and AI were the third- and fourth-ranked top trends identified by LinkedIn's survey respondents, they were not the first.[15] Most cited as "very" or "extremely important" to the future of workplace hiring was a commitment to diversity: 78% of respondents identified it as essential, and 53% already incorporated it as a focus in their recruiting efforts.[16] In fact, the trend of hiring by algorithm grew out of a cottage industry of tech start-ups seeking to help diversify Silicon Valley.[17] Algorithmic decision-making offers unprecedented potential to reduce the stereotypes and implicit biases that often infect human decisions.[18] If both an intention to use data analytics and a commitment to diversity in hiring are of high importance to the same majority of employers, surely the two objectives can be aligned.[19]

A small, but robust, body of legal scholarship has begun to raise concerns about the potential for algorithmic decision-making to result in protected class discrimination in employment.[20] Title VII of the Civil Rights Act of 1964 (Title VII) prohibits discrimination in hiring, firing, compensation, and other "terms, conditions, [and] privileges" of employment on the basis of protected classes, including race and sex.[21] The few legal scholars addressing algorithmic discrimination in the workplace agree that, while using algorithms to make employment decisions offers the promise of reducing the biases inherent in human subjective decision-making, it also poses a more significant, and dangerous, risk of reproducing existing ine-

14. *See* Barocas & Selbst, *supra* note 11, at 677–93; Kim, *Data-Driven*, *supra* note 13, at 883–92.

15. LINKEDIN TALENT SOLUTIONS, *supra* note 7, at 4.

16. *Id.*

17. *See infra* Subpart I.A.

18. *See infra* Subpart II.C.

19. These trends may be most important to the segment of hiring professionals who responded to LinkedIn's survey and have more limited appeal to others. But it is precisely that segment—recruiters and hiring managers likely to both use data analytics/AI and value diversity in hiring—to whom the issue of algorithmic discrimination most applies.

20. *See* Ajunwa, *supra* note 13; Barocas & Selbst, *supra* note 11; Kim, *Auditing Algorithms*, *supra* note 13; Kim, *Data-Driven*, *supra* note 13; Kim & Scott, *supra* note 13; Kroll et al., *supra* note 13; Sullivan, *supra* note 13; *see also* Matthew T. Bodie et al., *The Law and Policy of People Analytics*, 88 U. COLO. L. REV. 961 (2017) (surveying the field); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164 (2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2950018 (proposing a thought piece to illustrate limitations of existing law); Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555 (2016) (same).

21. 42 U.S.C. § 2000e-2 (2012) (prohibiting discrimination because of race, color, national origin, sex, or religion).

quality.[22] Worse still, current scholarship suggests, the apparent neutrality of algorithms and the "black box" nature of machine learning make this hiring trend a new way of doing business that could be unreachable by existing antidiscrimination law.[23]

While they agree on the problem, scholars have proposed two different, though complementary, solutions. One response focuses on regulating to improve the algorithms themselves, based on computer science techniques guided by Title VII.[24] By requiring accountability for the ways in which the underlying data may be flawed or the algorithmic process may incorporate bias, this approach seeks to help reduce and prevent algorithmic discrimination *ex ante*.[25] The second response focuses on improving antidiscrimination law's ability to reach algorithmic discrimination *ex post* by reinterpreting Title VII doctrine as applied to the context of algorithmic discrimination.[26]

Each response is a well-researched and thoughtful way to approach a difficult problem, and each stands to make an impact on the future of big data at work. Yet each also carries with it a limitation: a lack of current enforceability. An *ex ante* focus on improving the algorithms requires a governance structure that makes employers mitigate a problem for which, according to the solution's proponents, employers likely cannot be held liable under Title VII.[27] An *ex post* focus on improving the law requires unearthing new possibilities in Title VII doctrine that may run counter to current court precedent.[28]

This Article challenges the assumptions underlying existing scholarship on algorithmic discrimination and offers a third possibility in response to the problem. While the technology of algorithms and AI may be new, the legal issues it raises are not. Algorithmic decision-making is just the latest personnel management tool—not so different from past innovations like the rise of personality testing in the 1980s and the use of executive recruiters and staffing agencies in the 1990s.[29] In the fifty-five years since Title VII

22. *See infra* Subpart I.B; *see also* Barocas & Selbst, *supra* note 11, at 677–93; Kim, *Data-Driven*, *supra* note 13, at 883–92.

23. *See infra* Subpart I.B; *see also* Barocas & Selbst, *supra* note 11, at 694–713; Kim, *Data-Driven*, *supra* note 13, at 901–04.

24. *See infra* Section I.B.1; *see also* Barocas & Selbst, *supra* note 11, at 714–28; Kroll et al., *supra* note 13, at 678–95.

25. *See infra* Section I.B.1; *see also* Barocas & Selbst, *supra* note 11, at 714–28; Kroll et al., *supra* note 13, at 678–95.

26. *See infra* Section I.B.2; *see also* Kim, *Data-Driven*, *supra* note 13, at 909–36.

27. *See infra* Section I.B.1.

28. *See infra* Section I.B.2; *infra* notes 124–25 (discussing Connecticut v. Teal, 457 U.S. 440 (1982) and the issue of a "bottom line" defense to Title VII).

29. *See infra* Subpart I.A.

was enacted to prohibit employment discrimination, its doctrine has adapted to reach increasingly more subtle and complex forms of discrimination. In particular, courts have recognized that employment decisions that incorporate stereotypes associated with protected classes may be actionable, which raises new legal questions for the use of predictive algorithms.[30]

The Article identifies a gap in existing scholarship regarding the theoretical foundations underlying antidiscrimination law that, when filled, suggests a new path forward. Both scholarly camps identify that Title VII serves two main goals, each providing a theory of liability, yet neither effectively redressing the harm of algorithmic discrimination.[31] On the one hand, Title VII's anticlassification goal requires formal equality and gives rise to the disparate treatment framework of liability, under which discrimination occurs when an applicant or employee is intentionally treated differently than others based on protected class.[32] On the other hand, Title VII's antisubordination goal seeks substantive equality and gives rise to the disparate impact framework of liability, under which discrimination occurs when applicants or employees are treated the same—in a "facially neutral" manner—but the resulting outcomes have disproportionately negative results on certain protected classes.[33] Current scholarship approaches algorithmic discrimination as primarily a problem of disparate impact because it views algorithmic decision-making as a facially neutral practice applied equally to all applicants or employees.[34] Yet affirmative defenses available to employers under current Title VII disparate impact doctrine mean that, even if an employer's use of algorithmic decision-making results in a disparate impact by protected class, the impact could be excused as "job related" and consistent with "business necessity," making the employer likely to prevail.[35]

But anticlassification and antisubordination are not the only theories supporting Title VII. Existing scholarship overlooks an important third principle of antidiscrimination law: its antistereotyping goal. Under the antistereotyping approach, the law requires not just equal treatment or equal outcomes between protected groups but also individualized treatment even *within* one protected group. Individuals may not be judged for employment purposes based on stereotypes associated with a protected class. The stereotype framework of liability under Title VII is a particular species of claim arising under the disparate treatment framework. While disparate treatment

30. *See infra* Part III.
31. *See infra* Subpart II.A.
32. *See infra* Subpart II.A.
33. *See infra* Subpart II.A.
34. *See infra* Subpart I.B.
35. *See infra* Subpart III.A.

typically requires intentional discrimination, stereotype theory allows it to reach intentional actions that incorporate or are infected by even unrecognized bias.[36]

Applying an antistereotyping lens to the issue of algorithmic decision-making calls into question the underlying "neutrality" of algorithms and the big data on which they rely. This Article is the first to propose that some algorithmic discrimination may be challenged as disparate treatment using Title VII's stereotype theory of liability. When an individual is judged negatively based on or by comparison to a body of group data, the individual may have been unfairly stereotyped. Viewed this way, predictive analytics that seek to judge and match individuals to a possibly biased model of a "good employee" appear to be a form of stereotyping at hyperspeed. The fact that a computer, instead of a human, does the stereotyping should not insulate from liability the employer who relies on the stereotyped results if the employer's intentional use of an algorithm discriminates. Indeed, if AI is meant to model human decision-making, but on an autonomous and massive scale, theories of liability that apply to human decision-making should likewise apply.[37]

An antistereotyping approach also offers new lessons for preventing algorithmic discrimination at work. Current legal scholarship has identified how to reduce the data problems and discriminatory effects of algorithmic decision-making, and an entire field of computer science is focused on the technical aspects of this endeavor.[38] Title VII's antistereotyping principle offers additional guidance—not just on how to de-bias algorithms themselves but on how to think about the *uses* for which algorithms are appropriate. Workplaces can use algorithms in a wide variety of ways, some of which exacerbate reliance on stereotypes and others of which help counteract the effects of bias. The risk of liability for stereotyping may help discourage the former and encourage the latter. This could help shift the focus from ensuring that algorithms do not result in new discrimination toward fulfilling the promise of their design: to suppress human biases and increase diversity in hiring. It could help move beyond merely *non*discriminatory to actively *anti*discriminatory algorithms.[39]

Framing algorithmic discrimination as a problem of stereotyping and unequal treatment is also essential for redressing similar concerns outside of the employment context and the protections of Title VII. One prominent

36. *See infra* Subpart II.B.

37. *See infra* Subparts II.B, III.B. *But see* Sullivan, *supra* note 13, at 8 (arguing that a computer making decisions using AI "isn't human, so it can't 'intend' to discriminate," as required for disparate treatment liability).

38. *See infra* Section I.B.1.

39. *See infra* Subparts II.B, II.C.

example is any challenge to the use of algorithms in the criminal justice context raised under the U.S. Constitution's Equal Protection Clause, which does not recognize liability under a disparate impact framework.[40] This means that, should individuals wish to challenge, for example, predictive policing or algorithmic risk assessment for sentencing or probation, they would have to be able to demonstrate that they experienced discriminatory treatment—not merely the discriminatory effects of a facially neutral practice. The antistereotyping principle applies equally to antidiscrimination law under Title VII and the Constitution; in fact, this principle originated in cases brought by individuals under the Equal Protection Clause to challenge state and federal laws that enforced gender role stereotypes.[41] To the extent that algorithmic discrimination constitutes unlawful stereotyping in the workplace, similar arguments may apply to algorithmic discrimination in other facets of life.

This Article proceeds in three Parts. Part I provides some context for the rise of big data at work and some background on current common uses of algorithms in the workplace. It then summarizes existing scholarship on the issue, including proposed solutions to remedy algorithmic discrimination at work through regulating to improve algorithms or re-envisioning the reach of Title VII. Part II turns to the theories underlying antidiscrimination law, first identifying the two main theories addressed in current scholarship, anticlassification and antisubordination, then introducing a third, the antistereotyping principle. It also applies an antistereotyping lens to offer perspective on how algorithms are used in the workplace, suggesting that some uses are better than others for advancing racial and gender equality at work. Part III examines whether algorithmic discrimination in employment can be redressed under existing law, first revisiting existing arguments on the limitations of current doctrine, then proposing that some types of algorithmic discrimination could constitute disparate treatment under a stereotype theory of liability. This Part concludes with possible implications for algorithmic discrimination outside of the employment context, including challenges to the use of algorithms in the criminal justice context brought under the Equal Protection Clause.

Ultimately, the Article suggests that there is more room for redressing algorithmic discrimination under existing law than others have identified and more guidance to be gained from incorporating an antistereotyping perspective into the current debate. A focus on the risks of technology-aided decision-making is important, but it need not overshadow the potential rewards: algorithms may incorporate structural biases, but they also

40. *See infra* Subpart III.C.

41. *See infra* Subpart II.B.

suppress human biases. If both big data and diversity are important to the employers of the future, an antistereotyping approach can help align the two so that data helps, rather than hampers, greater workplace equality.

I. THE CHALLENGE OF ALGORITHMS AT WORK

Over the past decade, data analytics has made its way into human resources practices, raising concerns about the potential for data-based employment discrimination. This Part begins by providing some context and background on the rise of and current uses for algorithms in workplace decision-making. It then summarizes current scholarship on the legal implications of algorithmic discrimination at work, including suggested responses to the problem to date.

A. The Rise of Algorithms in the Workplace

1. Algorithmic Decision-Making: The Latest Personnel Management Innovation

The current rise of data- and AI-based decision-making at work comes within a long context of employer practices seeking to make better selection decisions faster and more cheaply.[42] Since the passage of civil rights laws, including Title VII, employers have had to incorporate a principle of nondiscrimination into their processes. These and other legal and regulatory requirements on employers helped spur the growth of the human resources field and, with it, an ever-evolving series of tools designed to assist employers.[43] Over the past five decades, employers have adopted a variety of practices for workforce management in areas like hiring, evaluation, and promotion. Yet despite evolving practices, Title VII has been applied in each era and to each practice. And where workforce management innovations concealed or exacerbated continuing discrimination, Title VII was adapted to meet the challenge.[44]

In the early twentieth century, psychologists developed personality tests that employers could use to evaluate applicants for qualities they de-

42. *See, e.g.*, Bodie et al., *supra* note 20, at 964–68; Pernilla Bolander & Jörgen Sandberg, *How Employee Selection Decisions Are Made in Practice*, 34 ORG. STUD. 285, 285–87 (2013) (describing existing research since the 1970s that "has focused on developing and testing tools intended to improve selection and make it more efficient").

43. *See generally* FRANK DOBBIN, INVENTING EQUAL OPPORTUNITY (2009) (documenting the role of personnel professionals in putting antidiscrimination law into practice).

44. *See, e.g.*, Griggs v. Duke Power Co., 401 U.S. 424, 436 (1971); *Fact Sheet on Employment Tests and Selection Procedures*, EEOC [hereinafter *Fact Sheet*], https://www.eeoc.gov/policy/docs/fact employment_procedures.html (last modified Sept. 23, 2010).

sired in employees.[45] The field of personality testing grew over time, in part to help assign soldiers to various duties during World Wars I and II.[46] In the late 1980s, employers' use of personality testing to evaluate job applicants became widespread, with dozens of different types of tests in use.[47] With the growth of this innovation came fears about its risks: commentators raised concerns about a variety of legal issues, including applicant privacy, due process, and discrimination.[48] Yet antidiscrimination law adapted. In caselaw, federal courts interpreted Title VII's statutory language on "ability tests" to include coverage of potential discrimination in employer personality testing.[49] The U.S. Equal Employment Opportunity Commission (EEOC), the federal agency responsible for enforcing Title VII, also weighed in, including its view that personality tests are among the "employment tests and selection procedures" Title VII covers.[50]

Likewise, the concept of executive search originated in the mid-twentieth century, but in the 1990s, employers began routinely outsourcing the process of hiring to recruiters and staffing firms.[51] Again, antidiscrimination law rose to the challenge. In caselaw, federal courts interpreted statutory language establishing that Title VII covers "employers" and "employment agencies" to include relationships between staffing agencies or recruiting firms and their employer clients.[52] In 2006, the EEOC addressed the hiring trend in a section entitled "Recruitment" in its updated Compliance Manual, citing earlier guidance and explaining the grounds upon which recruiters and staffing firms could be held liable for their own—or their employer-clients'—hiring discrimination.[53]

---

45. *See* Kimberli R. Black, *Personality Screening in Employment*, 32 AM. BUS. L.J. 69, 71 (1994); Bodie et al., *supra* note 20, at 964–68.

46. *See* Black, *supra* note 45, at 71–72; Bodie et al., *supra* note 20, at 964–68.

47. *See* Black, *supra* note 45, at 76–80.

48. *See id.* at 90–120; Susan J. Stabile, *The Use of Personality Tests as a Hiring Tool: Is the Benefit Worth the Cost?*, 4 U. PA. J. LAB. & EMP. L. 279, 299–308 (2002).

49. *See* Sujata S. Menjoge, *Testing the Limits of Anti-Discrimination Law: How Employers' Use of Pre-Employment Psychological and Personality Tests Can Circumvent Title VII and the ADA*, 82 N.C. L. REV. 326, 335–36 (2003) (citing cases interpreting 42 U.S.C. § 2000e-2(h), including Colbert v. H-K Corp., No. 11599, 1971 WL 215, at *1 (N.D. Ga. July 28, 1971)).

