Florida State University Law Review

Volume 47 | Issue 3 Article 3

Spring 2020

Stretching Human Laws to Apply to Machines: The Dangers of a "Colorblind" Computer

Zach Harned

Hanna Wallach

STRETCHING HUMAN LAWS TO APPLY TO MACHINES: THE DANGERS OF A "COLORBLIND" COMPUTER

ZACH HARNED* & HANNA WALLACH**

ABSTRACT

Automated decision-making has become widespread in recent years, largely due to advances in machine learning. As a result of this trend, machine learning systems are increasingly used to make decisions in high-stakes domains, such as employment or university admissions. The weightiness of these decisions has prompted the realization that, like humans, machines must also comply with the law. But human decision-making processes are quite different from automated decision-making processes, which creates a mismatch between laws and the decision makers to which they are intended to apply. In turn, this mismatch can lead to counterproductive outcomes.

We take antidiscrimination laws in employment as a case study, with a particular focus on Title VII of the Civil Rights Act of 1964. A common strategy for mitigating bias in employment decisions is to "blind" human decision makers to the sensitive attributes of the applicants, such as race. The same strategy can also be used in an automated decision-making context by blinding the machine learning system to the race of the applicants (strategy 1). This strategy seems to comply with Title VII, but it does not necessarily mitigate bias because machine learning systems are adroit at using proxies for race if available. An alternative strategy is to not blind the system to race (strategy 2), thereby allowing it to use this information to mitigate bias. However, although preferable from a machine learning perspective, this strategy appears to violate Title VII.

We contend that this conflict between strategies 1 and 2 highlights a broader legal and policy challenge, namely, that laws designed to regulate human behavior may not be appropriate when stretched to apply to machines. Indeed, they may even be detrimental to the very people that they were designed to protect. Although scholars have explored legal arguments in an attempt to press strategy 2 into compliance with Title VII, we believe there lies a middle ground between strategies 1 and 2 that involves partial blinding—that is, blinding the system to race only during deployment and not during training (strategy 3). We present strategy 3 as a "Goldilocks" solution for discrimination in employment decisions (as well as other domains), because it allows for the mitigation of bias while still complying with Title VII. Ultimately, any solution to the general problem of stretching human laws to apply to ma-

* Founder of Stanford Artificial Intelligence & Law Society, Stanford Law School.

** Senior Principal Researcher, Microsoft.

chines must be sociotechnical in nature, drawing on work in both machine learning and the law. This is borne out in strategy 3, which involves innovative work in machine learning (viz. the development of disparate learning processes) and creative legal analysis (viz. analogizing strategy 3 to legally accepted auditing procedures).

I. INTRODUCTION: GREAT MINDS MAY NOT THINK ALIKE

"[A]n A.I. system must be subject to the full gamut of laws that apply to its human operator."
Oren Etzioni[1]

1. Oren Etzioni, *How to Regulate Artificial Intelligence*, N.Y. TIMES (Sept. 1, 2017), https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html.

"[The A.I. chess program] doesn't play like a human, and it doesn't play like a program. . . . It plays in a third, almost alien, way."
Demis Hassabis[2]

In 2012, Xerox was looking to increase retention of its employees, so it contracted with a third-party vendor, Evolv Incorporated.[3] Evolv Inc. specialized in developing machine learning systems that made recommendations about which applicants to hire. For example, Xerox could specify desirable factors like education or experience, and these would be considered in Evolv Inc.'s recommendation system. To ensure that the system would not make discriminatory recommendations, it would not be provided with the race of the applicants so that it would be unable to exhibit any bias. The problem with this strategy, though, is that there are many proxies for race. For example, one of the system's most predictive features for the length of a prospective employee's tenure was how far her home was from the employer's office. But an applicant's zip code is often highly correlated with her race. Xerox quickly realized that Evolv Inc.'s system was therefore able to make discriminatory recommendations on the basis of race even though it was not explicitly provided with the race of the applicants. As a result, information about an applicant's distance from her home to the office had to be withheld from the system as well.

With recent advances in machine learning, humans are not the only ones making high-stakes decisions anymore. Now that machine learning systems are increasingly used to make decisions in domains such as employment or university admissions, many have called for these systems to be held accountable to the same laws as their human counterparts. Oren Etzioni succinctly summarizes this perspective in an opinion piece in the New York Times:

> First, an A.I. system must be subject to the full gamut of laws that apply to its human operator. This rule would cover private, corporate and government systems. We don't want A.I. to engage in cyberbullying, stock manipulation or terrorist threats; we don't want the F.B.I. to release A.I. systems that entrap people into committing crimes. We don't want autonomous vehicles that drive through red lights, or worse, A.I. weapons that violate international treaties.[4]

This is a seemingly reasonable request. But it fails to take into account a small but crucial detail: these laws were designed with human decision makers in mind, yet automated decision-making processes of-

2. Will Knight, *Alpha Zero's "Alien" Chess Shows the Power, and the Peculiarity, of AI*, MIT TECH. REV. (Dec. 8, 2017), https://www.technologyreview.com/s/609736/alpha-zeros-alien-chess-shows-the-power-and-the-peculiarity-of-ai/.

3. Joseph Walker, *Meet the New Boss: Big Data*, WALL ST. J. (Sept. 20, 2012), https://www.wsj.com/articles/SB10000872396390443890304578006252019616768.

4. Etzioni, *supra* note 1.

ten differ starkly from human decision-making processes. For example, consider DeepMind's sophisticated AI-based game-playing program Alpha Zero:

> The [Alpha Zero chess] program often made moves that would seem unthinkable to a human chess player. 'It doesn't play like a human, and it doesn't play like a program,' Hassabis said at the Neural Information Processing Systems (NIPS) conference in Long Beach. 'It plays in a third, almost alien, way.' Besides showing how brilliant machine-learning programs can be at a specific task, this shows that artificial intelligence can be quite different from the human kind. As AI becomes more commonplace, we might need to be conscious of such 'alien' behavior.[5]

Given the inherent differences between these decision-making processes, some strategies for complying with the law will work better for humans than for machines. These differences were not so important when computers were seldom involved in high-stakes decisions. But with today's prevalence of automated decision-making, these differences must now be considered. As we cede more decision-making authority to machine learning systems, we need to carefully decide what strategies they should implement to comply with the law; simply implementing strategies intended for human decision makers may have undesirable consequences.

II. PROTECTIONS AGAINST BIAS

"A computer will do what you tell it to do, but that may be much different from what you had in mind."
Joseph Weizenbaum

A. Implicit and Explicit Bias

Before delving into legal protections against discrimination, we first explain why bias is so difficult to protect against. Sadly, the human condition is one afflicted with bias, often centering around sensitive attributes such as race and gender. Stemming from evolutionary and societal pressures, bias and its manifestations have been well documented by cognitive and social psychologists.[6] Humans are routinely tasked with making decisions in high-stakes domains, such as employment or university admissions. When bias affects these decisions, it negatively impacts not only the applicants whose fates are being decided, but also the composition of organizations as a whole. Consider the literature showing that gender diversity in the boardroom is beneficial for the female board members who might otherwise

5. Knight, *supra* note 2.

6. For readable surveys of the literature, *see* JENNIFER L. EBERHARDT, BIASED: UNCOVERING THE HIDDEN PREJUDICE THAT SHAPES WHAT WE SEE, THINK, AND DO (2019); DANIEL KAHNEMAN, THINKING, FAST AND SLOW (2013).

have been prevented from serving due to bias, but also creates a more effective board that, in turn, better serves shareholders.[7]

Bias is particularly pernicious because it is often implicit. Explicit bias, such as overt racism, is typically easier to detect than implicit bias and thus can be easier to eliminate. For example, a human resources manager who is outspokenly racist can be readily identified, thereby preventing her bias from impacting applicant selection. However, when bias is instead implicit, it is more difficult to detect, and the person exhibiting bias is often unaware that she is doing so. Indeed, social psychologists have demonstrated that people making hiring or admissions decisions unwittingly inflate the importance of whatever qualification happens to favor their preferred applicant.[8] In other words, a decisionmaker may (unconsciously) engage in differential elevation of certain qualifications in order to rationalize her unwitting discrimination.