50. *See Fact Sheet*, supra note 44.

51. *See, e.g.*, Rich Williams, *The Evolution of Executive Search*, https://charlesaris.com/evolution-executive-search/ (last visited Oct. 31, 2018) (noting that "'[h]eadhunter' [became] a household word" in the "[l]ate 1980s/early 1990s," due to "several high-profile CEO searches for IBM, Coca-Cola and The Walt Disney Company").

52. *See, e.g.*, Reynolds v. CSX Transp., Inc., 115 F.3d 860, 869 n.12 (11th Cir. 1997), *rev'd on other grounds*, 524 U.S. 947 (1998); *see also* 42 U.S.C. §§ 2000e-2(b) (2012) (stating that Title VII applies to "employment agencies"), 2000e(c) (defining "employment agency").

53. *See* U.S. EQUAL EMP. OPPORTUNITY COMM'N, COMPLIANCE MANUAL § 15, IV.A (Apr. 9, 2006) (citing U.S. EQUAL EMP. OPPORTUNITY COMM'N, ENFORCEMENT GUIDANCE: APPLICATION OF EEO LAWS TO CONTINGENT WORKERS PLACED BY TEMPORARY EMPLOYMENT AGENCIES AND OTHER STAFFING FIRMS (Dec. 3, 1997)), https://www.eeoc.gov/policy/docs/race-color.html; Michael Harris,

While it poses some unique challenges, the rise of data-based analytics and the use of AI in hiring is just the most recent innovation, making it part of this same evolution. Indeed, even some types of algorithmic hiring tools model earlier personality testing by using AI to measure cues from applicants that indicate desirable personality traits.[54] As discussed in Subpart I.B, current scholarship on algorithmic discrimination expresses concern that Title VII may be unable to reach the discriminatory harms caused by this innovation. Yet, as this Article argues, if the law of Title VII has been able to adapt to reach earlier trends in employer hiring, there is reason to believe it can adapt to reach current innovations, too.

2. Current Uses of Algorithms at Work

To date, employers have used algorithms and AI in a wide variety of ways, and the legal implications depend on their use. Generally, when employers use algorithms, the goal is to gather and apply data to make decisions in a faster, more efficient, and more objective manner. Employers may use algorithms to track productivity, assist with performance evaluations, evaluate compensation, determine necessary training, manage workplace benefits, and more.[55] Yet, the most common use of algorithms in the workplace, and the most directly relevant to the current debate over algorithmic discrimination, is in hiring. Over the past five years, dozens of technology companies have been launched to offer data- or AI-based options for recruiting job applicants and making hiring decisions, now often referred to as "talent acquisition."[56] While an algorithm is simply a computerized formula that can be designed to do whatever an employer asks of it, the leading data-based recruitment services available to employers tend to serve two main roles: either data mining and predictive matching or skills-based testing and recruitment tracking.[57]

EEOC Is Watching You: Recruitment Discrimination Comes to the Forefront, ERE RECRUITING INTELLIGENCE (May 30, 2006), https://www.ere.net/eeoc-is-watching-you-recruitment-discrimination-comes-to-the-forefront/.

54. Hilke Schellmann & Jason Bellini, *Artificial Intelligence: The Robots Are Now Hiring*, WALL ST. J. (Sept. 20, 2018), https://www.wsj.com/articles/artificial-intelligence-the-robots-are-now-hiring-moving-upstream-1537435820 (describing how the companies DeepSense and HireVue are using new tools to detect personality traits).

55. *See, e.g.*, Josh Bersin, *9 HR Tech Trends for 2017*, SOC'Y FOR HUMAN RES. MGMT. (Jan. 25, 2017), https://www.shrm.org/hr-today/news/hr-magazine/0217/pages/9-hr-tech-trends-for-2017.aspx.

56. *See* Jenny Roper, *What Do We Mean When We Walk About Talent?*, HR MAG. (June 15, 2015), http://www.hrmagazine.co.uk/article-details/what-do-we-mean-when-we-talk-about-talent.

57. Like Kim, I set aside the issue of third-party liability and focus on employer responsibility for using a hiring process to make its decisions. *See* Kim, *Data-Driven*, *supra* note 13, at 916 ("[T]his exploration focuses on employer liability, leaving aside the question whether vendors who create these models and sell or license them to employers should bear any legal responsibility. . . . Regardless of whether vendors are directly liable, employers who face potential legal responsibility will have an in-

a. Mining-and-Matching Uses

The most prominent use of algorithms and AI in hiring is to mine available data for potential applicants and predict who will succeed in a given position based on matching applicants to a model employee. The model employee is either specified by human programming or is constructed automatically by the algorithm searching for patterns from a set body of data.[58] As one provider of such a service, Infor Talent Management (Infor), describes, its "cloud-based Predictive Talent Analytics . . . solution . . . leverag[es] large quantities of behavioral and performance data," which it "customize[s] into predictive models [that let businesses] better select, retain, and develop the right talent."[59] Infor claims to have access to 19% of the U.S. workforce in its database and advertises that it weighs thirty-nine behavioral characteristics in its algorithm.[60]

Likewise, a company called Entelo provides recruiting software that mines and collects data on potential job applicants to help employers "efficiently discover and qualify talent."[61] Entelo explains that its search engine "[f]ollow[s] the digital footprint of your candidates with social and professional information aggregated from over 50 sites across the web."[62]

b. Testing-and-Tracking Uses

In contrast to data mining and predictive matching, other hiring algorithms focus on measuring applicants' performance on skills-related challenges or tracking and improving upon employers' own hiring practices. A company called GapJumpers creates blind skills-based challenges for employers to use in evaluating candidates and uses algorithms to create and rank applicant results.[63] GapJumpers describes its "performance audition challenges" as a way to "evaluate candidates on work performance . . . rather than keywords on a resume" to "avoid discarding desirable talent that

---

centive to pressure vendors to avoid biased outcomes."). I also set aside what is known as the "cat's paw" problem, in which liability may attach when one decision maker unknowingly carries out the intentional discrimination of a second, for whom the first has served as a "cat's paw" (based on an Aesop fable). *See* Staub v. Proctor Hosp., 562 U.S. 411, 415–16, 422–23, n.1 (2011). For a discussion of liability issues related to third-party platforms and job advertisers, see generally Ajunwa, *supra* note 13, and Kim & Scott, *supra* note 13.

58. *See* Barocas & Selbst, *supra* note 11, at 673–93.

59. *Infor Talent Science*, INFOR, https://www.infor.com/products/talent-science/ (last visited Sept. 26, 2018).

60. *Id.*

61. *Recruiting Information Software*, ENTELO, https://www.entelo.com/products/ (last visited Sept. 24, 2018).

62. *Entelo Platform*, HR.COM, https://www.hr.com/buyersguide/product/view/entelo_entelo_platform (last visited Sept. 24, 2018).

63. *Increase Diversity by Interrupting Hiring Bias*, GAPJUMPERS, https://www.gapjumpers.me/how-it-works/employers (last visited Sept. 26, 2018).

[does] not fit pre-conceived notions."[64] GapJumpers claims that its services lead to "10% more diversity every quarter."[65]

Another company, Textio, uses algorithms to improve companies' job postings "[b]y analyzing the hiring outcomes of more than 10 million job posts a month [to] predict[] the performance of [the] listing" and providing "real-time guidance" on improvement.[66] Textio claims that, on average, using their "augmented writing" algorithms to improve job postings allows employers to "recruit 25% more people qualified enough to interview and 23% more women" and at a pace that is "17% faster" than without their tools.[67]

c. Combined Uses

As uses of algorithms are endless, employers may also combine mining-and-matching with testing-and-tracking uses throughout the hiring process. For example, the company Talent Sonar focuses on five hiring practices, including predictive analytics, to help employers "efficiently find the person who best fits each job from a broader, more qualified candidate pool."[68] While it offers "data-driven hiring decisions from [a scoring engine]," it also includes "blind resume review," "inclusive job descriptions," "structured interviews," and precommitment to hiring qualifications and priorities up front.[69] Talent Sonar claims that its process can "[a]ttract 30% more qualified candidates."[70]

* * *

Whether for mining and matching or testing and tracking, all hiring algorithms tend to incorporate some element of searching for patterns and predicting outcomes in an effort to improve hiring decisions. And regardless of the method used, virtually all algorithmic recruiting services claim to help an employer expand its pool of applicants—often with the stated goal of improving racial, ethnic, and gender diversity in the employer's workforce.[71] For that reason, algorithmic hiring stands to support broader

64. RESOURCE SOLUTIONS, RECRUITMENT OUTSOURCING INSIGHTS 18, https://www.robertwalters.com/content/dam/robert-walters/corporate/news-and-pr/files/whitepapers/resource-solutions-annual-insights-report-11.pdf (last visited Oct. 4, 2018).

65. GAPJUMPERS, https://www.gapjumpers.me (last visited Sept. 15, 2018).

66. *Textio*, WELCOMEAI, https://www.welcome.ai/products/human-resources-recruiting/textio (last visited Sept. 24, 2018); *see also* Kim, *Data-Driven*, *supra* note 13, at 872.

67. *Textio*, *supra* note 66.

68. TALVISTA, ABOUT TALENT SONAR, http://tsarchive.talvista.com/wp-content/uploads/2017/08/Talent-Sonar-Fact-Sheet.pdf (last visited Sept. 24, 2018).

69. *Id.*

70. *Talent Sonar*, TALVISTA, http://tsarchive.talvista.com (last visited Sept. 24, 2018).

71. *See supra* notes 59–70 and accompanying text.

workplace equality as compared to more traditional hiring methods. Yet, as addressed in Part II, when it comes to the potential for discriminatory results, not all algorithms are created equal: the likelihood that algorithmic hiring will result in discrimination varies based on both the type of algorithm and how it is used.

B. Existing Legal Scholarship on Algorithmic Discrimination at Work

Among legal scholarship on the rise of big data, a focus on algorithmic discrimination is a more recent development. In just the past three years, a handful of legal scholars have begun to focus on employment discrimination concerns raised by the use of algorithms at work.[72] This Subpart describes current scholarship documenting the scope of the problem and two distinct solutions to resolve it: improve the algorithms or improve the law.

While application of the law of Title VII to the problem of algorithmic discrimination is explored more fully in Part III below, brief definitions are needed to ground the discussion of the current scholarly debate. Under Title VII, employees who believe they have experienced race or sex discrimination in hiring can challenge the hiring decision using a disparate treatment theory of liability, a disparate impact theory, or both.[73] Employees allege disparate treatment when they believe that they were intentionally treated differently and experienced negative employment consequences because of their protected class status.[74] In contrast, they allege disparate impact when they believe that an employer's equal treatment of all employees resulted in disproportionately negative results for members of their protected class.[75] For the most part, current scholarship on algorithmic discrimination in the workplace characterizes the harm as one of disparate impact and its solutions as shaped by disparate impact concepts—a characterization that, in Part III, this Article challenges.

1. The "Improve the Algorithms" Approach

In their germinal work on the subject, *Big Data's Disparate Impact*, Solon Barocas and Andrew Selbst provide a comprehensive analysis of the way in which the use of algorithms in the workplace can lead to discriminatory results.[76] They then suggest that such discrimination will be difficult, if

72. *See supra* notes 13, 20.
73. *See infra* Part III.
74. *See infra* Subpart III.B.
75. *See infra* Subpart III.A.
76. *See* Barocas & Selbst, *supra* note 11, 677–93.

not impossible, to reach under the existing law of Title VII.[77] As a result, they propose a solution to the problem that focuses on regulating to improve the algorithms themselves, to reduce the incidence of algorithmic discrimination in the first place.[78]

Barocas and Selbst identify five different ways in which algorithms may be biased. As they explain, human programmers may inadvertently introduce bias into a machine-learning algorithm when they identify the goal the algorithm should seek to match (the "target variable") or when they provide it with sample data from which the algorithm "learns" the criteria in an applicant that matches the employer's desired outcomes (the "training data").[79] For example, if programmers train an algorithm to look for "good employees" by correlating "good" with criteria that incorporated bias in the past—such as past subjective performance assessments infected by human bias—or if they ask an algorithm to determine its own pattern for decision-making by matching past biased decisions—such as a data set in which no applicants from historically black colleges were hired—these biases will be reproduced in all future decisions.[80]

Discrimination may also occur, they suggest, when data mining results in "incorrect, partial, or nonrepresentative" data collection that disproportionately disadvantages certain protected classes, like racial minorities, who have less access to technology from which accurate data may be mined.[81] These inaccuracies are compounded by the fact that data mining generalizes from a limited sample set.[82] In addition, discrimination may result from a lack of a rich and specific set of decision-making factors.[83] This leads algorithms to unintended discrimination in two ways: overascribing meaning to each of a few data points from which broader generalizations are then made,[84] or relying on factors that effectively serve as a proxy for protected classes based on correlations that exist in society at large.[85] Lastly, Barocas and Selbst suggest, employers who wish to commit intentional discrimination may do so by manipulating the data involved and "masking" their intentions with algorithmic "neutrality" (though they acknowledge that the expense of such a cover-up makes it so unlikely as to be of little concern).[86]

77. *Id.* at 694–714.
78. *Id.* at 714–22.
79. *Id.* at 677–81.
80. *Id.* at 677–84.
81. *Id.* at 684–87.
82. *Id.* at 686.
83. *Id.* at 691–92.
84. *Id.* at 688–90.
85. *Id.* at 691–92.
86. *Id.* at 692–93.

Despite the many ways in which algorithmic decision-making may lead to discriminatory results, as Barocas and Selbst view it, under existing antidiscrimination law, "some, if not most, instances of discriminatory data mining will not generate liability."[87] Except for the unlikely case of an employer manipulating the data, they suggest that an employer cannot be held liable for algorithmic discrimination under Title VII's disparate treatment framework.[88] Disparate treatment requires proof of intentional discrimination, and by its very nature, Barocas and Selbst posit, algorithmic discrimination is "unintentional."[89] And, although algorithmic discrimination seems to fit under Title VII's disparate impact theory of liability, existing proof structures mean that any disparate impact created by an algorithm may likely be excused under an employer's affirmative defense of "business necessity."[90] In short, disparate treatment does not apply, and disparate impact likely cannot be successfully proven, leaving plaintiffs without a remedy in existing antidiscrimination law for documented patterns of algorithmic discrimination.[91]

As a result, Barocas and Selbst propose an alternative solution to redress algorithmic discrimination: focus on regulating the algorithms that employers use to mitigate discriminatory effects from the outset.[92] Noting that computer scientists are working on technical fixes, they propose a roadmap, grounded in antidiscrimination law, for reducing biased algorithms.[93] In particular, Barocas and Selbst suggest that employers can set better target variables by tightening the "nexus" between attributes that serve as proxies and the skills required by the job, or by experimenting with relying on a variety of different data points that give an accurate result to see which reduces disparate impact.[94] Employers can also attempt to improve training data, the authors suggest, by carefully removing data points that incorporate past bias or by "oversampling" to correct past inaccuracies where collected data is incomplete.[95] Yet, Barocas and Selbst concede, both options may prove technically challenging and possibly cost-prohibitive for

87. *Id.* at 675.

88. *Id.* at 694–701; *infra* Subpart III.A.

89. Barocas & Selbst, *supra* note 11, at 698 ("Except for masking, discriminatory data mining is by stipulation unintentional.").

90. *Id.* at 706–12; *infra* Subpart III.B.

91. *See* Barocas & Selbst, *supra* note 11, at 675, 729 ("By now, it should be clear that Title VII, and very likely other similarly process-oriented civil rights laws, cannot effectively address this situation [of data mining's disparate impact].").

92. *Id.* at 714–22.

93. *Id.* at 714.

94. *Id.* at 715–16.

95. *Id.* at 716–19.

data miners.[96] Ultimately, they conclude that there are few "obvious, complete, or welcome resolution[s]" to removing entirely the potential for disparate impacts of algorithmic decision-making, and that to do so "will necessitate open-ended exploration without any way of knowing when analysts have exhausted the possibility for improvement."[97]

While the challenge is steep and requires more than merely technical solutions, the authors propose that regulating the process for use of algorithmic decision-making can help. In a second work on the topic, *Accountable Algorithms*,[98] Barocas, Joshua Kroll, and coauthors develop what they describe as "a new technological toolkit to verify that automated decisions comply with key standards of legal fairness."[99] The authors provide an extensive explanation of how computer science tools can be used to reduce unfairness in algorithmic decision-making, and they identify the need for a governance structure in which law and policy makers work with computer scientists to ensure fairness.[100] In another work, *Disparate Impact in Big Data Policing*,[101] Selbst proposes one such model of algorithmic governance for the criminal justice context and beyond: requiring "algorithmic impact statements" modeled on a similar approach in environmental law.[102] Users of predictive algorithms would be required to document publicly the expected effectiveness and potential disparate impacts of their technological choices and its reasonable alternatives, subject to public notice and comment.[103]

Scholarship focusing on regulating to improve the algorithms has been invaluable, both in identifying the sources of algorithmic discrimination and, likely, as its most direct means of redress. Yet, without current enforceability, such proposals may fall short. The same scholars who aptly identify the underlying unfairness in the data also suggest that fixing discriminatory algorithms will be difficult and expensive, without a guarantee

96. *See id.* Employers can also seek to incorporate more granular data to reduce statistical discrimination or to remove more data that correlates with protected classes. Again, however, these approaches have a downside: the potential for more unfairness or less accuracy. *Id.* at 719–22.

97. *Id.* at 716–18, 722.

98. Kroll et al., *supra* note 13.

99. *Id.* at 633, 695–705.

100. *See id.* at 695–705. This call for collaboration is well-supported: outside of the legal literature, computer scientists (including Barocas) have launched an entire field of research designed to detect and correct the disparate impacts of algorithms from a technical perspective. *See, e.g.*, Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 KNOWLEDGE & INFO. SYS. 95 (2018), https://doi.org/10.1007/s10115-017-1116-3; Michael Feldman et al., *Certifying and Removing Disparate Impact*, ARXIV (July 16, 2015), https://arxiv.org/pdf/1412.3756.pdf. For a list of additional computer science scholarship on this topic, see *Scholarship*, FAT ML, https://www.fatml.org/resources/relevant-scholarship (last visited Nov. 2, 2018).

101. *See* Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2017).

102. *Id.* at 118–19, 168–82.

103. *Id.*

of making the process more accurate or fair.[104] At the same time, they demonstrate that, in their view, an employer whose use of algorithms results in discrimination likely could be excused from liability under antidiscrimination law.[105] If there is no potential for antidiscrimination liability, then these proposals rely on employers complying with new regulations without clear financial incentives to do so to achieve greater fairness that is difficult to measure.[106]

2. The "Improve the Law" Approach

In the second foundational work in this area, *Data-Driven Discrimination at Work*, Pauline Kim agrees with Barocas and Selbst about the potential risks of algorithmic discrimination but suggests an alternative solution: improve antidiscrimination law to reach this new harm.[107] To Barocas and Selbst's description of the five mechanisms for algorithmic discrimination,[108] Kim adds her own catalogue of harms of using big data at work, including intentional discrimination hiding behind a "legitimate business reason" of "the output of a computer model";[109] individual record errors that result in denial of employment opportunities; data-driven statistical bias that "coincides with systematic disadvantage to protected classes";[110] and algorithms that, while not flawed themselves, nevertheless reproduce structural disadvantage resulting in disparate impacts on protected classes.[111]

Like Barocas and Selbst, Kim questions whether current Title VII doctrine can address the unique challenges of algorithmic discrimination.[112] Kim identifies only one type of algorithmic discrimination that "easily fits within the conventional framework" of Title VII as disparate treatment: when an employer uses a seemingly neutral data model to justify its intent

104. *See* Barocas & Selbst, *supra* note 11, at 716–18.

105. *See id.* at 675, 729.

106. *See* Kim, *Data-Driven*, *supra* note 13, at 891–97 (describing how market-based solutions, alone, will not fix the problem of algorithmic discrimination); *id.* at 894 ("[D]ata models are more likely to exhibit bias, and market competition will not reliably eliminate them. . . . [because] biased data models may be accurate *enough* to persist in a competitive market, even though they are biased against certain groups. . . . [F]eedback effects may appear to confirm the accuracy of biased data models, entrenching their use. . . . [B]iased data models may be efficient precisely *because* they are discriminatory, and therefore pressures toward efficiency will not eliminate them.").

107. *See generally id.*

108. *See id.* at 875–78.

109. *Id.* at 884.

110. *Id.* at 887.

111. *See id.* at 884–90.

112. *Id.* at 903.

to discriminate.[113] While this may not be easy for an employee to prove, Kim notes, "[s]uch a scenario poses no particular conceptual challenge" under current doctrine.[114] Like Barocas and Selbst, Kim views all other algorithmic discrimination as a matter for disparate impact law and recognizes the challenges this poses for proving discrimination.[115] As described in Part III, an employer may prove, as an affirmative defense to a disparate impact claim, that its practice is "job related" and "consistent with business necessity" by validating the practice with proof that the practice is "statistically correlated" with success on the job.[116] Because algorithms are designed to find statistical correlations, Kim argues, the traditional approach to the question of business necessity becomes merely "tautological"—that is, the algorithm can always serve as its own validation, even if it discriminates.[117]

Instead of looking for a regulatory solution focused on improving the algorithms, however, Kim suggests that the law of Title VII may be read to better meet the challenge of proving algorithmic disparate impact.[118] Kim argues that a close reading of the statutory text of Title VII suggests an as-of-yet unrecognized prohibition on what she calls "classification bias"—"the use of classification schemes that have the effect of exacerbating inequality or disadvantage along [the] lines of . . . protected characteristics."[119] Title VII doctrine could "directly prohibit" this sorting bias, Kim suggests, or the current disparate impact framework could and should be altered in four ways to redress algorithmic discrimination.[120] First, employers should be allowed to retain and use information on protected class status from datasets as needed to assess the risks of biased outcomes.[121] Second, employees should be able to make out a prima facie case of disparate impact by using the model's training data rather than the law's current approach of

113. *Id.*; *infra* Subpart III.B.

114. Kim, *Data-Driven*, *supra* note 13, at 865.

115. *See id.* at 902–09; *infra* Subpart III.A.

116. Kim, *Data-Driven*, *supra* note 13, at 866, 908.

117. *Id.* at 866, 908.

118. *Id.* at 869, 902.

119. *Id.* at 890–91, 911. Kim explains that classification bias is a form of disparate impact, not disparate treatment, and distinct from anticlassification theory. *Id.* at 891–92 ("In speaking of classification bias, I do not mean to invoke what is sometimes referred to as 'anticlassification' theory. . . . [which] identifies discriminatory harm primarily in the use of classifications—like race—to make decisions. . . . in contrast to antisubordination theory, which aims to promote equality by redressing structures and practices that disadvantage historically subordinated groups, regardless of [intent]. . . . [T]he concept of classification bias proposed here looks at the consequences of employers' decisions. By asking whether neutral classification schemes work to systematically deprive already disadvantaged groups of opportunities, it shares the concerns of antisubordination theorists.").

120. *See id.* at 916–25.

121. *See id.* at 917–18.

using "relevant labor market," which—given that an algorithm assumes a complete data universe—poses a major obstacle for employee proof.[122] Third, an employer's defense to a disparate impact should require more than statistical correlation: the employer should be required to show that "no problems exist with the data or model construction that are biasing the results."[123] Lastly, employers who detect and correct algorithmic bias in their own decision-making processes should be able to rely on a "bottom-line" defense that their ultimate decisions show no discrimination[124]—an approach that the U.S. Supreme Court rejected in the 1980 case *Connecticut v. Teal*.[125]

In a later piece responding to Kroll and his coauthors' deep dive into technological efforts to correct the algorithms, Kim reinforced her belief in the importance of antidiscrimination law's norms in redressing algorithmic discrimination.[126] While, as Kroll and his coauthors suggest, transparency and auditing of algorithms is only a limited solution (relative to other forms of governance) from the perspective of computer science, Kim noted its importance from the perspective of law, explaining that "[t]echnical tools alone cannot reliably prevent discriminatory outcomes because the causes of bias often lie not in the code, but in broader social processes."[127]

As with the scholars focused on regulatory solutions to improve algorithms, Kim's work, focused on improving antidiscrimination law itself, both adds to our understanding of the problem of algorithmic discrimination and proposes a thoughtful vision for redressing it. Were Kim's proposals to be adopted by courts applying Title VII, algorithmic discrimination would, no doubt, decrease. Yet like Barocas and Selbst's, Kim's approach is limited by its current enforceability; even Kim suggests that, while possible under Title VII as enacted, her proposal requires "fundamentally rethinking antidiscrimination doctrine."[128]

* * *

Both the "improve the algorithms" and "improve the law" approaches have put the issue of algorithmic employment discrimination on the map

122. *See id.* at 918–20.

123. *Id.* at 920–23.

124. *See id.* at 923–25.

125. 457 U.S. 440, 445–56 (1982). *But see* Kim, *Data-Driven*, *supra* note 13, at 924–25 (arguing that this defense makes sense in the context of classification bias because it will incentivize "equality-promoting uses of data" by "encourag[ing] employers to audit the impact of . . . decision-making algorithms . . . to create processes that produce less biased results overall").

126. *See* Kim, *Auditing Algorithms*, *supra* note 13, at 189–91, 202–03.

127. *See id.* at 191.

128. Kim, *Data-Driven*, *supra* note 13, at 865.

and offered compelling proposals to redress this issue. Yet both approaches start from the assumption that antidiscrimination law involves only two theoretical paths: formal equality or substantive equality. Part II of this Article identifies a third principle key to the theoretical framework underlying antidiscrimination law—individual freedom from stereotypes—and raises its practical implications for big data at work. Part III then revisits remedying algorithmic discrimination under existing law using an antistereotyping approach.

II. ALIGNING ALGORITHMS WITH ANTIDISCRIMINATION LAW THEORY

Scholarship on algorithmic discrimination at work has centered on the inability of the law to redress the problem under either of two main theories underlying antidiscrimination law: the anticlassification goal that requires formal equal treatment and the antisubordination goal that seeks substantive equality of opportunities and outcomes. This Part explains that there is a third theory behind antidiscrimination law unexplored in the current debate: an antistereotyping goal that requires not just equal but also *individual* treatment, which may offer additional guidance on the use of algorithms in the workplace.

A. Anticlassification and Antisubordination Theories

Jurists and scholars have wrestled with the principles that should guide antidiscrimination jurisprudence for the past half-century. This debate developed in the context of how best to ensure the constitutional guarantee of Equal Protection after the *Brown v. Board of Education*[129] decision desegregated public education in 1954.[130] With the passage of Title VII in 1964, the same questions arose about applying the statutory law to employment discrimination.[131] Both scholarship and caselaw on the topic reflect a tension between what are recognized as the two main theories for achieving equality: anticlassification theory and antisubordination theory.[132]

129. 347 U.S. 483 (1954).

130. *See* Jack M. Balkin & Reva B. Siegel, *The American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIAMI L. REV. 9, 9–11 (2003); Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles Over Brown*, 117 HARV. L. REV. 1470, 1470–76 (2004).

131. *See* Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 994 (2012).

132. A large body of scholarship addresses this question, a full discussion of which is beyond the scope of this Article. *See generally, e.g.*, Areheart, *supra* note 131; Balkin & Siegel, *supra* note 130; Ruth Colker, *Anti-Subordination Above All: Sex, Race, and Equal Protection*, 61 N.Y.U. L. REV. 1003 (1986); Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107 (1976); Bar-

An anticlassification approach (also known as antidifferentiation) focuses on formal equality and equal treatment of individuals.[133] It defines discrimination as actions that treat individuals differently from one another based on protected class, viewing any acknowledgement of those differences as perpetuating discrimination.[134] In an anticlassification approach to equality, the goal is "colorblindness"—that the law should entirely ignore protected class status.[135] For example, in a 2007 Supreme Court decision striking down the use of race as a factor in assigning students to specific schools within a public school district to achieve racial diversity, Chief Justice John Roberts opined, "The way to stop discrimination on the basis of race is to stop discriminating on the basis of race."[136] Such an approach would treat all individuals the same, regardless of the fact that individuals to whom the same treatment is applied start from different positions of social status and advantage.[137]

An antisubordination approach (also known as antisubjugation) focuses on substantive equality and both equal opportunities and equal outcomes between groups.[138] It defines discrimination as actions that perpetuate social hierarchy and the oppression of historically disadvantaged groups.[139] Antisubordination theory recognizes that the law must consider how members of protected classes are situated differently within society based on the historical context of the social-status subordination of racial minorities and women.[140] For example, in his partial dissent from a 1978 decision on the use of race in public university medical school admissions to achieve racial diversity, Justice Harry Blackmun explained, "In order to get beyond racism, we must first take account of race[;]. . . in order to treat some persons equally, we must treat them differently."[141] This approach views formal equal treatment as insufficient to root out discrimination because, if racial

---

bara J. Flagg, *"Was Blind, but Now I See": White Race Consciousness and the Requirement of Discriminatory Intent*, 91 MICH. L. REV. 953 (1993); Charles R. Lawrence III, *The Id, the Ego, and Equal Protection: Reckoning with Unconscious Racism*, 39 STAN. L. REV. 317 (1987); Siegel, *supra* note 130; Reva B. Siegel, *From Colorblindness to Antibalkanization: An Emerging Ground of Decision in Race Equality Cases*, 120 YALE L.J. 1278 (2011).

133. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1005–06.

134. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1005–06.

135. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1005–06.

136. Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 748 (2007); *see also* Stacy L. Hawkins, *A Deliberate Defense of Diversity: Moving Beyond the Affirmative Action Debate to Embrace a 21st Century View of Equality*, 2 COLUM. J. RACE & L. 75, 90–98 (2012).

137. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1005–10.

138. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1007–10.

139. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1007–10.

140. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1007–10.

141. Regents of Univ. of Cal. v. Bakke, 438 U.S. 265, 407 (1978) (Blackmun, J., concurring in part, dissenting in part); *see also* Hawkins, *supra* note 136, at 96.

or gender minorities start from a disadvantaged position and are treated the same as more advantaged majorities, they will remain in a lower status position in perpetuity.[142]

The same theories of equality apply whether an antidiscrimination claim arises under constitutional law (Equal Protection) or under statutory employment law (Title VII); yet their role in the legal doctrine varies.[143] As explored in Part III, caselaw interpreting the Equal Protection Clause has, for the most part, limited its doctrine to an anticlassification approach.[144] Plaintiffs alleging protected class discrimination under the Equal Protection Clause may only pursue claims of unequal treatment; no disparate impact theory of liability is available.[145] In contrast, Title VII recognizes both the anticlassification principle, in its prohibition of disparate treatment,[146] and the antisubordination principle, in its prohibition of unjustified disparate impact.[147]

In the context of algorithmic discrimination, current scholarship focuses on the inability of anticlassification and antisubordination approaches to encompass the problem.[148] Because they view the use of algorithmic decision-making as a "neutral" employment practice, scholars agree that any resulting discrimination is primarily an issue of disparate impact that implicates antisubordination goals, rather than disparate treatment that implicates anticlassification goals.[149] They also agree that the disparate impact

142. *See* Balkin & Siegel, *supra* note 130, at 9–11; Colker, *supra* note 132, at 1005–10.

143. Since an amendment to Title VII in 1972 that extended the statute to cover state and federal governments, public sector employees may pursue an employment discrimination claim under either Title VII or the Equal Protection Clause. *See generally* Stephen M. Rich, *One Law of Race?*, 100 IOWA L. REV. 201 (2014) (discussing points of "convergence" and "divergence" between Title VII and Equal Protection doctrine).