This subtlety is what makes implicit bias so challenging to confront and to protect against. An implicitly biased decision maker is not even aware of her bias and, if queried, may simply cite the qualification she deemed most important in making her decision. This makes her particularly intransigent because she genuinely believes her decision is rooted in an objective metric, like grade point average. The problem is that she does not recognize that she varies the importance of the qualification depending on the race of the applicant under consideration. Worse still, when people are told that they will be held accountable for their decisions, this effect is potentiated, manifesting in increased rates of biased selection, as well as stronger recollections of details that favor their preferred applicants.[9] Although some studies have found that this effect can be mitigated by requiring a decision maker to commit in advance to the relative importance of different qualifications,[10] other studies have failed to replicate this finding.[11]

Additionally, the factors that influence bias can be quite subtle. One study showed that simple linguistic framing (e.g., "girls are as good at math as boys" vs. "boys are as good at math as girls") can perpetuate

7. *See* Kevin Campbell & Antonio Mínguez-Vera, *Gender Diversity in the Boardroom and Firm Financial Performance*, 83 J. BUS. ETHICS 435, 436 (2008).

8. Gordon Hodson, John F. Dovidio & Samuel L. Gaertner, *Processes in Racial Discrimination: Differential Weighting of Conflicting Information*, 28 PERSONALITY & SOC. PSYCHOL. BULL. 460, 461-62 (2002); Eric Luis Uhlmann & Geoffrey L. Cohen, *Constructed Criteria: Redefining Merit to Justify Discrimination*, 16 PSYCHOL. SCI. 474, 474 (2005).

9. *See* Michael I. Norton, Joseph A. Vandello & John M. Darley, *Casuistry and Social Category Bias*, 87 J. PERSONALITY & SOC. PSYCHOL. 817, 817 (2004).

10. Uhlmann & Cohen, *supra* note 8, at 478.

11. Norton, Vandello & Darley, *supra* note 9, at 820.

stereotypes.[12] Another study observed that cultural differences in emotion expression (e.g., desire for a job being expressed as excitement in European American applicants vs. calmness in Hong Kong Chinese applicants) can lead to biased hiring decisions.[13]

Despite its subtlety, implicit bias itself is nothing new, and neither is the notion that legal safeguards must be put in place in order to protect people from discrimination. We have had laws and mechanisms prohibiting discrimination since before the spate of social psychology research papers exploring the existence and nature of implicit bias. However, this research does emphasize that people—even when well intentioned—cannot be left as the sole arbiters of high-stakes decisions, given the human tendency toward implicit bias. Therefore, we require some form of sophisticated procedural oversight to ensure that bias does not negatively impact decisions. We now turn to the current legal regime that aims to provide such protection.

B. Constitutional Protections

Historically, legal mechanisms have played an active role in efforts to combat both explicit and implicit bias in the United States. For example, the Fourteenth Amendment of the U.S. Constitution contains the Equal Protection Clause, guaranteeing that no state shall deny an individual "the equal protection of the laws."[14] The Fourteenth Amendment also contains the Due Process Clause, which has been subsequently interpreted to cover both procedural due process and substantive due process.[15] The Due Process Clause prominently interdicts explicit bias on the part of state actors by requiring that they adhere to fair procedures before an individual may be deprived of life, liberty, or property.[16] This minimum requisite applied to all is intended to promote fairness and to combat discrimination. Perhaps even more directly related to explicit bias is substantive due process, which provides protection to vulnerable groups when a fundamental right is at stake, including the rights of "discrete and insular minorities."[17]

When a court must determine the constitutionality of a law, it typically applies rational basis review, under which there only need exist a hypothetical justification relating the law to a legitimate government

12. *See* Eleanor K. Chestnut & Ellen M. Markman, *"Girls Are as Good as Boys at Math" Implies That Boys Are Probably Better: A Study of Expressions of Gender Equality*, 42 COGNITIVE SCI. 2229 (2018).

13. Lucy Zhang Bencharit et al., *Should Job Applicants Be Excited or Calm? The Role of Culture and Ideal Affect in Employment Settings*, 19 EMOTION 377, 398 (2019).

14. U.S. CONST. amend. XIV.

15. Akhil Reed Amar, *The Bill of Rights and the Fourteenth Amendment*, 101 YALE L.J. 1193 (1992).

16. *See* U.S. CONST. amend. XIV, § 1.

17. United States v. Carolene Prod. Co., 304 U.S. 144, 153 n.4 (1938).

interest.[18] However, when the law relates to the fundamental rights of a vulnerable group, the court must instead apply strict scrutiny.[19] This doctrine was revived in the case of *Loving v. Virginia*, in which a Virginia law forbidding interracial marriage was struck down as unconstitutional.[20] Strict scrutiny requires the law's actual purpose to be compelling, unlike rational basis review in which a hypothetical justification will suffice.[21]

In short, the substantive Due Process Clause of the Fourteenth Amendment prohibits explicit bias on the basis of race because strict scrutiny forces legislators to justify their proposed law on explicitly non-prejudiced grounds.[22] However, this structure means that the Fourteenth Amendment is less well suited to prohibiting implicit bias. Additionally, the Fourteenth Amendment is limited by the State Action Doctrine, meaning that these protections are only enforceable against state actors, and not against private parties.[23] As we explain below, there are, however, statutory protections that protect more broadly against non-state actors. Because we take antidiscrimination laws in employment as a case study, and because most employers are not state actors, we therefore focus primarily in what follows on statutory protections.

C. Statutory Protections

To protect individuals from discrimination by private parties, Congress enacted the Civil Rights Act of 1964.[24] The constitutionality of this federal act was upheld in the case of *Heart of Atlanta Motel v. United States*, in which the Court ruled that Congress was authorized to pass such a law under the Commerce Clause.[25] The Civil Rights Act of 1964 was a landmark piece of social justice legislation. The Act is

18. *See* Nebbia v. New York, 291 U.S. 502 (1934).

19. United States v. Carolene Products Company, 304 U.S. 144 (1938); Korematsu v. United States, 323 U.S. 214 (1944).

20. Loving v. Virginia, 388 U.S. 1, 11-12 (1967).

21. We note that not all sensitive attributes are necessarily entitled to strict scrutiny. *See, e.g.* Bowers v. Hardwick, 478 U.S. 186, 194-96 (1986) (upholding an anti-sodomy law as constitutional, but failing to specify what level of scrutiny was applied to reach this decision); Craig v. Borden, 429 U.S. 190, 197 (1976) (wherein gender-based classifications required the application of intermediate scrutiny). Under intermediate scrutiny, it must be demonstrated that the law under consideration furthers the government's interest in a way that is substantially related to that interest. *See* Wengler v. Druggists Mutual Ins. Co., 446 U.S. 142, 150 (1980). *But see* Miss. Univ. for Women v. Hogan, 458 U.S. 718, 724 (1982) (requiring an "exceedingly persuasive justification" for gender-based discrimination, thereby applying intermediate scrutiny in a similar way to strict scrutiny).

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

23. Virginia v. Rives, 100 U.S. 313 (1880).

24. Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241 (1964).

25. Heart of Atlanta Motel, Inc. v. United States, 379 U.S. 241, 261 (1964).

broken into various titles, including Title VII, the focus of this paper, which provides antidiscrimination strictures that apply to employment practices:

> (a) Employer practices
>
> It shall be an unlawful employment practice for an employer -
>
> (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or
>
> (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin.[26]

Although we focus on Title VII and discrimination in employment decisions, our analysis is readily transferrable to other antidiscrimination doctrines, such as Title VI, which prohibits discrimination in admissions decisions for universities that receive Federal funding.[27] This is because "the main thrust of antidiscrimination law is fairly consistent across regimes."[28] Additionally, despite concentrating on racial discrimination, our analysis can be extended to discrimination on the basis of other protected attributes (e.g., national origin).

D. Disparate Treatment Under Title VII

The Civil Rights Act of 1964 provides two important causes of action: disparate impact and disparate treatment. Disparate impact claims generally involve individuals wronged by implicit bias, whereas disparate treatment claims are typically focused on explicit bias.

A disparate impact claim involves a practice or activity that appears facially neutral, but, has a disproportionately adverse impact on the protected group in question.[29] We do not focus on disparate impact claims for two reasons. First, machine learning and disparate impact

---

26. 42 U.S.C. § 2000e-2(a) (2012).