144. *See* Washington v. Davis, 426 U.S. 229 (1976); *infra* Subpart III.C. *But see* Balkin & Siegel, *supra* note 130, at 10–11 (suggesting that antisubordination principles are still present in the Court's application of strict scrutiny to facial discrimination in the context of affirmative action); Siegel, *supra* note 130, at 1541–42 ("It is generally assumed that when the Court required plaintiffs challenging facially neutral state action to prove discriminatory purpose, it was embracing anticlassification values and repudiating antisubordination values. Yet[,] . . . [in] the Court's affirmative action cases[,] . . . the judiciary has developed the concept of discriminatory purpose with sensitivity to the social status of groups that government benefits and burdens . . . [E]ven in the area of discriminatory purpose doctrine, the Equal Protection Clause has been interpreted in ways that vindicate concerns about group subordination. . . . [C]oncerns about subordination shape the concept of classification itself.").

145. *See Washington*, 426 U.S. at 242–52.

146. *See* 42 U.S.C. § 2000e-2(a)(1) (2012); McDonnell Douglas Corp. v. Green, 411 U.S. 792, 800–01 (1973); *infra* Subpart III.B.

147. *See* 42 U.S.C. § 2000e-2(a)(2); Griggs v. Duke Power Co., 401 U.S. 424 (1971); *infra* Subpart III.A.

148. *See* Barocas & Selbst, *supra* note 11, at 723–28; Kim, *Data-Driven*, *supra* note 13, at 891–92.

149. *See* Barocas & Selbst, *supra* note 11, at 723–28; Kim, *Data-Driven*, *supra* note 13, at 891–92; Kroll et al., *supra* note 13, at 692–94.

doctrine that animates antisubordination theory comes with a proof structure and affirmative defenses that could allow employers to escape liability for algorithmic discrimination at work.[150]

Commentators disagree, however, about the application of the antisubordination principle in Title VII jurisprudence to algorithmic hiring decisions. The debate centers on the U.S. Supreme Court's 2009 decision in *Ricci v. DeStefano*, in which the Court held that, where an employer threw out the results of a promotion exam that it believed had created a disparate impact on black and Latino applicants, those (mostly white) individual employees who would have been promoted had the exam been certified could allege disparate treatment.[151] Barocas and Selbst along with Kroll and his coauthors express concern that the *Ricci* holding might limit employers' ability to improve upon their own algorithms if they detect a resulting disparate impact because, once an algorithm is applied, rejecting its decisions could amount to *Ricci*-style disparate treatment against those whom the algorithm favors.[152] In contrast, Kim explains that such a scenario would be factually distinct from *Ricci*, in which, but for the employer's rejection of the exam results, the specific plaintiffs who brought the lawsuit were guaranteed promotion based on a pre-set plan under a union contract.[153] As Kim rightly observes, because no particular applicants would be identified and guaranteed a job as a result of an employer merely using a hiring algorithm, *Ricci* would pose no obstacle to the employer improving upon its own algorithm should it detect disparate effects.[154]

Regardless of the difference in interpretation of *Ricci*, however, the scholarly disagreement focuses solely on whether and how the antisubordination approach could apply to algorithmic discrimination. Both views share the same starting assumption that antisubordination is the appropriate theoretical frame for the problem.

---

150. *See infra* Subpart III.A.

151. 557 U.S. 557, 576–93 (2009).

152. *See* Barocas & Selbst, *supra* note 11, at 724–28; Kroll et al., *supra* note 13, at 692–94.

153. *See* Kim, *Data-Driven*, *supra* note 13, at 925–32.

154. *See id.* ("Unlike the situation in *Ricci*, prohibiting the use of a biased algorithm does not constitute a disparate treatment violation because there has been no adverse employment action. No employee has been deprived of a job to which he is entitled because no employee has any right or legitimate expectation that an employer will use any particular model. . . . Because disparate treatment violations occur only when employees' legitimate entitlements are disrupted, nothing in *Ricci* . . . prohibit[s] employer attempts to identify and avoid such bias. . . . An employer might not be permitted to fire an employee solely because she was selected using a biased data model. However, Title VII should not be read to prohibit the employer from ceasing to use that model once it discovers the bias.").

B. Antistereotyping Theory

For decades, the debate over how to balance anticlassification and antisubordination principles has dominated much of the discussion of antidiscrimination law. Yet there is a third goal reflected in both the Equal Protection Clause and Title VII that is directly relevant to the issue of algorithmic discrimination: the antistereotyping principle, which requires that people be treated not only equally but also individually under the law.[155] An anticlassification approach requires formal equal treatment of individuals who are members of different groups. An antisubordination approach seeks to equalize opportunities or outcomes between members of different groups. In contrast, an antistereotyping approach requires individual treatment even *within* one's own group. It is, in part, a subspecies of anticlassification theory in that one way to treat everyone equally is to treat each of us as an individual.[156] But it also, independently, does more: it requires that individuals not be held to or judged against stereotypes associated with any protected classes, including those protected classes to which they belong.[157] Anticlassification requires that we treat Woman A and Man B the same, and antisubordination requires that we ensure that All Women are not disadvantaged as compared to All Men. But antistereotyping also requires that we treat Woman A as an individual and not make work-related judgments about her as compared to All Women.

A focus on individualized treatment has long been a part of both Equal Protection and Title VII jurisprudence. In the constitutional context, the U.S. Supreme Court has consistently interpreted the language of the Four-

155. *See* Stephanie Bornstein, *The Law of Gender Stereotyping and the Work-Family Conflicts of Men*, 63 HASTINGS L.J. 1297, 1301–12 (2012) [hereinafter Bornstein, *Gender Stereotyping*]; Stephanie Bornstein, *Unifying Antidiscrimination Law Through Stereotype Theory*, 20 LEWIS & CLARK L. REV. 919, 937–42 (2016) [hereinafter Bornstein, *Unifying*]; Cary Franklin, *Inventing the "Traditional Concept" of Sex Discrimination*, 125 HARV. L. REV. 1307, 1354–58 (2012); Cary Franklin, *The Anti-Stereotyping Principle in Constitutional Sex Discrimination Law*, 85 N.Y.U. L. REV. 83, 120 (2010) [hereinafter Franklin, *Anti-Stereotyping*]. Note that Reva Siegel has proposed her own third principle guiding modern Equal Protection Law, which she terms "antibalkanization"—a discussion of which is beyond the scope of this Article. *See* Siegel, *supra* note 130, at 1281–82 ("Over the decades, observers of the Court have come to describe the dispute in binary terms[:]. . . a colorblind anticlassification principle, premised on the belief that the Constitution protects individuals, not groups, and so bars all racial classifications, except as a remedy for specific wrongdoing [versus] . . . an antisubordination principle that identifies racial stratification (rather than classification) as the wrong and endeavors to rectify the forms of group inequality that race-based and race-salient policies have caused. . . . [T]his binary framework obscures the views of the Justices [in the middle,] who . . . reason from an . . . independent view more concerned with social cohesion than with colorblindness.").

156. *See* Anita Bernstein, *What's Wrong with Stereotyping?*, 55 ARIZ. L. REV. 655, 671, 687, 718–21 (2013) (suggesting that stereotyping is "a technology of actionable discrimination" held to be unlawful where it constrains individuals' freedom).

157. *See* Bornstein, *Gender Stereotyping*, *supra* note 155; Bornstein, *Unifying*, *supra* note 155, at 937–42; Franklin, *Anti-Stereotyping*, *supra* note 155, at 88, 90–91.

teenth Amendment that "[n]o State shall . . . deny to any person within its jurisdiction the equal protection of the laws"[158] as "personal rights" that are "guaranteed to the individual."[159] Under Title VII, the statutory text itself focuses on the individual, making it unlawful for an employer to "discriminate against any individual" or "deprive any individual of employment opportunities."[160] The Court has interpreted Title VII to protect an individual against protected class discrimination even when other members of the protected class, or the protected class as a whole, may not have suffered harm.[161]

The idea that individualized treatment also protects individuals from being held to stereotypes associated with their own protected class first appeared in Equal Protection cases.[162] In a series of cases—litigated by Ruth Bader Ginsburg throughout the 1970s in her then-role as the Director of the Women's Rights Project of the ACLU—the Supreme Court established that antidiscrimination law also served an antistereotyping purpose.[163] The Court held that state laws could not establish rules that support gender role stereotypes, thereby punishing individual men who did not behave like other men or individual women who did not behave like other women.[164] During this time, the Court invalidated state or federal laws that recognized that men were preferable to women as estate administrators for deceased family members;[165] allowed women, but not men, a caregiver's tax deduction;[166] required only men, and not women, to prove dependency on their spouses to receive military or social security survivor benefits;[167] and denied social security benefits to the children of widowed men, but not of widowed

158. U.S. CONST. amend. XIV, § 1.

159. Shelley v. Kraemer, 334 U.S. 1, 22 (1948) (first citing McCabe v. Atchison, Topeka Santa Fe Ry. Co., 235 U.S. 151, 161–62 (1914); then citing Missouri *ex rel.* Gaines v. Canada, 305 U.S. 337, 351 (1938); and then citing Oyama v. California, 332 U.S. 633 (1948)).

160. 42 U.S.C. § 2000e-2(a) (2012).

161. *See, e.g.*, Connecticut v. Teal, 457 U.S. 440, 455–56 (1982) (rejecting the "bottom line" defense).

162. *See* Bornstein, *Gender Stereotyping*, *supra* note 155, at 1301–12; Franklin, *Anti-Stereotyping*, *supra* note 155, at 84–88.

163. *See* Bornstein, *Gender Stereotyping*, *supra* note 155, at 1301–12; Franklin, *Anti-Stereotyping*, *supra* note 155, at 119–42.

164. *See* Bornstein, *Gender Stereotyping*, *supra* note 155, at 1301–12; Franklin, *Anti-Stereotyping*, *supra* note 155, at 119–42.

165. Reed v. Reed, 404 U.S. 71, 77 (1971); Bornstein, *Gender Stereotyping*, *supra* note 155, at 1302–04.

166. Moritz v. Comm'r, 469 F.2d 466, 470 (10th Cir. 1972), *cert. denied*, 412 U.S. 906 (1973); Bornstein, *Gender Stereotyping*, *supra* note 155, at 1304–06.

167. Califano v. Goldfarb, 430 U.S. 199, 216–17 (1977) (social security benefits); Frontiero v. Richardson, 411 U.S. 677, 688–91 (1973) (military housing and medical benefits); Bornstein, *Gender Stereotyping*, *supra* note 155, at 1306–09.

women.[168] Through this series of cases, the Court established that applying laws that incorporated stereotypes associated with gender roles to individuals constituted disparate treatment in violation of the Equal Protection Clause.

In the four decades since, the Court has reinforced and extended the antistereotyping approach to Equal Protection when state actors assume that individuals will or should conform to their protected class stereotype.[169] In 1982, in *Mississippi University for Women v. Hogan*, the Court sided with a male plaintiff who sought admission to the state's female-only nursing school.[170] In invalidating the school's exclusive admissions policy, the Court explained that it "perpetuate[d] the stereotyped view of nursing as an exclusively woman's job[,] . . . lend[ing] credibility to the old view that women, not men, should become nurses, and mak[ing] the assumption that nursing is a field for women a self-fulfilling prophecy."[171] In 1994, in *J.E.B. v. Alabama ex rel. T.B.*, the Court held that, in a paternity and child custody suit, a state prosecutor's use of preemptory challenges to disqualify all male potential jurors based on the assumption that individual female jurors would be more sympathetic to the mother constituted unconstitutional sex discrimination.[172] Noting that it was "reaffirm[ing] what, by now, should be axiomatic," the Court explained that "discrimination [that] serves to ratify and perpetuate invidious, archaic, and overbroad stereotypes about the relative abilities of men and women" violates Equal Protection.[173] In *United States v. Virginia*, a 1996 decision authored by Justice Ginsburg, the Court held that a state military college could not constitutionally exclude women from admission by "rely[ing] on overbroad generalizations about the different talents, capacities, or preferences of males and females" when hundreds of individual women had applied for admission.[174] And in 2003, in *Nevada Department of Human Resources v. Hibbs*, the Court held that a man who was fired while on leave to care for his injured wife could sue his public employer under the federal Family and Medical Leave Act because the statute was enacted to remedy sex discrimination by providing family leave to both men and women so as to overcome "mutually reinforcing [gender] stereotypes" and "a self-fulfilling cycle of discrimination that forced women . . . [into] the role of primary family caregiver" and men out

168. Weinberger v. Wiesenfeld, 420 U.S. 636, 650–53 (1975); Bornstein, *Gender Stereotyping*, *supra* note 155, at 1309–12.

169. *See* Franklin, *Anti-Stereotyping*, *supra* note 155, at 142–72.

170. 458 U.S. 718, 729–30 (1982).

171. *Id.*

172. 511 U.S. 127, 137–46 (1994).

173. *Id.* at 130–31.

174. 518 U.S. 515, 533 (1996); *see also* Franklin, *Anti-Stereotyping*, *supra* note 155, at 143–46.

of it.[175] Beyond the Supreme Court, several federal district courts and circuit courts of appeals have now applied the antistereotyping approach to Equal Protection to strike down a wide array of state actions that punish individuals for failure to conform to gender stereotypes.[176]

In the context of Title VII, references to protected class stereotypes appeared in employment cases throughout the 1970s as well.[177] In an early and influential case, *Los Angeles Department of Water & Power v. Manhart*,[178] the U.S. Supreme Court held that an employer could not require female employees to contribute more than male employees to its pension fund, despite the fact that women lived longer than men on average and, therefore, received greater pension payouts.[179] Noting that "employment decisions cannot be predicated on mere 'stereotyped' impressions about the characteristics of males or females," the Court explained that assuming any one individual would meet the stereotype of the group violated the law.[180] "The statute's focus on the individual is unambiguous," the Court explained; thus, "[e]ven a true generalization about the class is an insufficient reason for disqualifying an individual to whom the generalization does not apply."[181]

Then, in 1989, in the case of *Price Waterhouse v. Hopkins*,[182] the Court first articulated what has become known as the "stereotype theory" of liability under Title VII. Plaintiff Ann Hopkins sued her employer, a top accounting firm, for sex discrimination after she was passed over for promotion to partner.[183] Hopkins had been an outstanding employee with

175. 538 U.S. 721, 736 (2003); *see also* Franklin, *Anti-Stereotyping*, *supra* note 155, at 149–54.

176. *See, e.g.*, Glenn v. Brumby, 663 F.3d 1312, 1317 (11th Cir. 2011) (holding that a state employer's termination of a transgender employee whose sex assigned at birth was male was a penalty for failure to conform to masculine gender stereotype, in violation of Equal Protection Clause); Smith v. City of Salem, 378 F.3d 566, 572 (6th Cir. 2004) (same); Knussman v. Maryland, 272 F.3d 625, 636–37 (4th Cir. 2001) (holding that a state employer's denial of "primary caregiver" leave to a male employee violated the Equal Protection Clause, like other "[g]ender classifications based upon generalizations about typical gender roles in the raising and nurturing of children" (first citing Caban v. Mohammed, 441 U.S. 380 (1979); and then citing Stanley v. Illinois, 405 U.S. 645 (1972))); Free the Nipple v. City of Fort Collins, 237 F. Supp. 3d 1126, 1133 (D. Colo. 2017) (holding that a municipal ordinance penalizing women, but not men, for exposing their breasts in public likely violated Equal Protection Clause because it was "based on an impermissible gender stereotype that results in a form of gender-based discrimination"), *appeal filed*, No. 17-1103 (10th Cir. Mar. 21, 2017).

177. *See, e.g.*, City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 704–11 (1978); Sprogis v. United Air Lines, Inc., 444 F.2d 1194, 1198 (7th Cir. 1971).

178. *Manhart*, 435 U.S. at 702.

179. *Id.* at 704–11.

180. *Id.* at 707–08.

181. *Id.* at 708.

182. 490 U.S. 228, 235, 251 (1989) (plurality opinion), *superseded in part by statute on other grounds*, Civil Rights Act of 1991 § 107(a), 42 U.S.C. § 2000e-2(m) (2012), *as recognized in* Burrage v. United States, 571 U.S. 204, 213 n.4 (2014).