27. Title VI asserts, "No persons in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance." 42 U.S.C. § 2000d (2012). The purpose of Title VI is to make sure that no federal funds are used to subsidize, foster, or further discrimination. *See* Guardians Ass'n v. Civil Serv. Comm'n, 463 U.S. 582, 588-89 (1983); *see also* Alexander v. Choate, 469 U.S. 287, 292-93 (1985). Title VI notably applies to admissions practices for universities, since virtually all of them receive some form of federal financial assistance (e.g., via Pell grants, NSF grants, etc.). However, Title VI extends to myriad other institutions as well (e.g., Health and Human Services). Office for Civil Rights, *Civil Rights Requirements Title VI of the Civil Rights Act*, HHS.GOV, https://www.hhs.gov/civil-rights/for-individuals/special-topics/needy-families/civil-rights-requirements/index.html (last visited Feb. 9, 2020).

28. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 694 (2016).

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

have already received a thorough and thoughtful treatment in the literature.[30] Second, a disparate impact analysis will not differ across the various machine learning strategies that we consider, so analyzing this cause of action would not be particularly illuminating. We therefore place disparate impact outside of the scope of this paper and instead focus on the second type of claim that could be brought under Title VII: disparate treatment. We do, however, note that the line we draw between disparate impact and disparate treatment is, in reality, not quite so clear.[31]

A disparate treatment claim involves the intentional differential treatment of an individual on the basis of a protected attribute, such as race.[32] This prohibition of intentional discrimination is based on the Equal Protection Clause of the Fourteenth Amendment and hence involves a similar analysis.[33] In bringing a disparate treatment claim under Title VII, the plaintiff attempts to prove intentional discrimination via either direct or circumstantial evidence.

1. Direct Evidence

Direct evidence establishes discriminatory intent without the need for "inference or presumption."[34] The clearest forms of direct evidence are express classifications. These are situations in which the defendant either admits to using or explicitly uses a protected attribute, such as race, in her decision-making process.[35] In these situations, the defendant need not possess "bad faith, ill will, or any evil motive."[36] Any purposeful use of race invokes the highest level of judicial scrutiny, regardless of whether the motives leading to the discrimination were malevolent or benign.[37] Rather than trying to discern the defendant's motives, the investigation focuses on the adverse racial classification

30. Barocas & Selbst, *supra* note 28.

31. *See* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Barocas & Selbst, *supra* note 28; Richard A. Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341 (2010); George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313 (2006).

32. *See Guardians*, 463 U.S. at 607; *Alexander*, 469 U.S. at 292-93.

33. *See* Grutter v. Bollinger, 539 U.S. 306, 343–44 (2003) but recall the aforementioned discussion (*supra* note 21), noting that protection may differ across various sensitive attributes (viz. race and gender).

34. Davis v. Chevron, U.S.A., Inc., 14 F.3d 1082, 1085 (5th Cir. 1994).

35. Miller v. Johnson, 515 U.S. 900, 904–05 (1995).

36. Williams v. City of Dothan, 745 F.2d 1406, 1414 (11th Cir. 1984); Bangerter v. Orem City Corp., 46 F.3d 1491, 1501 (10th Cir. 1995); Ferrill v. Parker Grp., Inc., 168 F.3d 468, 473 n.7 (11th Cir. 1999).

37. City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493 (1989); Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 226 (1995).

of the plaintiff—that is, the "explicit terms of the discrimination."[38] "Put another way, direct evidence of intent is 'supplied by the policy itself.'"[39]

Additionally, it is not necessary for racial classification to be the sole basis for the defendant's decision.[40] Therefore, even if the defendant's explicit use of race may have only partially motivated her decision, this can still be sufficient to justify a direct-evidence disparate treatment claim. Title VII does, however, note that explicit consideration of race (and hence disparate treatment) is sometimes permissible (e.g., when taking affirmative action).[41]

There are also forms of direct evidence other than express classifications. For example, isolated comments by the defendant intimating the role that race played in her decision about the plaintiff can also be probative, so long as they are contemporaneous, suspiciously timed, or causally related to the decision.[42] However, we note that stray remarks or even derogatory comments are not direct evidence unless they too are related to the decision.[43] But in our analysis, express classification is the most pertinent form of direct evidence.[44]

2. *Circumstantial (or Indirect) Evidence*

In practice, most Title VII litigation centers around issues of circumstantial (or indirect) evidence because direct evidence is typically hard to find. This is because humans are rarely explicit in their use of racial classifications for decision-making. In other words, the subtlety of implicit bias and the factors that influence it means that there is rarely a "smoking gun."[45]

---

38. Int'l Union, United Auto. Aerospace & Agric. Implement Workers of Am. v. Johnson Controls, Inc., 499 U.S. 187, 199 (1991).

39. Hassan v. City of N.Y, 804 F.3d. 277, 295 (3d Cir. 2015) (quoting Massarsky v. Gen. Motors Corp., 706 F.2d 111, 128 (3d Cir. 1983) (Sloviter, J., dissenting)).

40. EQUAL EMP'T OPPORTUNITY COMM'N, COMPLIANCE MANUAL, SECTION 15: RACE AND COLOR DISCRIMINATION 15-9 (April 19, 2006), https://www.eeoc.gov/policy/docs/race-color.html (forbidding the use of a protected attribute as "all or part of the motivation for an employment decision"); Doe ex rel. Doe v. Lower Merion Sch. Dist., 665 F.3d 524, 548 (3d Cir. 2011).

41. *See infra* Part II.E.

42. Kennedy v. Schoenberg, Fisher & Newman, Ltd., 140 F.3d 716, 723 (7th Cir. 1998); Troupe v. May Department Stores, 20 F.3d 734, 736 (7th Cir. 1994).

43. Price Waterhouse v. Hopkins, 490 U.S. 228, 277 (1989); Fuentes v. Perskie, 32 F.3d 759, 767 (3d Cir. 1994). Even if such remarks do not serve as direct evidence, they can be used as circumstantial evidence of discriminatory intent. Fitzgerald v. Action, Inc., 521 F.3d 867, 877 (8th Cir. 2008).

44. *See infra* Section II.C.

45. Aman v. Cort Furniture Rental Corp., 85 F.3d 1074, 1081-82 (3d Cir. 1996) ("It has become easier to coat various forms of discrimination with the appearance of propriety, or to ascribe some other less odious intention to what is in reality discriminatory behavior. In other words, while discriminatory conduct persists, violators have learned not to leave the proverbial 'smoking gun' behind.").

There are two major frameworks used in presenting circumstantial evidence for disparate treatment. The first is the burden-shifting *McDonnell-Douglas* framework.[46] In this framework, the plaintiff attempts to demonstrate that the defendant gave differential treatment to similarly situated individuals on the basis of their race, color, or national origin.[47] This is achieved via three burden-shifting steps.[48] In the first step, the plaintiff must, by a preponderance of the evidence, establish a prima facie case. This is typically achieved by producing evidence of a similarly situated individual who was treated differently from the plaintiff.[49] After this step, the burden shifts to the defendant, who must articulate a legitimate non-discriminatory reason for acting in such a manner.[50] This is a burden of production, not of persuasion.[51] Although this burden is a fairly low bar, the defendant's reason must be "clear and reasonably specific."[52] If successful, the burden is then placed back on the plaintiff to demonstrate that the defendant's proffered reason is false[53] and exists merely as pretext for her true discriminatory intent.[54] There are a number of different ways in which the plaintiff can attempt to prove pretext. First, the plaintiff can identify "weaknesses, implausibilities, inconsistencies, incoherencies, or contradictions" in the defendant's reason.[55] Second, she can demonstrate that the defendant deviated from a written or unwritten policy regarding the decision-making process.[56] Finally, she can produce evidence that the defendant's reason is nothing more than a "*post hoc* fabrication."[57]

---

46. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 802 (1973).

47. *See* Brewer v. Bd. of Trs. of Univ. of Ill., 479 F.3d 908, 921 (7th Cir. 2007) (where the inability to identify a similarly situated individual destroyed the disparate treatment claim).

48. McDonnell Douglas Corp. v. Green, *supra* note 46.

49. *See* Brewer v. Bd. of Trs. of Univ. of Ill., 479 F.3d 908, 921 (7th Cir. 2007).

50. EEOC v. Boeing Co., 577 F.3d 1044, 1049 (9th Cir. 2009).

51. St. Mary's Honor Ctr. v. Hicks, 509 U.S. 502, 506-07 (1993).

52. *See* Tex. Dep't of Cmty. Affairs v. Burdine, 450 U.S. 248, 254–55, 258 (1980). For example, reference to selecting the "best qualified" applicant is insufficiently vague. Instead specific factors such as education, comparable work experience, or seniority *inter alia* must be cited. Steger v. Gen. Elec. Co., 318 F.3d 1066, 1075–76 (11th Cir. 2003).

53. Finding the defendant's proffered reason false allows, but does not require, a finding of discrimination. *See* Anderson v. Baxter Healthcare Corp., 13 F.3d 1120, 1123 (7th Cir. 1994); St. Mary's Honor Ctr v. Hicks, 509 U.S. 502, 511 (1993). However, there is some uncertainty about how many of the defendant's proffered reasons must be successfully falsified. *See* Monroe v. Children's Home Ass'n, 128 F.3d 591, 593 (7th Cir. 1997) (reasoning that not all of the defendant's reasons need to be falsified); Coco v. Elmwood Care, Inc., 128 F.3d 1177, 1178-79 (7th Cir. 1997) (reasoning that falsifying only one of several of the defendant's reasons may be insufficient to survive summary judgment against the plaintiff).

54. Brooks v. Cty. Comm'n of Jefferson Cty., 446 F.3d 1160, 1162 (11th Cir. 2006).

55. *Id.* at 1163 (quoting Jackson v. Ala. State Tenure Comm'n, 405 F.3d 1276, 1289 (11th Cir. 2005)).

56. *See* Plotke v. White, 405 F.3d 1092, 1102 (10th Cir. 2005).

57. Fuentes v. Perskie, 32 F.3d 759, 764 (3d Cir. 1994).

The second framework is the *Arlington Heights* framework, in which a number of factors are considered for their probative value in establishing the defendant's discriminatory intent.[58] Relevant factors include, but are not limited to,[59] statistical evidence evincing a clear pattern of discrimination, the specific sequence of events that occurred prior to the case, and relevant legislative or administrative history.[60] The *Arlington Heights* framework is often most usefully employed when litigation reveals that there are a variety of different forms of evidence.[61]

E. Affirmative Action

There are some situations, however, in which the explicit consideration of race is permissible. For example, the Equal Opportunity Employment Commission (EEOC) cites the following three circumstances in which voluntary affirmative action is appropriate:

> (a) Adverse effect. Title VII prohibits practices, procedures, or policies which have an adverse impact unless they are justified by business necessity. In addition, Title VII proscribes practices which "tend to deprive" persons of equal employment opportunities. Employers, labor organizations and other persons subject to Title VII may take affirmative action based on an analysis which reveals facts constituting actual or potential adverse impact, if such adverse impact is likely to result from existing or contemplated practices.
>
> (b) Effects of prior discriminatory practices. Employers, labor organizations, or other persons subject to Title VII may also take affirmative action to correct the effects of prior discriminatory practices. The effects of prior discriminatory practices can be initially identified by a comparison between the employer's work force, or a part thereof, and an appropriate segment of the labor force.
>
> (c) Limited labor pool. Because of historic restrictions by employers, labor organizations, and others, there are circumstances in which the available pool, particularly of qualified minorities and women, for employment or promotional opportunities is artificially limited. Employers, labor organizations, and other persons subject to Title VII may, and are encouraged to take affirmative action in such circumstances.[62]

---

58. *See* Vill. of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252, 266-68 (1977).

59. Pac. Shores Props., LLC v. City of Newport Beach, 730 F.3d 1142, 1158-59 (9th Cir. 2013).

60. *Id.*; Sylvia Dev. Corp. v. Calvert Cty., 48 F.3d 810, 819 (4th Cir. 1995).

61. CIVIL RIGHTS DIVISION, TITLE VI LEGAL MANUAL, U.S. DEPARTMENT OF JUSTICE VI.B.2., https://www.justice.gov/crt/case-document/file/923551/download (last visited Apr. 29, 2020).

62. Circumstances under which voluntary affirmative action is appropriate, 29 C.F.R. § 1608.3(a)-(c) (2019).

Affirmative action has been used by employers to increase diversity with regard to underrepresented minorities.[63] In short, affirmative action—when appropriately executed—is a permissible instance of disparate treatment. However, the use of sensitive attributes to fulfill quotas, even for the purpose of affirmative action, is prohibited. Indeed, the rote filling of racial quotas was deemed a violation of the Fourteenth Amendment's Equal Protection Clause.[64]

We note that the practical utility of appealing to affirmative action as a defense for disparate treatment may be limited given certain political (and possibly constitutional) realities. As Selbst and Barocas point out in their paper on machine learning and disparate impact, "[p]olitically, anything that even hints at affirmative action is a nonstarter today, and to the extent that it is permissible to enact such policies, their future constitutionality is in doubt."[65]

F. Legal Protections Operationalized: Organizational Monitoring Strategies

To maintain compliance with legal protections against discrimination, humans have operationalized laws like Title VII via various organizational monitoring strategies. This approach is evidenced by the rise and prevalence of strategies such as internal dispute resolution, mandatory arbitration, and antidiscrimination training materials and hiring policies.[66]

Perhaps the most common strategy is mandatory diversity or implicit bias training, which studies have shown to be ineffective.[67] For this reason, we do discuss it, despite its widespread popularity. Another strategy that is sometimes used is perspective taking.[68] This seemingly simple strategy asks individuals to engage in imaginative

---

63. In addition to Title VII and employment cases, Title VI case law has stipulated that remedying historical discrimination and promoting fairness in higher education can constitute compelling government interest and is therefore permissible. *See* United States v. Paradise, 480 U.S. 149, 167 (1987); Grutter v. Bollinger, 539 U.S. 306, 343 (2003). *See also* Nondiscrimination; Equal Employment Opportunity; Policies and Procedures, 28 C.F.R. § 42.104(b)(6)(1) (2019) (asserting that when "administering a program regarding which the recipient has previously discriminated against persons on the ground of race, color, or national origin, the recipient must take affirmative action to overcome the effects of prior discrimination").

64. Regents of Univ. of Cal. v. Bakke, 438 U.S. 265, 307 (1978).

65. Barocas & Selbst, *supra* note 28, at 715.

66. *See generally* Cynthia Estlund, *Rebuilding the Law of the Workplace in an Era of Self-Regulation*, 105 COLUM. L. REV. 319 (2005) for a thoughtful discussion of this trend and its consequences.

67. *See* Lisa Legault, Jennifer N. Gutsell & Michael Inzlicht, *Ironic Effects of Antiprejudice Messages: How Motivational Interventions Can Reduce (but Also Increase) Prejudice*, 22 PSYCHOL. SCI. 1472, 1476 (2011); Frank Dobbin & Alexandra Kalev, *Why Diversity Programs Fail*, HARV. BUS. REV. (2016), https://hbr.org/2016/07/why-diversity-programs-fail.

68. *See generally* Theresa K. Vescio, Gretchen B. Sechrist & Matthew P. Paolucci, *Perspective Taking and Prejudice Reduction: The Mediational Role of Empathy Arousal and Situational Attributions*, 33 EUR. J. SOC. PSYCHOL. 455 (2003).

exercises in which they envision themselves in the shoes of a person belonging to a group that they are biased against. People who engage in such exercises have been shown to have improved attitudes toward the group in question.[69] This attitude improvement is mediated by increased empathy and situational attributions (*viz.* minimization of the fundamental attribution error,[70] in which people tend to attribute the actions of others to internal dispositions but attribute their own actions to external situations). In short, perspective taking increases empathy toward others and makes people less susceptible to attribution bias. Perspective taking attenuates bias even in quick, unconscious cognitive processing.[71] Studies of perspective taking in social psychology have focused on bias against women, African Americans, members of the LGBT community, the indigent, and the clinically obese.[72] We emphasize that these studies do not vaunt perspective taking as a "cure" for bias. Although perspective taking can temper bias, it cannot eliminate it entirely.