183. *Id.* at 231–32.

superior qualifications, but she was criticized for failing to conform to assumptions about how, as a woman, she should behave at work.[184] The decision makers criticized Hopkins for being too "aggressive," "macho," and "masculine," and suggested that she should look and behave more "femininely" if she wanted to be selected for partnership.[185] The Court held in Hopkins's favor, finding that the assessment of her work performance was impermissibly influenced by gender stereotypes.[186] As the Court explained, "[i]n forbidding employers to discriminate against individuals because of their sex, Congress intended to strike at the entire spectrum of disparate treatment of men and women resulting from sex stereotypes."[187]

In the three decades since the *Price Waterhouse* decision, the stereotype theory of liability has become a significant part of Title VII jurisprudence.[188] While evidence on the operation of stereotypes has played a role in countless Title VII cases, the antistereotyping principle in Title VII has been particularly recognized by courts where plaintiffs allege discrimination on the basis of family caregiving responsibilities or transgender status.[189] Many federal district courts and circuit courts of appeals have applied the U.S. Supreme Court's *Price Waterhouse* theory to hold that, when employees are penalized at work based on stereotypes about how they will or should behave, that constitutes disparate treatment.[190] Courts have held that, when a female employee is denied a promotion or otherwise penalized at work based on assumptions related to her status as a mother—for example, that she will be (or should be) less committed to or focused on work—an employer has violated Title VII's prohibition of disparate treatment under a stereotype theory.[191] Likewise, courts have held that an employer violates Title VII when it fires or otherwise penalizes transgender employees based on assumptions related to their gender presentation—that

184. *Id.* at 235, 250–55.

185. *Id.* at 235.

186. *Id.* at 255–58.

187. *Id.* at 251 (citations omitted).

188. *See generally* Bernstein, *supra* note 156 (analyzing stereotype theory under Title VII); Bornstein, *Unifying*, *supra* note 155 (same); Kerri Lynn Stone, *Clarifying Stereotyping*, 59 U. KAN. L. REV. 591 (2011) (same); Kimberly A. Yuracko, *Soul of a Woman: The Sex Stereotyping Prohibition at Work*, 161 U. PA. L. REV. 757 (2013) (same); *see also* Franklin, *Anti-Stereotyping*, *supra* note 155 (analyzing stereotype theory under Equal Protection).

189. *See* cases cited *infra* notes 191–92; *see also* Bornstein, *Unifying*, *supra* note 155, at 937–42; Bornstein, *Gender Stereotyping*, *supra* note 155, at 1301–12.

190. *See* cases cited *infra* notes 191–92; *see also* Bornstein, *Unifying*, *supra* note 155, at 937–42, 962–63 (describing prescriptive and descriptive stereotyping); Bornstein, *Gender Stereotyping*, *supra* note 155, at 1301–12.

191. *See, e.g.*, Chadwick v. WellPoint, Inc., 561 F.3d 38, 42 n.4, 48 (1st Cir. 2009); Back v. Hastings on Hudson Union Free Sch. Dist., 365 F.3d 107, 121–22 (2d Cir. 2004).

is, assumptions that a person whose sex assigned at birth was male should look or behave according to masculine gender stereotypes.[192]

Along with antisubordination and anticlassification, the antistereotyping principle is now a well-developed theory in antidiscrimination law, arising in cases brought under both the Equal Protection Clause and Title VII. In the context of algorithmic discrimination, two pieces of stereotype doctrine are of particular importance. First, while case law has developed mostly in the context of sex stereotyping, courts have not so limited the theory, which applies to race, sex, and other protected classes equally under both Title VII and the Equal Protection Clause.[193] While alleged less often, employee plaintiffs can and do succeed on claims of racial stereotyping under Title VII.[194]

Second, in cases brought under a stereotype theory, courts have made clear that a plaintiff may make out a case of disparate treatment even without providing what is known as "comparator evidence"—evidence that others who were similarly situated outside of the relevant protected class were treated better.[195] The so-called comparator requirement grew out of the U.S. Supreme Court's statement in *McDonnell Douglas Corp. v. Green*—the original case laying out disparate treatment theory under Title VII—in which the Court stated that, to prove race discrimination, evidence of a similarly situated white employee who was treated better than the black plaintiff would be "[e]specially relevant."[196] Despite several later statements by the Court clarifying that such evidence was not required, some lower courts continue to misapply the original holding, requiring a plaintiff to provide comparator evidence to prevail.[197] In cases alleged under a ste-

192. *See, e.g.*, EEOC v. R.G. & G.R. Harris Funeral Homes, 884 F.3d 560, 572 (6th Cir. 2018), *petition for cert. filed*, No. 18-107 (U.S. July 20, 2018); Glenn v. Brumby, 663 F.3d 1312, 1320–21 (11th Cir. 2011); Barnes v. City of Cincinnati, 401 F.3d 729, 737–38 (6th Cir. 2005); Smith v. City of Salem, 378 F.3d 566, 572 (6th Cir. 2004). *But see* Etsitty v. Utah Transit Auth., 502 F.3d 1215, 1221–22, 1224 (10th Cir. 2007) (holding that "sex" under Title VII does not encompass transgender status discrimination, without addressing gender nonconformity under a sex stereotyping theory).

193. *See* Bornstein, *Unifying*, *supra* note 155, at 941 n.118; *see also* J.E.B. v. Alabama *ex rel.* T.B., 511 U.S. 127, 139–40 (1994) (holding that "gross generalizations" are constitutionally impermissible under the Equal Protection Clause, whether made on the basis of race or gender); Price Waterhouse v. Hopkins, 490 U.S. 228, 251–52 (1989) (plurality opinion) (holding that, in a Title VII case, "[b]y focusing on Hopkins' specific proof . . . we do not suggest a limitation on the possible ways of proving that stereotyping played a motivating role in an employment decision"), *superseded in part by statute on other grounds*, Civil Rights Act of 1991 § 107(a), 42 U.S.C. § 2000e-2(m) (2012), *as recognized in* Burrage v. United States, 571 U.S. 204, 213 n.4 (2014).

194. *See* Bornstein, *Unifying*, *supra* note 155, at 963–72 (citing, e.g., Thomas v. Eastman Kodak Co., 183 F.3d 38, 59 (1st Cir. 1999); Kimble v. Wis. Dep't of Workforce Dev., 690 F. Supp. 2d 765 (E.D. Wis. 2010)).

195. *Id.*

196. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 804–05 (1973); *see also* Bornstein, *Unifying*, *supra* note 155, at 942–45.

197. *See* Bornstein, *Unifying*, *supra* note 155, at 942–45.

reotype theory, however, numerous courts have established that a plaintiff may make out a case of disparate treatment under stereotype theory without comparator evidence because, where a work-related decision is made on the basis of a stereotype associated with a protected class, that "can by itself and without more be evidence of an impermissible, [protected class]-based motive."[198] This makes doctrinal sense: if antistereotyping requires that we treat Woman A as an individual rather than as a member of All Women, comparator evidence to Man B or to All Men is irrelevant.

As discussed in Part III, the ability to argue a stereotyping theory for disparate treatment based on any protected class and without having to provide comparator evidence may help victims of algorithmic discrimination, for whom such evidence is unavailable. Moreover, the antistereotyping principle offers an important additional perspective from which to consider the use of algorithms at work.

C. Lessons from Theory: Toward Antidiscriminatory Algorithms

As described in Part I, how algorithms are designed and used to help employers recruit and hire varies, which means the possibility that they will lead to discriminatory results will vary as well. Focusing on anticlassification and antisubordination theories is helpful for identifying how algorithms discriminate and how to make them less discriminatory. Including antistereotyping theory takes this analysis one step further, to offer lessons not just for preventing algorithms from actively discriminating, but also for using algorithms in a way that actually improves upon current human decision-making—to make them affirmatively *anti*discriminatory.

Despite their concerns that structural bias is the greater risk, current scholars do recognize that the use of algorithms offers the potential to reduce human cognitive biases and widen candidate pools.[199] For this reason, they also suggest that, in addition to improving the algorithms or the law itself, employers should change the ways in which they use algorithms to greater support Title VII's antisubordination goals.[200] Barocas and Selbst agree that, by understanding the potential for algorithmic discrimination, employers can improve their models to reduce their disparate impacts by "mak[ing] more effective use of the tools that computer scientists have begun to develop."[201] Kim offers that workforce analytics "can be a useful

198. Back v. Hastings on Hudson Union Free Sch. Dist., 365 F.3d 107, 122, 124 (2d Cir. 2004); *see also* Bornstein, *Unifying*, *supra* note 155, at 944–50.

199. *See* Barocas & Selbst, *supra* note 11, at 673–74, 731–32; Kim, *Data-Driven*, *supra* note 13, at 869–71.

200. *See* Barocas & Selbst, *supra* note 11, at 720–23.

201. *Id.* at 731–32.

tool for diagnosing both cognitive and structural forms of bias."[202] Citing Textio—the algorithm that detects biased language in job postings—as an example, Kim suggests that, instead of relying on data tools to make employment-related decisions, employers could use them to analyze the "decision-making process itself," with the potential to discover and correct "hidden biases."[203] Whether data is helpful or harmful, Kim suggests, "depends a great deal on how the algorithms are constructed and deployed."[204]

Adding an antistereotyping lens to the issue of algorithmic decision-making at work offers another benefit: it may help sort stereotype-activating uses of algorithms from stereotype-suppressing uses, even when the antisubordination principle seems met. Mining-and-matching uses of algorithms appear to be the most likely to rely on stereotypes because, as described in Part III, predictive matching algorithms judge individuals against a model employee that may incorporate protected class stereotypes. Yet if a mining-and-matching use of an algorithm is combined with the specific goal of increasing workforce diversity, it may ultimately result in bottom-line employment decisions that meet an antisubordination goal.[205] This may obscure the fact that the algorithm itself relies on potentially questionable stereotyping.

For example, Entelo advertises that its predictive matching algorithm allows employers to highlight diverse candidates to increase workforce racial or gender diversity.[206] Among the data it mines, Entelo also provides a "More Likely To Move™" feature, which it claims spots applicants with a likelihood of changing jobs "within the next 90 days."[207] The ability to change jobs—particularly if it requires geographic relocation—is a factor that may disadvantage women in relation to men, based on women's greater family caregiving responsibilities and, if relevant, reliance on a male

202. Kim, *Data-Driven*, *supra* note 13, at 871–72 (emphasis omitted).

203. *Id.*

204. *Id.* at 874.

205. *Cf.* Connecticut v. Teal, 457 U.S. 440, 442 (1982) (explaining lack of a "bottom line" defense).

206. *Diversity Recruiting Software*, ENTELO, https://www.entelo.com/products/platform/diversity/ (last visited Oct. 4, 2018); *Entelo Diversity*, ENTELO, https://www.entelo.com/wp-content/uploads/2017/09/diversity_DS-6.27.16.pdf (last visited Nov. 2, 2018) ("Entelo's proprietary algorithm [helps companies] . . . [f]ind candidates from underrepresented groups based on gender, race/ethnicity, and veteran status[.] . . . Entelo Diversity allows companies of all sizes to reap the benefits of building strong, diverse teams. Additionally, since information is layered on top of a candidate's skill-set, the solution provides a level of objectivity as it relates to your hiring practices.").

207. *Recruiting Automation Platform*, ENTELO, https://www.entelo.com/products/platform/ ("Focus your efforts where it counts. Entelo analyzes dozens of variables to predict candidates' receptiveness to new opportunities. Candidates who are More Likely To Move™ are 2X more likely to make change within the next 90 days, than other candidates." (emphasis omitted)) (last visited Oct. 4, 2018).

partner's job income.[208] An unstated requirement that applicants leave a current job to take the one for which they applied is not actionable under Title VII, and a stated requirement that, to be hired or promoted, applicants must be willing to relocate multiple times might be actionable only as disparate impact; neither raises any issue of stereotyping. But an unspoken and unstated assumption that an individual applicant is unlikely to move in the future based on data by comparison to past group behavior, which preemptively knocks the applicant out from consideration, raises a potential issue of stereotyping. If women as a group tend to change jobs less, and algorithms look for patterns among group data, then any individual applicant who looks like a woman may be excluded from consideration on the assumption that she will not change jobs in the future. She may be disadvantaged in relation both to men and to women whose past work behavior conforms to a stereotypically masculine career trajectory. If the candidate is qualified, the employer should ask her about the likeliness of moving, rather than having an algorithm "predict" (that is, "assume") the answer based on future conformity to past behavior—just as in *Manhart*, the employer could not predict that any individual woman would live longer than a man based on past group data.[209] Even if, at the end of the process, the employer hires a woman—thus meeting an antisubordination goal—the antistereotyping goal can help flag that incorporating this factor into an algorithm might be problematic.

Likewise, applying stereotype theory helps identify that testing-and-tracking uses of algorithms may be more likely than mining-and-matching uses of algorithms to suppress protected class stereotypes and human cognitive biases. Instead of solely expanding a pool of candidates and allowing employers to ultimately factor in diversity, both GapJumpers and Textio adopt strategies known from social science research to help reduce the operation of stereotypes: "blinding" decision makers to candidates' protected class status and "interrupting" actions that may unknowingly incorporate bias.[210] GapJumpers was, in fact, modeled on a famous study documenting how blind auditions can reduce the operation of sex-based stereotypes in

208. *See* Naomi Schoenbaum, *The Family and the Market at* Wal-Mart, 62 DEPAUL L. REV. 759, 759–60 (2013).

209. *See* City of L.A. Dep't of Water & Power v. Manhart, 435 U.S. 702, 704–11 (1978); text accompanying *supra* notes 177–81.

210. *See* Joan C. Williams, *Hacking Tech's Diversity Problem*, HARV. BUS. REV. (Oct. 2014), https://hbr.org/2014/10/hacking-techs-diversity-problem; *Bias Interrupters*, CTR. FOR WORKLIFE LAW: UNIV. OF CAL. HASTINGS COLL. OF THE LAW, http://worklifelaw.org/projects/bias-interrupters (last visited Sept. 10, 2018); *Fall 2017 Corporate Program Meeting Recap*, VMWARE WOMEN'S LEADERSHIP INNOVATION LAB: STANFORD UNIV., https://womensleadership.stanford.edu/blueprint (last visited Sept. 10, 2018).

hiring decisions.[211] Similarly, TalentSonar's feature to help employers "precommit" to stated hiring criteria was based on studies documenting how selection criteria may shift over time, often unintentionally, to favor those criteria held by candidates in dominant protected classes.[212]

Whether an algorithm could result in exacerbating or, alternatively, reducing protected class biases will depend on both how it is created and how it is used. Certainly any attempt to reduce human decision-making and widen and diversify pools of job candidates could do more to serve Title VII's antidiscrimination goals than relying on traditional hiring methods. But considering Title VII's antistereotyping principle also offers important lessons for suppressing the operation of biases and stereotypes within algorithmic decision-making that may escape detection with only a bottom-line focus on antisubordination principles.

III. REMEDYING ALGORITHMIC DISCRIMINATION UNDER EXISTING LAW

Considering antidiscrimination theory as a way to improve algorithms and the uses to which they are applied in the workplace, while useful, is not enough to spark needed change. To that end, this Part revisits the idea of employer liability for algorithmic discrimination under existing antidiscrimination law. Current scholarship characterizes algorithmic discrimination as mainly a problem of disparate effects yet also suggests that it could escape the reach of disparate impact doctrine. This Part questions both the characterization of the problem as one of only disparate impact and the limitations of the law's reach. It then concludes with implications for algorithmic discrimination in contexts other than employment, for which disparate impact liability is not an option.

A. Title VII Disparate Impact

In interpreting Title VII, the U.S. Supreme Court has read the statute to allow two main legal frameworks for proving discrimination: disparate treatment and disparate impact.[213] As described previously, disparate impact reflects antisubordination principles of substantive equality by recognizing that, in some cases, treating all people the same when they are

211. *See* Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715 (2000).