In practice, strategies like perspective taking have not been widely adopted, likely because of difficulties in implementing and enforcing their use, as well as limited efficacy. A promising new development is the use of virtual reality technology to assist with perspective taking.[73] However, it remains to be seen whether virtual reality technology can help mitigate bias, not least because there are several barriers to its use. The first is access, given that virtual reality technology is not ubiquitous. The second is that it may raise ethical and legal concerns. Consider using virtual reality technology to temper gender bias in employment decisions by building empathy toward people who have been sexually harassed during interviews.[74] The more realistic a virtual en-

---

69. Adam D. Galinsky & Gordon B. Moskowitz, *Perspective-Taking: Decreasing Stereotype Expression, Stereotype Accessibility, and in-Group Favoritism*, 78 J. PERSONALITY & SOC. PSYCHOL. 708, 708 (2000).

70. Edward E. Jones & Victor A. Harris, *The Attribution of Attitudes*, 3 J. EXPERIMENTAL SOC. PSYCHOL. 1 (1967).

71. *See* Andrew R. Todd, Galen V. Bodenhausen & Adam D. Galinsky, *Perspective Taking Combats the Denial of Intergroup Discrimination*, 48 J. EXPERIMENTAL SOC. PSYCHOL. 738, 739 (2012).

72. *See, e.g.*, Nilanjana Dasgupta & Shaki Asgari, *Seeing Is Believing: Exposure to Counterstereotypic Women Leaders and Its Effect on the Malleability of Automatic Gender Stereotyping*, 40 J. EXPERIMENTAL SOC. PSYCHOL. 642, 654 (2004); Jamie Lee Gloor & Rebecca M. Puhl, *Empathy and Perspective-Taking: Examination and Comparison of Strategies to Reduce Weight Stigma*, 1 STIGMA & HEALTH 269, 271 (2016); Colin Tucker Smith et al., *Perspective Taking Explains Gender Differences in Late Adolescents' Attitudes Toward Disadvantaged Groups*, 45 J. YOUTH & ADOLESCENCE 1283, 1284-85 (2015).

73. *See* Natalie Salmanowitz, *The Impact of Virtual Reality on Implicit Racial Bias and Mock Legal Decisions*, 5 J. L. & BIOSCI. 174, 176 (2018); Lara Maister et al., *Changing Bodies Changes Minds: Owning Another Body Affects Social Cognition*, 19 TRENDS IN COGNITIVE SCI. 6 (2015).

74. *See* Amanda Holpuch & Olivia Solon, *Can VR Teach Us How to Deal with Sexual Harassment?*, GUARDIAN (May 1, 2018, 6:00 PM), https://www.theguarian.com/world/2018/may/01/sexual-assault-training-program-vantage-point-virtual-reality-video-games.

counter is, the more this intervention might seem like employer-sanctioned virtual harassment of its employees, which is obviously not a particularly tenable strategy.

A more easily deployed strategy for mitigating bias is to "blind" human decision makers to the sensitive attributes of the applicants, such as race. Numerous individuals and organizations have recognized the importance of such blinding. For example, teachers often impose blind grading practices on themselves, while reviewers for academic journals and conferences routinely review paper submissions that have been stripped of the authors' identities.

Modern orchestral hiring practices serve as an informative example of self-imposed blinding.[75] Orchestras have long suffered from the problem of being disproportionately male. This imbalance is due to a longstanding anti-female stereotype, leading to female applicants being unfairly rejected. In the 1970s and 1980s, many orchestras imposed changes to their hiring practices intended to minimize this gender bias. One of these changes was to institute blind auditions. The judges responsible for making hiring decisions recognized that their implicit bias meant that they could not be impartial, so they needed to blind themselves to the gender of the applicants. This was typically achieved by using a large cloth screen to hide the applicants. But the judges quickly realized that they could see the shoes of the applicants under the screen, so the screen was extended down to the floor. However, even with this modification, the judges could hear the applicants' shoes as they walked across the stage, so they laid down carpet or made the applicants remove their shoes. In other words, shoe appearance and sound were proxies for gender, so the judges had to blind themselves to these signals as well. Ultimately, the more blind to gender the judges were, the less gender biased their decision-making process became.

This example illustrates a (somewhat) successful implementation of blinding, while also introducing the idea of proxies. Many factors can serve as proxies for the sensitive attributes of the applicants. In other words, blinding is difficult because information can still leak in, and humans are adroit at detecting subtle signals of group membership. In this example, the applicants' shoes served only as a proxy for gender and were not relevant to the decision-making process. Therefore, the judges could blind themselves to this proxy without impairing their decisions. However, in practice, pure proxies are rare, which complicates the use of blinding as a strategy for mitigating bias. So, even the most promising strategy of blinding falls short of our hopes of eliminating bias from human decision-making processes. Perhaps technology can help?

---

75. *See* Claudia Goldin & Cecilia Rouse, *Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians*, 90 AM. ECON. REV. 715, 716 (2000).

III. MACHINE LEARNING TO THE RESCUE?

"But they [computers] are useless. They can only give you answers." Pablo Picasso[76]

A. Bias in Machine Learning

It is not possible to perfectly blind human decision makers to sensitive attributes (and their proxies), which is one of the reasons why bias has not been eradicated from human decision-making processes. The dawn of automated decision-making processes shone a ray of hope onto this seemingly intractable problem.[77] Finally, an impartial arbiter that could prevent discrimination because, unlike humans, machines are unafflicted by animus, sentiment, and implicit bias. The prevailing view was that computers were cold calculation machines, uninhibited by evolutionary and societal pressures, and thus incapable of discrimination. However, this myth that "numbers are neutral" paints a dangerously inaccurate picture.

Machine learning has established itself as an impressive tool for synthesizing data to make accurate predictions.[78]Machine learning was intended to move beyond the hassle of rule-based artificial intelligence, in which programmers would have to tell the computer exactly what to do. Instead, programmers could rely on mathematics and an abundance of data to "teach" the computer how to make predictions.[79] A common misconception is that because machine learning is just math, it cannot exhibit bias—a computer only does what its programmers tell it to do, and its programmers are not telling it to be biased. However, this narrative glosses over some of the nuance of machine learning methods, and the many ways that bias can still creep in.[80]

First, machine learning systems are designed by humans, and the human condition is one afflicted with bias. Humans—either programmers or others involved in development and deployment—make value

---

76. William Fifield, *Pablo Picasso: A Composite Interview*, PARIS REV. 32 (Summer-Fall 1964).

77. Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. (CLINICAL RES. ED.), 657, 657 (1988) (providing a classic example of this initial optimism being dashed by the unintended consequences of an automated decision-making process gone awry).

78. See, e.g., Alon Halevy, Peter Norvig and Fernando Pereira, *The Unreasonable Effectiveness of Data*, 24 IEEE INTELLIGENT SYSTEMS 8 (2009).

79. For the reader who is unfamiliar with machine learning and is looking for an accessible, yet thorough, primer, *see* Andrew Ng, *Machine Learning*, COURSERA, https://www.coursera.org/learn/machine-learning (last visited Apr. 30, 2020).

80. For an introductory overview to the various means by which bias can be perpetuated by a machine learning system, *see* Karen Hao, *This is how AI bias really happens—and why it's so hard to fix*, MIT TECH. REV. (Feb. 4, 2019), https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/. *See also* Batya Friedman & Helen Nissenbaum, *Bias in Computer Systems*, 14 ACM TRANS. INFO. SYS'S. 330, 332 (1996).

judgements and assumptions about a system and its operation in the world, all of which are necessarily impacted by any implicit bias affecting those humans.[81]

Second, humans produce the data with which machine learning systems are trained. Not only do humans select the specific datasets to use, but the data itself comes from society and society, on the whole, is biased. In other words, there is no such thing as "unbiased" data—or, at least, it is rare and very hard to find. There is a growing literature in machine learning demonstrating that a machine learning system trained using biased data will necessarily make biased predictions.[82] If proper care is not taken when developing and deploying machine learning systems, their use runs the risk of not only perpetuating existing bias, but further entrenching it or even creating new forms of bias. This danger is only exacerbated by automation bias.[83] Despite considerable attention given to people's aversion toward machines, people appear to uncritically prefer and accept decisions made by machines over decisions made by humans.[84]

Laws have been historically been designed with human decision makers in mind. However, automated decision-making processes are very different from human decision-making processes. This creates a mismatch between laws and the decision makers to which they are intended to apply, which can lead to counterproductive outcomes. This mismatch is particularly salient now that machine learning systems are increasingly used to make decisions in high-stakes domains such as employment, university admissions, and even criminal justice,[85] and has even engendered a fierce debate over the fairness of machine learning systems and their use.[86]

Given this mismatch, machine learning developers are left confused as to the best way to create systems that comply with the law. In other

---

81. *See* Selbst & Barocas, *supra* note 28, at 677-93 for a nice overview of the various ways in which bias can be imputed by a machine learning system during the development process.

82. *See* Tolga Bolukbasi et al., *Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings*, ARXIV, Jul. 21, 2016, https://arxiv.org/pdf/1607.06520.pdf; Joy Boulamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RES. 1, 1 (2018).

83. *See generally* Jennifer Logg, Julia Minson & Don A. Moore, *Algorithm Appreciation: People Prefer Algorithmic To Human Judgment*, 151 ORG. BEHAV. & HUM. DECISION PROCESSES 90 (2019); Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 HUM. FACTORS 230 (1997); Linda J. Skitka, Kathleen L. Mosier & Mark Burdick, *Does Automation Bias Decisionmaking?*, 51 INT'L. J. HUM.-COMPUTER STUDS. 991 (1999); Mary T. Dzindolet et al., *The Role of Trust in Automation Reliance*, 58 INT'L. J. HUM.-COMPUTER STUDS. 697 (2003).

84. *See generally* Logg et al., *supra* note 83, at 90.

85. *See* John Monahan & Jennifer L. Skeem, *Risk Assessment in Criminal Sentencing*, 12 ANN. REV. CLINICAL PSYCHOL. 489 (2016).

86. Julia Angwin et al. *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

words, developers want to know how to best to operationalize antidiscrimination laws so that their systems conform to relevant legal requirements. A reasonable starting point would be to look at how humans have operationalized laws. As we explained earlier, this is often accomplished via various organizational monitoring strategies, including diversity or implicit bias training, perspective taking, and blinding.[87] Although these strategies are all appropriate for human decision makers, the one that is most applicable to automated decision makers is blinding. As we explain next, there are three ways to do this in the context of machine learning: total blinding, no blinding, and partial blinding.

B. The Three Strategies for Blinding

One might think it possible to make a machine learning system unbiased by completely blinding it to sensitive attributes (strategy 1). This strategy is well-intentioned, but as we explained via the orchestra example above, proxies for sensitive attributes abound, and as adroit as humans are at detecting these proxies, they pale in comparison to machines. If a machine learning system is trained using a dataset that has been stripped of sensitive attributes, the system can still "reconstruct" this information via proxies.[88] Furthermore, for high-dimensional datasets, sensitive attributes may be redundantly encoded across many features, and there is no easy way to determine when a feature is too correlated with a sensitive attribute.[89] Worse still, if sensitive attributes are discarded, it is much harder to detect and isolate their effects, which will be more diffuse.[90] In other words, this strategy is not particularly effective at mitigating bias.

From a machine learning perspective, there are better ways to mitigate bias. Counterintuitively, these strategies rely on giving the system access to sensitive attributes. One strategy is to not blind the system to sensitive attributes at all, thereby allowing the system to use this information to mitigate bias (strategy 2). This strategy encompasses various techniques for facilitating fair decisions, such as post-

87. *See supra* Part II.F.

88. *See e.g.* Sam Corbett-Davies and Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV, Aug. 14, 2018, https://arxiv.org/pdf/1808.00023.pdf (reviewing the shortcomings of various anticlassification approaches).

89. Cynthia Dwork et al., *Fairness Through Awareness,* ARXIV, Nov. 29, 2011, https://arxiv.org/pdf/1104.3913.pdf.

90. *See* Rich Caruana et al., *Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission, in* Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 1721 (2015); Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS & PROC. 22 (2018).

processing techniques,[91] decoupled classifiers,[92] or a single classifier with different decision thresholds.[93] It is a more active strategy than strategy 1, and it also has the virtue of making it easier to correct for existing bias reflected in the data. But as we will discuss in further detail, using strategy 2 may be problematic from a legal perspective.[94]

Finally, we believe there lies a middle ground between strategies 1 and 2, in which the system is blinded to sensitive attributes only during deployment and not during training (strategy 3). Early techniques along these lines involved culling decision rules from an expert system on the basis of their relation to a sensitive attribute.[95] There are several more modern techniques that also implement this strategy, such as including the sensitive attribute as a feature during training and then removing its effect prior to deployment various ways of preprocessing the data used to train the system, or directly training the system to simultaneously maximize prediction accuracy and a parity-based fairness metric.[96] Regardless of the specific technique, though, the end result is a machine learning system that can be deployed without access to sensitive attributes.[97]

---

91. Moritz Hardt, Eric Price & Nathan Srebro, *Equality of Opportunity in Supervised Learning*, ARXIV, Oct. 7, 2016, https://arxiv.org/pdf/1610.02413.pdf.

92. Cynthia Dwork et al., *Decoupled Classifiers for Fair and Efficient Machine Learning*, ARXIV, Jul, 20, 2020, https://arxiv.org/pdf/1707.06613.pdf.

93. *See* Kleinberg et al., *supra* note 90.

94. *See infra* Part II.C.b.

95. *See* Dino Pedreschi, Salvatore Ruggieri & Franco Turini, *Discrimination-Aware Data Mining*, *in* Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 560 (2008).

96. *See* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J.: ECON. POL. 206 (2011); Alekh Agarwal et al., *A Reductions Approach to Fair Classification*, ARXIV, Jul. 16, 2018, https://arxiv.org/pdf/1803.02453.pdf; Faisal Kamiran & Toon Calders, *Classifying without Discriminating*, *in* 2009 2nd International Conference on Computer, Control, and Communication 1 (2009); Faisal Kamiran, Toon Calders & Myloka Pechenizkiy, *Discrimination Aware Decision Tree Learning*, *in* 2010 IEEE International Conference on Data Mining 869 (2010); Muhammad Bilal Zafar et al., *Fairness Constraints: Mechanisms for Fair Classification*, ARXIV, Mar. 23. 2017, https://arxiv.org/pdf/1507.05259.pdf. *But see* Zachary C. Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?[v3]*, *in* 32ND CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, ARXIV, Jan. 11, 2019, https://arxiv.org/pdf/1711.07076.pdf (expressing concerns over the efficacy of such methods using disparate learning processes).

97. It is often hypothesized that there is a tradeoff between fairness and accuracy, but in fact that may only be with regard to the specific datasets used to train and test the system, not the data encountered during deployment. *See* Aditya Krishna Menon & Robert C. Williamson, *The Cost of Fairness in Binary Classification*, *in* Proceedings 1st Conference on Fairness, Accountability, and Transparency 10 (2018); Dwork et al., *supra* note 92. Moreover, viewing the system's deployment context more broadly and considering factors not captured by the data, it might be the case that the hypothesized fairness–accuracy tradeoff is in fact illusory, and everyone is better off when bias is mitigated. This will be an interesting issue to explore empirically as machine learning systems that explicitly mitigate bias are deployed in the wild. *See* Michael Veale & Reuben Binns, *Fairer Machine Learning in the Real World: Mitigating Discrimination without Collecting Sensitive Data*, BIG DATA & SOC'Y,

It is important to note that the impact of using strategy 2 or 3 depends greatly on the specific definition of fairness that is used. A number of different definitions have been proposed in the machine learning literature, focusing on both outcome-based[98] and (much less frequently) process-based[99] conceptions of fairness.[100] However, there is much disagreement about these definitions, and if and when they should be used. This disagreement is further complicated by research demonstrating that some definitions, although facially valid, can be mathematically incompatible, thus forcing a choice between having and eating our proverbial cake.[101]

C. *Legal Analysis of The Three Strategies*

1. Human vs. Automated Decision Makers: Two Key Differences in Disparate Treatment

Before we analyze the three strategies, we note two important differences between human decision makers and automated decision makers when it comes to disparate treatment claims. First, recall that a disparate treatment claim involves the intentional differential treatment of a similarly situated individual on the basis of a protected attribute.[102] One might be tempted to think that because a machine has no motives at all, let alone wicked ones, it cannot intentionally discriminate. However, intentionality is defined broadly in the context of Title VII claims, in that "ill will, enmity, or hostility are not prerequisites of intentional discrimination."[103]

Second, when looking for evidence of disparate treatment, we cannot directly access a human's thoughts, so we turn to external indicia of discriminatory intent. However, humans are rarely explicit in their

---

Jul.–Dec. 2017, at 1; Kenneth Holstein et al., *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, in* 2019 ACM CHI Conference on Human Factors in Computing Systems (2019); *See generally* Lydia T. Liu et al., *Delayed Impact of Fair Machine Learning, in* Proceedings 35th International Conference on Machine Learning (2018).