212. *See* Eric Luis Uhlmann & Geoffrey L. Cohen, *Constructed Criteria: Redefining Merit to Justify Discrimination*, 16 PSYCHOL. SCI. 464 (2006).

213. Other litigation theories are categorized as falling within these two main divisions; for example, retaliation, harassment, and pattern-or-practice claims are all types of disparate treatment. *See, e.g.*, Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 986–89 (1988).

situated differently, as members of different protected classes, may have discriminatory results. The disparate impact framework was first articulated by the U.S. Supreme Court in the 1971 case *Griggs v. Duke Power Co.*,[214] based on the statutory text of Title VII making it unlawful for an employer "to limit, segregate, or classify . . . employees or applicants . . . in any way which would deprive . . . any individual of employment opportunities or otherwise adversely affect his status as an employee, because of . . . race, color, religion, sex, or national origin."[215] Disparate impact applies when an employer adopts a "facially neutral" policy or practice that it applies to all applicants or employees equally, but which results in a disproportionately negative impact on members of one protected class.[216] Because the harm focuses on the discriminatory result of a seemingly fair practice, proof of an employer's intent in adopting the practice is not required.[217]

To prove a disparate impact case under Title VII, a plaintiff or class of plaintiff applicants or employees must first make out a prima facie case by showing that the defendant employer's practice resulted in a statistically significant disparity against a protected class when compared to the proper labor market.[218] This showing creates an inference of discrimination and shifts the burden of proof to the employer, who can rebut the statistics to show that there is, in fact, no disparate impact to meet plaintiffs' burden of proof.[219] Alternatively, the employer can prove the affirmative defense that the employment practice is "job related for the position in question and consistent with business necessity," which will excuse the disparate impact as lawful.[220] Even if the employer proves business necessity, plaintiffs may still prevail by showing that there is a less discriminatory "alternative employment practice" that the employer "refuses to adopt."[221]

Current scholars agree that algorithmic decision-making may lead to disparate impact discrimination, but raise concerns that disparate impact law as currently construed may allow employers to escape liability.[222] First, assuming that plaintiffs can show that an employer's use of algorithmic decision-making created a disparate impact by protected class,[223] a court

214. 401 U.S. 424 (1971).

215. 42 U.S.C. § 2000e-2(a)(2) (2012).

216. *Griggs*, 401 U.S. at 430.

217. *Id.*

218. 42 U.S.C. § 2000e-2(k)(1)(A)(i); *see also Griggs*, 401 U.S. at 432.

219. 42 U.S.C. § 2000e-2(k)(1)(B)(ii).

220. *Id.* § 2000e-2(k)(1)(A)(i).

221. *Id.* § 2000e-2(k)(1)(A)(ii).

222. *See* Barocas & Selbst, *supra* note 11, at 675; Kim, *Data-Driven*, *supra* note 13, at 902–09.

223. Barocas and Selbst do not focus on the challenges to the plaintiff of making out a prima facie case, instead focusing on the other two parts of disparate impact doctrine. Kim suggests that it may be difficult for plaintiffs to identify the "[proper] labor market" for comparison given that an algo-

could allow an employer to prevail by using the algorithm itself to prove its affirmative defense.[224] Traditionally, though not required by law, employers prove that a practice is job related and consistent with business necessity through what is known as a "validation study" that demonstrates that the practice is a valid measure for job performance.[225] Among possible validation studies are "criterion" or "construct" validation studies that prove how closely a challenged practice measures the likelihood of success on the job in question.[226] As scholars correctly explain, a predictive matching algorithm that is trained on data of a model "good employee" and then selects candidates for hire that match the model is self-validating.[227] Because a predictive matching algorithm is, in essence, a criterion validation study, even if it incorporates discriminatory data, a court could choose to excuse any disparate impact it causes because it selects employees in a way that is inherently job-related.[228]

That said, there may be room under existing law to require more demanding proof from an employer before the employer is found to satisfy its affirmative defense. While an algorithm will always meet the "job related" half of the defense by design, it may not necessarily meet the "business necessity" half. It is true that, as Barocas and Selbst suggest from their survey of relevant caselaw, courts have not been exacting in applying the business necessity standard, instead allowing some amount of "job-relatedness" to suffice.[229] They describe the affirmative defense as lying "somewhere in the middle of two extremes": "that the hiring criteria bear a

---

rithm assumes a closed universe of data. To remedy this, she proposes, as one of her four suggestions to improve the law, that plaintiffs be able to show a disparate impact on the training data rather than in comparison to the relevant labor market. *See* Kim, *Data-Driven*, *supra* note 13, at 919–20; *see also supra* Section I.B.2.

224. *See* Barocas & Selbst, *supra* note 11, 704–09; Kim, *Data-Driven*, *supra* note 13, at 866–67, 905–09.

225. *See* EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §§ 1607.5, 1607.15–.16 (1978).

226. *Id.*

227. *See* Barocas & Selbst, *supra* note 11, at 704–09; Kim, *Data-Driven*, *supra* note 13, at 866–67, 905–09.

228. *See* Barocas & Selbst, *supra* note 11, at 708–09 ("[T]here must be statistical significance showing that the result of the model correlates to the trait . . . determined to be an important element of job performance[]. This is an exceedingly low bar for data mining because data mining's predictions necessarily rest on demonstrated statistical relationships. . . . Thus, there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense."); Kim, *Data-Driven*, *supra* note 13, at 908 ("If an employer could meet this burden simply by showing that an algorithm rests on a statistical correlation with some aspect of job performance, then the test is entirely tautological, because, by definition, data mining is about uncovering statistical correlations. Any reasonably constructed model will satisfy the test, and the law would provide no effective check on data-driven forms of bias.").

229. *See* Barocas & Selbst, *supra* note 11, at 705 (describing that "all circuits seem to accept varying levels of job-relatedness rather than strict business necessity").

'manifest relationship' . . . or . . . be 'significantly correlated' to job performance" on one side and that it "accurately—but not perfectly—ascertains an applicant's ability to perform [the job] successfully" on the other.[230]

Yet because the law does not *require* a criterion validation study, it also need not be satisfied by one. In addition to criterion and construct validation, there is a third type of validation study, "content" validation, which tests the validity of a selection practice by testing how closely it measures successful performance on the actual tasks of a job rather than characteristics of successful employees in the abstract.[231] Given that a hiring algorithm could serve as a criterion validation of itself but still discriminate, courts could hold that such a study is not enough to meet the employer's burden of proof and instead require content validation. Similarly, Kim proposes that, under existing law, in the context of algorithmic discrimination, statistical correlation alone should not be enough to meet an employer's defense of business necessity; instead, courts should require employers to "defend the accuracy of the correlations" as unbiased.[232] An alternative to Kim's suggestion is for courts to hold that, in cases challenging algorithmic discrimination, criterion validation studies do not satisfy the employer's burden of proof for the business necessity affirmative defense. Both proposals may be possible under existing doctrine: while the basic affirmative defense was created by statute, it was not further defined, and whether it has been met in a particular case is a matter for the fact finder.[233]

What is more, requiring a content validation study rather than a criterion validation study could encourage employers toward stereotype-reducing uses of algorithms and away from stereotype-activating ones.[234] For example, a content validation study would more closely resemble a blind-skills-challenge algorithm that recognizes and tests applicants' skills for the job rather than their qualities in the abstract. If only content, but neither criterion nor construct, validation studies were allowed to support a business necessity defense in a lawsuit alleging algorithmic disparate impact, employers may move away from relying on the more problematic mining-and-matching algorithms that seek to match an abstract "good employee" profile—or at least toward a combined use of both mining-and-matching

230. *Id.* at 704–05 (footnote omitted).

231. *See* EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §§ 1607.5, 1607.15–.16.

232. Kim, *Data-Driven*, *supra* note 13, at 921; *see also supra* Section I.B.2.

233. *See, e.g.*, Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478 (3d Cir. 1999).

234. *See supra* Subpart II.C.

and testing-and-tracking algorithms—to reduce bias in their hiring processes.[235]

In addition, even if a court defers to an employer's business necessity proof, plaintiffs may still prevail if they can prove that the employer refused to adopt a less discriminatory alternative practice.[236] While acknowledging that this may be the most promising option for liability under existing disparate impact law, Barocas and Selbst remain less than optimistic. As they describe it, "a plaintiff could argue that the obvious alternative employment practice would be to fix the problems with the models," yet because "[f]ixing the models . . . is not a trivial task," the chance that a plaintiff could prevail in such a circumstance "seems slim."[237] In particular, they express concern that, even if plaintiffs could identify how to cure a model of its bias, they must still prove that the employer refused to adopt it.[238]

Again, this concern may understate the reach of existing disparate impact law. While it is true that, as Barocas and Selbst identify, "[n]either Congress nor courts have specified what it means for an employer to 'refuse' to adopt the less discriminatory procedure," caselaw has made clear that proof that an employer could use *its own* selection device in a less discriminatory manner may suffice.[239] Thus, plaintiffs may show either that the employer could have used the same algorithm in a less discriminatory way—for example, as one factor in hiring instead of as the entire hiring process—or that the employer could have removed certain biased factors from the algorithm that would have reduced its discriminatory effect.[240] Of course, as Barocas and Selbst argue, isolating the relevant factors from an algorithm will not be easy as a matter of proof.[241] Yet given efforts in the field of computer science to advance technical tools to reduce algorithmic bias, it may become possible to discover, adjust, and recreate an existing algorithm to reduce its disparate impact.[242] If so, this may be the best opportunity yet to reach algorithmic discrimination under existing disparate impact law: providing expert evidence that, by making adjustments to *an employer's own algorithm*, it is possible to create a less discriminatory alternative.

235. *See supra* Subpart I.A. (describing the different types of algorithms used at work).
236. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012).
237. *See* Barocas & Selbst, *supra* note 11, at 706, 709–10.
238. *Id.* at 718.
239. *Id.* at 710; *see, e.g.*, Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478, 504–05 (3d Cir.1999).
240. *See, e.g.*, *Lanning*, 181 F.3d at 504–05.
241. *See* Barocas & Selbst, *supra* note 11, at 709–10.
242. *See, e.g.*, Feldman et al., *supra* note 100.

Under existing disparate impact doctrine, courts can require more from an employer to satisfy its business necessity defense than mere statistical correlation, and courts can side with plaintiffs where an employer's own algorithm could be used in a less discriminatory way. Still, that a biased algorithm may be able to serve as its own justification for any disparate impact it creates calls into question the underlying neutrality of the algorithm, and calls for a closer examination of framing algorithmic discrimination as disparate treatment.

B. Title VII Disparate Treatment, Using Stereotype Theory

The second major legal framework for proving discrimination under Title VII is disparate treatment. Traditional disparate treatment reflects anticlassification principles of formal equal treatment. It was first articulated by the U.S. Supreme Court in the 1973 case *McDonnell Douglas Corp. v. Green*,[243] based on the statutory text of Title VII making it unlawful for an employer "to fail or refuse to hire . . . or otherwise to discriminate against any individual . . . [in] compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin."[244] In contrast to disparate impact's application to facially neutral practices, disparate treatment has been interpreted by courts to require proof of "intentional" discrimination and to apply when protected class was a "motivating factor" in an adverse employment action.[245]

Importantly, while disparate treatment requires "intentional" acts, Title VII's definition of discriminatory "intent" is broader than its terminology implies. A great deal of legal scholarship has addressed this issue, a full analysis of which is beyond the scope of this Article.[246] However, several key guidelines are clear. First, to be intentional, disparate treatment does not require protected class animus; in fact, it does not even require acting with conscious awareness that you are discriminating.[247] For example, if an employer adopts a practice of hiring candidates by subjective review of a

243. 411 U.S. 792 (1973).

244. 42 U.S.C. § 2000e-2(a)(1) (2012).

245. Desert Palace, Inc. v. Costa, 539 U.S. 90, 94 (2003); *McDonnell Douglas*, 411 U.S. at 802 (1973).

246. *See* Stephanie Bornstein, *Reckless Discrimination*, 105 CALIF. L. REV. 1055, 1077–83 (2017); *see also id.* at 1077 n.134 (describing the literature on the topic and noting that, as of August 2016, a "Westlaw search resulted in over 4,600 cases and over 2,100 journal and law review articles discussing intentional discrimination or intent to discriminate under Title VII, dating back to 1967").

247. *See id.* at 1077–83; Noah D. Zatz, *Managing the Macaw: Third-Party Harassers, Accommodation, and the Disaggregation of Discriminatory Intent*, 109 COLUM. L. REV. 1357, 1364 (2009); Michael J. Zimmer, *A Chain of Inferences Proving Discrimination*, 79 U. COLO. L. REV. 1243, 1248, 1289–94 (2008).

resume and interview and acts intentionally when applying that process to each individual applicant, applicants who believe they were evaluated worse and not hired because of their race can sue for disparate treatment, even if the employer did not believe the decision was racially motivated.[248] Second, intentional disparate treatment is virtually always proven by circumstantial evidence: defendants rarely admit that a protected characteristic entered into their decision-making process, nor must they.[249] Third, under what is known as the stereotype theory of liability, acting intentionally on the basis of a stereotype associated with a protected class constitutes disparate treatment on the basis of that protected class.[250]

While traditional disparate treatment reflects anticlassification principles, modern disparate treatment incorporates antistereotyping principles from *Price Waterhouse v. Hopkins* that require individuals to be treated individually, even within their protected class.[251] As cases brought under a stereotype theory of disparate treatment have held, an employer who judges an employee for work-related purposes on the basis of stereotypes associated with a protected class may commit intentional discrimination, even if the employer does not recognize its own bias or is operating on the basis of benevolent motives.[252] Indeed, in hundreds of cases over five decades, Title VII jurisprudence has demonstrated that intent is a broad concept that can be proven with a variety of circumstantial evidence, even despite an employer's stated intention or policy of nondiscrimination.[253]

Title VII disparate treatment claims can be brought on an individual or a class-wide basis. For individual claims, the plaintiff follows the burden-shifting framework established in *McDonnell Douglas*.[254] Plaintiff applicants or employees must show a prima facie case of disparate treatment that (1) they are a member of a protected class, (2) they were qualified for the position, (3) they experienced an adverse employment action, and (4) the adverse action occurred under circumstances giving rise to an inference of discrimination.[255] The prima facie burden is not a hard one to meet: the plaintiff can use any relevant evidence to meet step 4, including comparator

248. *See, e.g.*, *Desert Palace*, 539 U.S. at 92; Price Waterhouse v. Hopkins, 490 U.S. 228 (1989) (plurality opinion), *superseded in part by statute on other grounds*, Civil Rights Act of 1991 § 107(a), 42 U.S.C. § 2000e-2(m) (2012), *as recognized in* Burrage v. United States, 571 U.S. 204, 213 n.4 (2014); *McDonnell Douglas*, 411 U.S. at 802.

249. *See McDonnell Douglas*, 411 U.S. at 802; Bornstein, *supra* note 246, at 1077–83.

250. *See Price Waterhouse*, 490 U.S. at 250.

251. *See id.* at 250–52.

252. *See supra* Subpart II.B; *see, e.g.*, Chadwick v. WellPoint, Inc., 561 F.3d 38, 48 (1st Cir. 2009); Back v. Hastings on Hudson Union Free Sch. Dist., 365 F.3d 107, 121–22 (2d Cir. 2004).

253. *See* Bornstein, *supra* note 246, at 1077–83.

254. *See McDonnell Douglas*, 411 U.S. at 802.