98. *See* Dwork et al., *supra* note 92; Hardt, Price,& Srebro, *supra* note 91.

99. *See* Nina Grgić-Hlača et al., *The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making, in* Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems (2016).

100. *See* The Conference on Fairness, Accountability, and Transparency, *FAT* 2018 Translation Tutorial: 21 Definitions of Fairness and Their Politics*, YOUTUBE (April 18, 2018), https://www.youtube.com/watch?v=wqamrPkF5kk.

101. For the seminal papers and their simultaneous discovery of this surprising result, *see* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV, Nov. 17, 2016, https://arxiv.org/pdf/1609.05807.pdf; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, ARXIV, Oct. 24, 2016, https://arxiv.org/abs/1610.07524.

102. *See supra* Part II.D.

103. *See* Ferrill v. Parker Grp., Inc., 168 F.3d 468, 473 n.7 (11th Cir. 1999); *supra* Part II.D.1; *supra* note 36.

use of racial classifications for decision-making. This means that "direct evidence of intentional discrimination is hard to come by."[104] In contrast, machines do not know that concealing their consideration of race is the prudent or politically correct thing to do. So long as race is included as a feature, a machine learning system will wear its explicit bias on its sleeve, free for anyone with access to its internals to analyze as direct evidence of intentional discrimination.[105] This difference reveals an asymmetry: circumstantial evidence is most likely to assist with disparate treatment claims involving human decision makers, whereas direct evidence most likely to assist with disparate treatment claims involving automated decision makers.

Finally, we note that sensitive attributes (or proxies) do not fully determine most machine learning systems' predictions. In other words, race (or proxies for race) are typically not dispositive in a machine learning context. However, this does not foreclose a disparate treatment claim, because racial classification does not need to be the sole basis for the decision.[106]

2. *Analysis of the Three Strategies: A Problematic Tension*

We now analyze the three strategies in relation to disparate treatment claims. Each strategy is accompanied by a slightly different legal risk profile. Strategy 1, which involves blinding the machine learning system to the race of the applicants, would appear to be insulated against disparate treatment claims under Title VII because the system does not explicitly consider race in its decision-making process.[107] But, as we noted earlier, if the reason for using this strategy is to mitigate bias, this is an imprudent strategy.[108] Blinding the system to race does not necessarily mitigate bias because this information can be "reconstructed" via proxies. Moreover, if race is omitted as a feature, then detecting and isolating its effect is much harder. As a result, although

104. *See* Price Waterhouse v. Hopkins, 490 U.S. 228, 271 (1989); *supra* Part II.D.1.

105. There are two important exceptions, where it is not so easy to access and analyze the system's internals. The first is that the sophistication and complexity of a machine learning system may render it quite opaque. This relates to a vast literature on transparency and interpretability in machine learning. *See* Zachary C. Lipton, *The Mythos of Model Interpretability*, ARXIV, May 6, 2017, https://arxiv.org/pdf/1606.03490.pdf, for an overview of various meanings of interpretability in machine learning. *See also* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018) (making important distinctions between inscrutability of a machine learning system and non-intuitiveness). The second important exception stems from precedent protecting against divulgence of the internals of machine learning systems—even in high-stakes domains—on the basis of trade secret law. *See* State v. Loomis, 881 N.W.2d 749, 760-61 (Wis. 2016). Both of these exceptions are important and interesting, but unfortunately lie outside the scope of this paper.

106. Doe ex rel. Doe v. Lower Merion Sch. Dist., 665 F.3d 524, 548 (3d Cir. 2011).

107. However, we note that such a strategy is likely to be subject to disparate impact liability.

108. *See* Kleinberg et al., *supra* note 90.

this strategy appears to introduce little risk of a disparate treatment claim, it does run a serious risk of yielding a biased system. In short, strategy 1 is "too hot."

Strategy 2 is the opposite of strategy 1 in that the system is not blinded race at all, thereby allowing it to use this information to mitigate bias via various machine learning techniques for facilitating fair decisions. This strategy's explicit use of race appears to result in disparate treatment under Title VII. In particular, it creates direct evidence of disparate treatment because the proof of intent to make decisions on the basis of race is "supplied by the policy itself," where the policy is the machine learning system.[109] We note that several legal scholars have explored creative arguments in an attempt to show that strategy 2 can be compliant with Title VII.[110]

However, these arguments are not uncontroversial,[111] and, more importantly, we believe such legal gymnastics are not necessary given the sociotechnical nature of strategy 3, which draws on work in both machine learning and the law. In short, strategy 2 is "too cold."

Thus, a problematic tension emerges between Title VII, which forbids the explicit use of race in decision-making, and the machine learning perspective, which recognizes that a system must be given access to sensitive attributes in order to mitigate bias. This tension arises from our attempt to stretch human laws to apply to machines even though human decision-making processes are quite different from automated decision-making processes. In other words, when stretched to apply to machines, laws designed to regulate human behavior may even be detrimental to the very people that they were designed to protect. It is sadly ironic that machine learning techniques

109. *See* Hassan v. City of New York, 804 F.3d. 277, 295 (3d Cir. 2015) (quoting Massarsky v. Gen. Motors Corp., 706 F.2d 111, 128 (3d Cir. 1983) (Sloviter, J., dissenting)).

110. *See* Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L. J. (forthcoming) (providing an in-depth investigation of various legal arguments that attempt to press strategy 2 into compliance with Title VII, including analysis of the "strong basis in evidence" justification, as well as ancillary dicta from *Ricci v. DeStefano*, 557 U.S. 557 (2009). Additionally, Pauline Kim does a thorough job exploring a nonconventional reading of Title VII. *See* Kim, *supra* note 31. Kim argues that the standard reading of Title VII prohibits the use of sensitive attributes in machine learning systems, but, notes that this is to the detriment of fairness. She says, "If developers purge demographic variables such as race and sex from the dataset, it becomes more difficult, if not impossible, to determine whether a model is systematically biased. Preserving these variables allows a model to be tested to determine its effect on the distribution of opportunities among different groups." *Id.* at 918. This nicely summarizes the points we make regarding strategy 1. However, she argues that a close (and nonconventional) reading of Title VII reveals the statutory permissibility of including and using sensitive attributes, such as race. The notion of algorithmic affirmative action has also been explored by machine learning researchers. *See, e.g.*, Dwork et al., *supra* note 92 (proposing a definition of fair affirmative action linked to the fairness definition discussed in their paper, namely, individual fairness).

111. For example, Pauline Kim frames her argument in contradistinction to the position taken by Barocas & Selbst, *supra* note 28. *See* Kim, *supra* note 31, at 909.

for facilitating fair decisions and Title VII both strive to achieve the same goal but stand diametrically opposed to one another with regard to how to achieve this goal.

3. The Goldilocks Solution

Neither strategy 1 nor 2 seem tenable, but just like for Goldilocks, there is a third option that is "just right," namely, strategy 3. Strategy 3 seems to be a middle ground between strategies 1 and 2 because the system is blind to race during deployment, like strategy 1, but is able to use race during training to mitigate bias, like strategy 2. Therefore, strategy 3 seems to be a way to balance the problematic tension between the machine learning perspective and Title VII.

A possible concern regarding strategy 3 is whether its explicit consideration of race during training constitutes disparate treatment. We argue that this concern can be ameliorated by analogizing strategy 3 to legally accepted auditing procedures. Auditing procedures that protect against discrimination are commonly used in the context of employment. For example, the mandatory employer reporting requirements implemented by the EEOC require any employer subject to Title VII with 100 or more employees file an Employee Information Report EEO-1 containing information about employee demographics.[112] This information is collected for auditing purposes to guard against biased hiring practices. Similarly, universities routinely collect sensitive attributes from applicants.[113] This information is not used for decision-making, but, can instead be used to audit the process for bias in order to make it fairer for future applicants.