255. *Id.*

evidence if available, statistical evidence if relevant,[256] and, if a case is brought under the stereotype theory, evidence that protected class stereotypes played a role in the decision.[257] Once the plaintiff has met the prima facie case, a burden of production shifts to the defendant employer to articulate a legitimate nondiscriminatory reason for the adverse action—a low bar to meet, as any nondiscriminatory reason will usually suffice.[258] Finally, the burden of persuasion shifts back to the plaintiff to show that the defendant's stated reason for the adverse employment action was a "pretext" and not the real reason for the decision, which was the plaintiff's protected class.[259]

For class-wide disparate treatment claims, known as "pattern or practice" claims, the proof structure is different. As established in 1977 in the U.S. Supreme Court cases *Hazelwood School District v. United States* and *International Brotherhood of Teamsters v. United States*, the plaintiff applicants or employees must show that an employer's regular employment practices are discriminatory using statistics and anecdotal evidence to create an inference of discrimination.[260] This statistical part of the prima facie case is met in the same way that plaintiffs prove disparate impact, described above, by showing a statistically significant disparity by protected class from expected results using the relevant labor market.[261] Yet plaintiffs must also usually provide other anecdotal evidence to support the statistics in order to infer discriminatory intent: for example, anecdotal evidence from individuals who experienced discrimination or other evidence that stereotypes were at play.[262] Once plaintiffs have met the prima facie case, the burden shifts to the employer to rebut the inference of discrimination by either challenging plaintiffs' statistics or offering legitimate reasons for the disparity.[263] As compared to the defendant's burden in responding to a disparate impact claim, the rebuttal burden on a defendant in a pattern-or-

256. *Id.* at 803–05.

257. Price Waterhouse v. Hopkins, 490 U.S. 228 (1989) (plurality opinion), *superseded in part by statute on other grounds*, Civil Rights Act of 1991 § 107(a), 42 U.S.C. § 2000e-2(m) (2012), *as recognized in* Burrage v. United States, 571 U.S. 204, 213 n.4 (2014).

258. *McDonnell Douglas*, 411 U.S. at 802.

259. *Id.* at 804; *see also* Tex. Dep't of Cmty. Affairs v. Burdine, 450 U.S. 246, 256 (1981). Plaintiffs can also allege disparate treatment under a "mixed motive" framework of disparate treatment, for which they need only show that protected class status was "a motivating factor" rather than the sole factor for the decision, but the employer has a complete defense to damages if they can show that they would have made the same decision anyway. 42 U.S.C. §§ 2000e-2(m), 2000e-5(g)(2)(B) (2012).

260. Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 307–09 (1977); Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 337–40 (1977).

261. *See Hazelwood Sch. Dist.*, 433 U.S. at 307; *Int'l Bhd. of Teamsters*, 431 U.S. at 339; *supra* Subpart III.A.

262. *See, e.g.*, Gay v. Waiters' & Dairy Lunchmen's Union, 694 F.2d 531, 552–53 (9th Cir. 1982).

263. *See Hazelwood Sch. Dist.*, 433 U.S. at 310; *Int'l Bhd. of Teamsters*, 431 U.S. at 340.

practice case is relatively lower; the ultimate burden of persuasion remains with plaintiffs, to prove that the employer's practices were discriminatory.[264]

With one exception, current scholarship views the harm of algorithmic discrimination as outside of the realm of disparate treatment and instead a matter for the disparate impact framework. Barocas and Selbst acknowledge that, in the unlikely event that an algorithm is used with the purpose of covering up intentional discrimination, this "masking" would be actionable disparate treatment.[265] This would occur if an employer were to include protected class membership as a variable and intentionally manipulate pieces of an algorithm to get the discriminatory result it desired.[266] Likewise, Kim suggests that only one fact pattern of algorithmic discrimination "easily fits within the conventional [disparate treatment] framework": "[w]hen an employer intends to discriminate but relies on an apparently neutral data model to justify its decisions."[267]

By creating scenarios that assume animus on the part of the decision maker, both examples understate the reach of "intent" in disparate treatment law. If an employer acts intentionally in a way that applies stereotypes associated with protected classes to individuals, and the result of that action is protected class discrimination, that may be actionable disparate treatment under a stereotype theory.[268] Indeed, Barocas and Selbst and Kim start from the assumption that the data on which an algorithm is based may be inherently discriminatory. If, as they suggest, "data are not neutral,"[269] and "data mining can reproduce existing patterns of discrimination [or] inherit the prejudice of prior decision makers,"[270] why should we treat algorithms as "facially neutral"? Under Title VII, "unintentional" does not mean the same thing as "facially neutral," and "intent" is its own animal.[271]

---

264. *See* EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 309 (7th Cir. 1988) (citing *Burdine*, 450 U.S. at 254) (explaining that "[defendant] had a rebuttal burden, but to meet that burden, [defendant] needed only to produce evidence that 'raise[d] a genuine issue of fact as to whether it discriminated'").

265. *See* Barocas & Selbst, *supra* note 11, at 692–93, 701 ("[A]side from rational racism and masking (with some difficulties), disparate treatment doctrine does not appear to do much to regulate discriminatory data mining.").

266. *Id.* at 692–93. However, they suggest that "litigation arising from it likely would be tried under a 'mixed-motive' framework, which asks whether the same action would have been taken without the intent to discriminate." *Id.* at 693 n.85.

267. *See* Kim, *Data-Driven*, *supra* note 13, at 903.

268. *See* Bornstein, *Unifying*, *supra* note 155, at 928; Bornstein, *supra* note 246, at 1083–85.

269. Kim, *Data-Driven*, *supra* note 13, at 860, 883 ("Data mining models are thus far from neutral. Choices are made at every step of the process—selecting the target variable, choosing the training data, labeling cases, determining which variables to include or exclude—and each of these choices may introduce bias along the lines of race, sex, or other protected characteristics.").

270. Barocas & Selbst, *supra* note 11, at 674.

271. *See* Bornstein, *supra* note 246, at 1083–85.

For this reason, at least some forms of algorithmic discrimination may be litigable as disparate treatment under the stereotype theory of Title VII.[272]

Despite their focus on a disparate impact approach, Barocas and Selbst lay the foundation for a stereotyping argument by identifying two types of human biases or stereotypes that may be introduced—in foundational ways—into algorithmic decision-making.[273] First, when selecting target variables and training data on which to train a machine-learning algorithm to find "good employees," programmers must define what constitutes a "good" employee.[274] If the algorithm is trained on data that itself incorporates subjective biases—for example, past performance evaluation scores or other subjective qualities selected by the employer—it will incorporate such bias into any correlations it makes.[275] Second, if the algorithm is designed to determine its own pattern of decision-making based on past biased decisions, it will reproduce such bias in future decisions.[276] Barocas and Selbst explain that "[a]utomating the process in this way would turn the conscious prejudice or implicit bias of individuals involved in previous decision making into a formalized rule that would systematically alter the prospects of all future applica[tions]"—for example, they suggest, rejecting all applicants from historically black colleges because the employer had done so consistently in the past.[277] Both of these patterns can be described, as Barocas and Selbst suggest, by the "adage in computer science: 'garbage in, garbage out,'" where the "garbage out" is discriminatory employment decisions.[278]

Although current scholarship fails to recognize it, an employer intentionally applying a system to individuals that starts from a "garbage in" position infected by protected class bias may constitute disparate treatment. In particular, the use of predictive matching by algorithm to find applicants that fit a model "good employee" may pose a problem of protected class

272. According to Kim, "[r]eliance on algorithms will typically be a facially neutral employment practice . . . [but d]ata models that do not explicitly categorize on the basis of race or other protected categories may nevertheless operate as 'built-in headwinds' for disadvantaged groups" and be challenged under disparate impact. Kim, *Data-Driven*, *supra* note 13, at 905. Just because a practice does not "explicitly categorize" by protected class does not make it neutral; many practices that do not do so can be challenged using disparate treatment, for example human subjective decision-making processes.

273. *See* Barocas & Selbst, *supra* note 11, at 677–87.

274. *Id.* at 679.

275. *Id.* at 679–80.

276. *Id.* at 680–84.

277. *Id.* at 682.

278. *Id.* at 683–84, 697 n.113 ("The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.").

stereotyping. If an employer creates a model employee based on past subjective decision-making that incorporates protected class stereotypes and then applies that model to each future applicant, seeking to hold each individual to the stereotype of "good employee," that may no longer be a "facially neutral" practice. Just because a computer formula is making the decisions instead of a human does not wash away prior bias and make its application "neutral." This is demonstrably different than being a "facially neutral" practice for which a disparate impact analysis is appropriate.[279] The employer is not saying, for example, "We want people who score eighty percent on this test"; instead the employer is saying, "We want people who match this type of person." It is, in effect, like a hiring manager looking for applicants who "fit in" or looking for someone who is "management material." It is subjective decision-making by formula.[280]

Moreover, because AI is meant to mirror the human brain, subjective decision-making by a computer programmed by a human or modeled on past human decisions should be treated the same, for liability purposes, as subjective decision-making by a human. The employer who relies on its discriminatory results should be held liable to a similar extent under the law. Barocas and Selbst seemingly entertain, and reject, this possibility; but their analysis is based on an overly narrow view of the definition of intent in Title VII, presuming that "discriminatory data mining is by stipulation unintentional."[281] While they are correct that "the law does not adequately address unconscious disparate treatment,"[282] they use this statement to prove too much. And while "the doctrine [of Title VII] focuses on human decision makers,"[283] a human decision maker's application of a biased model may be enough.[284]

---

279. *See* Kim, *Data-Driven*, *supra* note 13, at 906–07 (discussing the difference between algorithmic discrimination and traditional testing for which disparate impact was designed); *see also* Bodie et al., *supra* note 20, at 1027–28 (suggesting that, if Uber's customer performance rating system is vulnerable to bias, aggregating the data in an application makes the data look neutral, but it is not).

280. *Cf.* Bodie et al., *supra* note 20, at 1017–18 ("[P]redicting future behavior based on characteristics of people who behaved in desirable or undesirable ways in the past . . . [or p]rofiling contains risks, in large part because classification and division is literally discrimination. Its purpose is to allow judgments to be made based on someone's membership in a group rather than based on their own individual merits. In fact, profiling can create new stereotypes on which people are judged.").

281. *See* Barocas & Selbst, *supra* note 11, at 698.

282. *See id.* at 698 (citing Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 9 (2006); Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM. & MARY L. REV. 911, 1000 (2005)).

283. *Id.* at 699 (emphasis omitted).

284. *See* Bornstein, *supra* note 249, at 1083–85. *But see* Barocas & Selbst, *supra* note 11, at 700 (suggesting that, "to be found liable under current doctrine, the employer would likely both have to know that this is the specific failure mechanism of the model and choose it based on this fact"); Sullivan, *supra* note 13, at 8 (arguing that a computer making decisions using AI "isn't human, so it can't 'intend' to discriminate," as required for disparate treatment liability).

This is, by no means, an easy or clear harm to identify. Algorithmic decision-making that incorporates stereotypes may fall into a gray area between what looks like disparate treatment and what looks like disparate impact.[285] Nevertheless, unearthing the potential for unlawful stereotyping is essential to ensure bias-free algorithms. For a stereotyping claim, it is not enough that an employer selects candidates using an objective factor that merely correlates with protected class. For example, if an employer has a stated preference for hiring military veterans, just because demographically more men may be veterans does not mean that a woman could sue for unequal treatment or stereotyping, as the U.S. Supreme Court held in *Personnel Administrator of Massachusetts v. Feeney*.[286] A female applicant disadvantaged by this identifiable and objective data point—you either are a veteran or you are not—could only allege disparate impact.[287]

But imagine that the preference is not an objective data point—for example, the employer prefers to hire employees it believes have "leadership abilities" or will be "aggressive." This is no longer a purely objective factor; determining whether a candidate has these qualities requires a subjective assessment of who is or is not a leader or aggressive. If a human decision maker looks at a female candidate's resume and assumes that, because it shows her to have a traditionally feminine background, she lacks these qualities so the decision maker does not hire her, the applicant could now allege gender stereotyping—even if the decision maker was not aware of its own biases. If an algorithm is trained to look for "aggressive leaders" by matching individuals to the group data of its successful employees with traditionally masculine backgrounds, and it assumes future behavior based on these past masculine cues, excluding those who do not match, the algorithm is now doing the stereotyping.

Under this formulation, an individual who believes they experienced discrimination when compared against a biased algorithmic model could sue for individual disparate treatment under a stereotype theory. Like Ann Hopkins, the plaintiff in *Price Waterhouse*, individuals judged against a model that incorporates and reproduces disadvantage by race or sex by defining "good employee" could argue that they were rejected based on a failure to match a biased stereotype regardless of their ability to do the

285. *Cf.* Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 773–82 (2006) (discussing the limiting impact disparate impact doctrine has had on disparate treatment theory and concepts of intent).

286. 442 U.S. 256, 276–78 (1979).

287. *See id.*

job.[288] To pursue this claim, plaintiffs would apply the same framework as would a plaintiff alleging Barocas and Selbst's or Kim's masking approach, but would prove stereotyping instead of animus to establish intent. Plaintiffs would argue that they were members of a protected class, met the minimum qualifications for the position, were not hired, and that a suspect algorithm is to blame, which they would show using statistics or factors in the algorithm that incorporated stereotyping. When the employer raises the algorithm itself as the legitimate nondiscriminatory reason, if plaintiffs can prove that the algorithm is biased, they will have proven pretext.

Plaintiffs could also, or instead, allege a class claim of pattern or practice disparate treatment using a stereotype approach. Like the plaintiffs in *Teamsters* and *Hazelwood*, plaintiffs who can show a statistically significant disparity in hiring by protected class from the relevant labor market could argue that an employer's algorithmic hiring constituted a pattern or practice of disparate treatment.[289] To pursue this claim, plaintiffs would create a prima facie case the same way as would plaintiffs alleging disparate impact. The rebuttal burden would then shift to the employer, who could either rebut the statistics or provide a legitimate nondiscriminatory reason for the disparity. Again, however, an algorithm shown to be infected with bias will not be considered "nondiscriminatory." While the challenge for plaintiffs is that the ultimate burden of persuasion remains with them to overcome defendants' argument, the advantage when compared to disparate impact is that there is no opportunity for the employer to raise a "business necessity" defense. Either the plaintiff proves the algorithm is unfairly biased or the employer demonstrates that it is not; validation studies do not suffice in pattern-or-practice disparate treatment claims.[290]

Of course, establishing the evidence necessary to prove either individual or pattern-or-practice disparate treatment liability poses significant challenges to plaintiffs that cannot be overstated. The challenges that current scholars identify in accessing the necessary proof in the context of disparate impact apply similarly to the context of disparate treatment.[291] Plaintiffs will likely need access to complex algorithms that may be inaccessible or from which individual factors may not be parsed, particularly if they in-

288. *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 235 (1989) (plurality opinion), *superseded in part by statute on other grounds*, Civil Rights Act of 1991 § 107(a), 42 U.S.C. § 2000e-2(m) (2012), *as recognized in* Burrage v. United States, 571 U.S. 204, 213 n.4 (2014).

289. *See* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 307–08 (1977); Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 339 (1977).

290. *See Hazelwood Sch. Dist.*, 433 U.S. at 310; *Int'l Bhd. of Teamsters*, 431 U.S. at 358; EEOC v. Sears, Roebuck & Co., 839 F.2d 302, 342 (7th Cir. 1988).

291. *See* Barocas & Selbst, *supra* note 11, at 701–13; Kim, *Data-Driven*, *supra* note 13, at 901–09.

volve "black box" machine learning.[292] Nevertheless, practical matters of proof should not determine whether a theory of litigation can apply to a particular fact pattern; such matters may determine whether plaintiffs will prevail under a given theory, but that is a separate matter. To the extent that Barocas and Selbst and Kim identify challenges in proving a case of "masking" but acknowledge that masking is a matter of disparate treatment conceptually,[293] the same can be said of disparate treatment under a stereotype theory. And to the extent that defining a relevant labor market or specific factors within an algorithm that could be less discriminatory poses challenges in cases that Barocas and Selbst and Kim identify as disparate impact cases,[294] if a case is properly framed as a pattern-or-practice disparate treatment case, similar challenges in proving the case should not change that frame.

When the use of predictive matching algorithms has a discriminatory result due to biases built into the data model, framing the problem as one of disparate treatment rather than disparate impact is important for several reasons. First, as a matter of law it is wrong to excuse discriminatory algorithms as "facially neutral" when they are more like subjective decision-making by computer. While it is possible to litigate the discriminatory results of subjective decision-making practices as disparate impact,[295] it is theoretically more accurate to do so as disparate treatment.[296] The fact that a flawed algorithm can be its own proof to satisfy the business necessity defense illustrates the problem of starting from a position of "neutrality." Second, so properly framed, employers facing liability under current law—or at least the specter of having to rebut a flawed algorithm as nondiscriminatory—will be more motivated to correct problematic algorithms, or, as discussed in Part II, to use algorithms differently, in ways that suppress rather than reinforce stereotypes. A disparate treatment lawsuit carries with it a stronger message of discriminatory culpability and the prospect of greater damages for plaintiffs[297]—both of which provide stronger incentives to employers to take preventative action. Lastly, identifying that, in

292. *See* Barocas & Selbst, *supra* note 11, at 701–13.

293. *See id.* at 696; Kim, *Data-Driven*, *supra* note 13, at 903–04.

294. *See* Barocas & Selbst, *supra* note 11, at 701–13; Kim, *Data-Driven*, *supra* note 13, at 901–09.

295. *See* Watson v. Fort Worth Bank & Tr., 487 U.S. 977, 989–93 (1988).

296. *See* Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1219, 1229–37 (1995) ("The disparate impact paradigm . . . is an inappropriate analytical tool for addressing the intergroup biases inherent in subjective decisionmaking. . . . From a phenomenological standpoint, subjective practices discrimination is a disparate treatment problem, not a disparate impact problem, and it requires a disparate treatment solution." (emphasis omitted)).

297. Compensatory and punitive damages are only available for "intentional" discrimination; thus, they are not available in disparate impact lawsuits. *See* 42 U.S.C. § 2000e-5(g)(1) (2012).

some cases, algorithmic discrimination may constitute disparate treatment under a stereotype theory is essential for those seeking to challenge algorithmic discrimination outside of the employment context, where a disparate impact theory of liability may not be available.

C. An Antistereotyping Approach to Algorithmic Discrimination Beyond the Workplace

Computer algorithms are now being used to make decisions in all areas of life. If, as current scholarship has documented, algorithms may incorporate and even magnify the biases inherent in the data on which they rely, the potential for algorithmic discrimination exists well beyond the workplace. Yet even if an algorithm can be proven to have disparate effects by protected class, not all antidiscrimination law recognizes a legal claim for disparate impact. Importantly, unless a relevant statute applies,[298] discrimination challenges brought against state or federal governments' uses of algorithms—for example, in decisions about criminal justice, government benefits, or even tax auditing—will be brought under the constitutional law of Equal Protection, which does not allow a plaintiff to sue for disparate impact.[299] For this reason, considering an antistereotyping approach to algorithmic discrimination is essential.

As described previously, the antistereotyping principle in antidiscrimination law grew directly out of Equal Protection jurisprudence: "the basic principle that the Fifth and Fourteenth Amendments to the Constitution protect persons, not groups," and that "all governmental action based on . . . a group classification . . . should be subjected to . . . judicial inquiry to ensure that the personal right to equal protection of the laws has not been infringed."[300] A commitment to rooting out protected class stereotypes and to striking down state action that reinforces them is as important to Equal Protection as it is to Title VII.

298. For example, in the context of government action in employment or housing, disparate impact is available by statute under Title VII (employment) or the Fair Housing Act, 42 U.S.C. §§ 3601–3619 (2012) (housing). *See* Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, 135 S. Ct. 2507, 2525 (2015) (holding that disparate impact claims are cognizable under the Fair Housing Act); Griggs v. Duke Power Co., 401 U.S. 424 (1971) (holding, for the first time, that disparate impact claims are cognizable under Title VII).

299. *See* Washington v. Davis, 426 U.S. 229 (1978). Note that constitutional challenges to governmental uses of algorithms may also be brought on other grounds, including under the Due Process Clause. This is beyond the scope of the Article's focus on protected class discrimination.

300. Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 227 (1995) (emphasis omitted).

While a full analysis of the application of constitutional law to algorithmic discrimination is beyond the scope of this Article,[301] a brief explanation of the law and one example serve to illustrate the importance of stereotype theory in such a situation. Under U.S. Supreme Court precedent, the Equal Protection Clause prohibits disparate treatment by protected class but does not provide redress for disparate impact. In *Washington v. Davis*[302]—a case about government hiring practices, but which was brought under the Equal Protection Clause, not Title VII[303]—the Court rejected the proposition that "a law, neutral on its face and serving ends otherwise within the power of government" discriminates "simply because it may affect a greater proportion of one race than of another."[304] While "[d]isproportionate impact is not irrelevant," the Court held, "it is not the sole touchstone of an invidious racial discrimination forbidden by the Constitution" and, "[s]tanding alone," it is not actionable.[305] To challenge algorithmic discrimination under Equal Protection, then, a government use of algorithms would have to constitute unequal treatment based on protected class.

One prominent example of a governmental use of algorithms that could be subject to an Equal Protection challenge is algorithmic risk assessment, used in the criminal justice context for things like predictive policing or sentencing and parole decisions.[306] As with any algorithm, the specific variables incorporated and the way the algorithm is used will determine its legality. But, some scholars[307]—and at least one court[308]—have suggested

301. A separate and growing body of scholarship has begun to address the constitutionality of algorithms in the context of criminal justice under the Equal Protection and Due Process Clauses and the Fourth Amendment. *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017); Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327 (2015); Andrew Guthrie Ferguson, *Predictive Prosecution*, 51 WAKE FOREST L. REV. 705 (2016); Melissa Hamilton, *Risk-Needs Assessment: Constitutional and Ethical Challenges*, 52 AM. CRIM. L. REV. 231 (2015); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. (forthcoming 2019); John Lightbourne, *Damned Lies & Criminal Sentencing Using Evidence-Based Tools*, 15 DUKE L. & TECH. REV. 327 (2017); Selbst, *supra* note 101; Dawinder S. Sidhu, *Moneyball Sentencing*, 56 B.C. L. REV. 671 (2015); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014); Sandra G. Mayson, *Bias in, Bias Out* (July 20, 2018) (draft on file with the author).

302. 426 U.S. 229 (1978).

303. At the time this case was filed in 1970, Title VII did not apply to the public sector; thus, the plaintiffs alleged discrimination under the Equal Protection Clause of the Fifth Amendment (applying to the D.C. government). In a 1972 amendment, Title VII was extended to cover government employers. Equal Employment Opportunity Act of 1972, Pub. L. No. 92-261, § 11, 86 Stat. 103, 111–13 (1972) (codified as amended at 42 U.S.C. § 2000e-16 (2012)).

304. *Washington*, 426 U.S. at 242.

305. *Id.*

306. *See, e.g.*, Lightbourne, *supra* note 301, at 337–42; Starr, *supra* note 301, at 821–41.

307. *See, e.g.*, Lightbourne, *supra* note 301, at 337–42; Starr, *supra* note 301, at 821–41.

that explicitly including protected class status, such as gender, in an algorithmic criminal risk assessment could give rise to a disparate treatment challenge under the Equal Protection Clause.

Once explicit consideration of a protected class is removed, however, an antistereotyping approach may be a plaintiff's only option if a predictive risk assessment algorithm still has discriminatory results. As in the employment context, individual plaintiffs in the predictive policing context could argue that they are being penalized on the basis of stereotypes associated with protected class groups. Of course, as in employment, a measurable objective factor that correlates with protected class may give rise to only disparate impact, not disparate treatment harms. For example, if the choice to police a certain neighborhood known to have a higher crime rate disproportionately affects African Americans, that is a disparate impact problem—a person arrested there is either inside the neighborhood or not. However, where the choice of who to arrest involves a subjective assessment—for example, an assessment of who is more likely to affiliate with gang members[309]—stereotyping may be at play. Much as there is a fine line between applying a stated preference for veterans (disparate impact) and assuming from group data that an individual is or is not likely to be a "leader" or "aggressive" (possible stereotyping), there is a fine line between applying actual data on incidents of crime (disparate impact) and assuming from group data that an individual is or is not likely to commit a crime (possible stereotyping).[310] The same argument could apply: algorithms based on group data that may incorporate human bias or past discrimination are not "neutral" just because they are made by a computer.

That said, just because an algorithm relies on protected class stereotypes—or even includes explicit consideration of a protected class—does not mean that its use will be ruled unconstitutional. A predictive risk assessment algorithm itself, or its use in a policing, sentencing, or parole decision, may likely survive a constitutional challenge depending on the requisite level of scrutiny applied—strict scrutiny if race or national origin ("narrowly tailored" to further a "compelling" government interest)[311] or intermediate scrutiny if gender ("substantially related" to an "important" government interest).[312] Indeed, even the court that considered whether ex-

308. *See* State v. Loomis, 881 N.W.2d 749, 766–67 (Wis. 2016) (discussing the issue but analyzing it under the Due Process Clause because the defendant had not raised an Equal Protection claim).

309. For a discussion of what Andrew Selbst describes as "person-based" (as opposed to "place-based") predictive policing, see Selbst, *supra* note 101, at 137–40.

310. *See supra* Subpart III.B (comparing preference for military veterans and preference for "aggressiveness" or "leadership" assumptions).

311. *Id.*

312. Craig v. Boren, 429 U.S. 190, 197 (1976).

plicitly including sex as a factor in an algorithmic criminal risk assessment *could* raise constitutional concerns held that it did not do so in the facts before it, where the risk assessment was used as only one factor in the plaintiff's parole decision.[313]

Moreover, as with algorithmic decision-making in the employment context, there is reason to believe that using algorithms in criminal risk assessment may actually *better* serve equality goals relative to current criminal justice practices by improving upon biased subjective human decision-making.[314] But where such algorithms do indeed result in discrimination, recognizing the potential for stereotype theory to give rise to a disparate treatment challenge under the Equal Protection Clause may influence governmental agencies to attempt to correct potential algorithmic biases, based not only on explicit protected classes but on associated stereotypes, too.

CONCLUSION

The rise of algorithms and AI offer a great deal of promise to help make better decisions faster by adding objectivity with data and correcting for human biases. But any improvement to traditional decision-making that relies on data will depend on what data is being used and how. Handled properly, algorithms can suppress, interrupt, or remove protected class stereotypes from decisions; handled improperly, they become a form of stereotyping, making assumptions about the future behavior of individuals by judging them against composite data on the past behavior of a group.

As current scholarship on algorithmic discrimination in the workplace has convincingly demonstrated, while algorithms may suppress human biases, they may reproduce and even exacerbate structural biases. Proposed solutions to date focus on requiring those using algorithmic decision-making to document and mitigate their own biases or on reinterpreting antidiscrimination law to meet the new challenge of algorithmic discrimination. Both proposals would make significant improvements to reducing and

313. *See Loomis*, 881 N.W. 2d at 766–67. Note, however, that the plaintiff in this case did not raise the challenge under the Equal Protection Clause, so the Wisconsin Supreme Court considered and rejected his gender discrimination claim under the Due Process Clause. *Id.*

314. *See, e.g.*, Ferguson, *Big Data and Predictive Reasonable Suspicion*, *supra* note 301, at 389–90 (noting that, while big data policing has downsides, human policing judgments "include explicit and implicit biases…and the frailties of human perception," and "racial stereotypes can influence suspicion" so that "[r]eplacing those generalized intuitions" with real data "should result in a more accurate policing strategy"). *Cf.* Selbst, *supra* note 101, at 115–16 ("Like other sectors' use of data mining, predictive policing is sold in part as a way to counteract the conscious or unconscious prejudices of human decision-makers—in this case the police. And it has the potential to do so. But . . . express consideration of race is not necessary for data mining to have a disproportionate racial impact."); Starr, *supra* note 301, at 850–55 (noting that, while advocates of data-based predictive tools for sentencing argue they are "superior to available alternatives," to say "that actuarial prediction outperforms clinical prediction is . . . a generalization that is not true in every case").

redressing algorithmic discrimination at work if adopted. Yet both also start from the assumption that algorithms are "facially neutral"—even when those algorithms incorporate human biases and stereotypes or prior past discrimination. This assumption unnecessarily hampers the ability of current antidiscrimination law to address an issue that is not just a problem of disparate impact.

Antidiscrimination law requires that people be treated equally, but it also requires that they be treated individually and not be judged against stereotypes associated with protected classes. Applying an antistereotyping lens to the problem of algorithmic discrimination suggests that, when individuals are judged against a model of group data that incorporates protected class stereotypes and then rejected for failing to conform to those stereotypes, that may constitute intentional stereotyping. Having a computer execute what amounts to composite subjective decision-making does not make otherwise biased action "facially neutral." If AI is meant to model human decision-making and, in fact, incorporates human biases and stereotypes, an employer should be held accountable for the discrimination it creates to the same extent that the employer would be if it relied on a biased human decision maker.

The suggestion that algorithmic discrimination can be considered intentional discrimination will likely be subject to criticism. The most obvious counterargument is that algorithms just combine objective data points about candidates and then compare individuals to make hiring decisions. Indeed, if the data is truly objective, this is a valid criticism. For this reason, testing-and-tracking algorithms that measure the candidate's performance on job-related tasks or that blind decision-makers to potential sources of bias may not be implicated. If the data is truly objective, then there has been no stereotyping and no such theory of liability would apply. But where an algorithm includes subjective data, like prior performance evaluations or past hiring decisions, or seeks to match candidates to a model employee that incorporates protected-class-related stereotypes, it is not neutral. If the algorithm discriminates, employment decisions that rely upon it should be challenged accordingly.

A second counterargument is that, even if you agree that certain types of algorithmic decision-making *should* be treated like human subjective decision-making, reaching biased human subjective decision-making poses challenges of its own under current Title VII doctrine. In the U.S. Supreme Court decision *Wal-Mart Stores, Inc. v. Dukes*, for example, the Court refused to find that widespread subjective decision-making, without more, could constitute a common question to support a pattern-or-practice claim

of sex discrimination.[315] Thus, even if a machine-learning algorithm could be treated like a human brain, in the context of subjective decision-making that results in discrimination, it is not easy to establish disparate treatment liability for decisions made by a human brain either. Any challenge to algorithmic employment discrimination under existing law will not be easy, particularly given that the challenges of producing proof to satisfy a plaintiff's burdens in any Title VII case are made all the more difficult by the "black box" of predictive algorithms and AI. But, if a predictive matching algorithm makes subjective assessments of candidates that result in discrimination, it may, ironically, be *easier* to expose human cognitive biases and stereotypes when they are incorporated into an algorithm than when they are held inside the brain of a human.[316] And the algorithm itself would be the uniform way in which human subjectivity was exercised—the "glue" that the *Wal-Mart* majority said was lacking to support a pattern-or-practice disparate treatment claim of sex discrimination in that case.[317]

Workforce analytics have become a multibillion-dollar industry, which is likely to continue to grow in the future.[318] Regulating to require employers to document their algorithmic choices before they use them is an important, direct solution. Strengthening existing law to help remedy algorithmic disparate impacts after they occur would, no doubt, help reduce and redress its occurrence. Applying an antistereotyping approach under existing law offers additional incentives and guidance. Framing algorithmic discrimination as actionable disparate treatment under a stereotype theory may raise the specter of potential liability under current law enough to motivate the costly and complex efforts that may be needed to rid certain algorithms of bias. Moreover, if the stated goal of data-based tools for hiring is to reduce discrimination and increase diversity, incorporating an antistereotyping approach can be of help. The technology has the potential to improve upon human decision-making by suppressing or removing human biases. It is a mistake, then, to excuse algorithms that incorporate human stereotypes and structural discrimination as "neutral." Considering an antistereotyping approach can help discourage the use of algorithms in ways that exacerbate bias and instead help unleash their antidiscriminatory potential.

315. Wal-Mart Stores, Inc. v. Dukes, 564 U.S. 338 (2011).

316. *See* Kroll et al., *supra* note 13, at 634 ("The . . . biases of human decisionmakers can be difficult to find and root out, but we can peer into the 'brain' of an algorithm: computational processes and purpose specifications can be declared . . . and verified . . . ." (emphasis omitted)).

317. *Wal-Mart Stores, Inc.*, 564 U.S. at 352.

318. *See* Bersin, *supra* note 55.