These auditing procedures are remarkably similar to strategy 3. The main difference is that strategy 3 considers the sensitive attribute *ex ante*, whereas auditing considers it *ex post*. But both do so for the same purposes. Therefore, there seems to be a strong analogy between strategy 3 and legally accepted auditing procedures. We note that this analogy does not hold for strategy 2. Strategy 2 considers the sensitive attribute during decision-making, while auditing does not.

In summary, strategy 3 is a "Goldilocks" solution. Strategy 1 is too hot because although it appears to comply with Title VII, it runs a risk of yielding a biased system.[114] Strategy 2 is too cold because although it is more effective at mitigating bias than strategy 1, it runs afoul of Title VII by explicitly considering race during deployment.[115] Strategy 3 is just right. It appears to comply with Title VII because the system

112. *See* Records and Reports, 29 C.F.R. 14 §1602.7 (1976).

113. *See* the most recent Status and Trends in the Education of Racial and Ethnic Groups Report conducted by the National Center for Education Statistics (NCES). U. S. Department of Education, *Status and Trends in the Education of Racial and Ethnic Groups 2018*, NCES (Feb. 2019), https://nces.ed.gov/pubs2019/2019038.pdf.

114. *See supra* Section II.C.b.

115. *See supra* Section II.C.b.

is blind to race during deployment, yet the system is still able to use race during training to mitigate bias. Strategy 3 avoids disparate treatment claims because it does not use racial classifications for decision-making; meanwhile, its consideration of race during training can be analogized to legally accepted auditing procedures, with the main difference being that strategy 3 considers the sensitive attribute *ex ante* rather than *ex post*.

IV. OBJECTIONS

"'Objection, your Honor!' 'Overruled' 'No, no. I strenuously object.' 'Oh! You strenuously object. Then I'll take some time and reconsider.'"
A Few Good Men[116]

A. Harms Arising from "Fairness"

One might think that strategy 3 is solving a non-existent problem because strategy 2 is not causing any harm and therefore does not violate Title VII. In other words, if strategy 2 mitigates bias, then no one is harmed—and with no harm, how can there be a disparate treatment claim? After all, by using strategy 2, the system no longer adversely affects the applicants who might have been harmed otherwise, even though the system does use racial classifications.

But there are several legal bases on which strategy 2 could still cause harm. First, the stigma involved in racial classifications can constitute a cognizable harm.[117] Therefore, even an applicant who is ultimately selected can be said to be harmed if her selection involved classifying the applicant on the basis of her race. Perhaps the most noted proponent of this view is Justice Clarence Thomas, who asserted that "These [affirmative action] programs stamp minorities with a badge of inferiority."[118] Legal scholars have noted that Thomas's opinion in *Grutter v. Bollinger* "open[ed] the door to an argument that these stigmatizing criteria should be understood as a form of discrimination that is cognizable and remediable at law."[119] It is for this reason that techniques such as decoupled classifiers[120] are likely to run afoul of legal protections against discrimination. Only in very limited circumstances

116. A FEW GOOD MEN (Columbia Pictures 1992).

117. *See* Johnson v. California et al., 543 U.S. 499 (2005); Brown v. Bd. of Educ., 347 U.S. 483 (1954) (holding racial segregation to be harmful in and of itself, regardless of its results); Shaw et al. v. Reno, 509 U.S. 630, 643 (1993) (asserting that classification on the basis of race "threaten[s] to stigmatize individuals by reason of their membership in a racial group").

118. *See* Adarand Constructors, Inc. v. Peña, 515 U.S. 200, 241 (1995). Additionally, there may be fear of harms to recipients of affirmative action for university admissions after graduation. *See* Sampath Kannan, Aaron Roth & Juba Ziani, *Downstream Effects of Affirmative Action*, ARXIV, Aug. 29, 2018, https://arxiv.org/pdf/1808.09004.pdf.

119. *See* Tomiko Brown-Nagin, *The Transformative Racial Politics of Justice Thomas?: The Grutter v. Bollinger Opinion*, 7 U. PA. J. CONS. L. 787, 806 (2005).

120. *See* Dwork et al., *supra* note 92.

can two different standards (including classifiers or decision thresholds) be legally applied on the basis of a sensitive attribute.[121] This harm-from-stigma argument is further substantiated by the social psychology literature on stigma and affirmative action. A number of studies have demonstrated that affirmative action can result in the beneficiary being stigmatized as incompetent both by the beneficiary and by others.[122] Differential treatment—even "positive" treatment—seems to still result in harm.

Even if the courts did not apply this standard of racial classifications as a cognizable harm to cases involving machine learning systems, other types of harm might still be established. For example, we typically assume that the plaintiff in a disparate treatment claim will be an underrepresented minority, but this does not have to be the case. Title VII provides protection from disparate treatment on the basis of race for everyone, not just for particular races. Moreover, a machine learning system that mitigates bias against one group might result in harm for a different group. The question then becomes one of fairness *toward whom*? Consider a system that explicitly uses race to make employment decisions that mirror racial population proportions from U.S. Census data. One might think that given the nature of this system, a disparate treatment claim could not be brought under Title VII because no one was harmed, aside from the aforementioned psychological harm of racial classifications. But there are additional complexities. First, although mirroring racial population proportions may seem facially "fair," it is only one fairness definition among many, and it could certainly be interpreted as unfair by those who prefer a more meritocratic definition of fairness. Second, mirroring racial population proportions may not be a desirable end goal. For example, historical injustices may require overrepresentation rather than proportional representation in certain circumstances. This is a thorny issue that falls outside the scope of this paper, but it is worth noting, nonetheless.

---

121. *See Bauer v. Lynch*, 812 F.3d 340, 351 (4th Cir. 2016) (holding it permissible under Title VII for the FBI to instate gender-normed physical fitness admissions tests with two different push-up standards for men and women).

122. *See* Madeline E. Heilman, *Affirmative Action: Some Unintended Consequences for Working Women*, 16 RES. ORG. BEHAV. 125 (1994); Madeline E. Heilman, Caryn J. Block & Jonathan A. Lucas, *Presumed Incompetent? Stigmatization and Affirmative Action Efforts*, 77 J. APPLIED PSYCHOL. 536 (1992); Madeline E. Heilman, Michael C. Simon & David P. Repper, *Intentionally Favored, Unintentionally Harmed? Impact of Sex-Based Preferential Selection on Self-Perceptions and Self-Evaluations*, 72 J. APPLIED PSYCHOL. 62 (1987); Pamela Stanush, Winfred Arthur, Jr. & Dennis Doverspike, *Hispanic and African American Reactions to a Simulated Race-Based Affirmative Action Scenario*, 20 HISP. J. BEHAV. SCI. 3 (1998). For a nice review of the literature over the last several decades, *see also* David A. Harrison et al., *Understanding Attitudes toward Affirmative Action Programs in Employment: Summary and Meta-Analysis of 35 Years of Research*, 91 J. APPLIED PSYCHOL. 1013 (2006).

A case highlighting some of these complexities in the context of university admissions is currently being brought by Students for Fair Admission, a group that includes a number of Asian-Americans who assert that Harvard University discriminated against Asian-American applicants.[123] This case is still being litigated and involves a number of difficult legal questions, as highlighted by the Department of Justice's recent statement of interest opposing Harvard's motion for summary judgement.[124] As an informative example, consider a simpler machine learning hypothetical. Suppose that the proportion of high-achieving Asian-American Harvard applicants is higher than the nationwide proportion of Asian-Americans according to U.S. census data. Suppose also that admissions decisions are made by a machine learning system that explicitly considers race in order to mirror racial population proportions from U.S. census data.[125] This definition of fairness is intended to protect underrepresented minorities. However, some high-achieving Asian-Americans are likely to be denied admission with the decision being made by a system that explicitly considers race. Of course, it would be up to a court to decide, but this example seems to involve a harm that could constitute grounds for a disparate treatment claim.[126] This is not a clear-cut issue, but the example serves to illustrate that strategy 2 can still cause harm and therefore violate Title VII, even when the system is behaving in a "fair" fashion.

B. Proxies, Disparate Treatment, and Circumstantial Evidence

The appeal of strategy 3 is that the applicants receive identical treatment by the system. This is relevant because a paradigmatic case of disparate treatment involves the intentional differential application

123. *See* Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll. *(Harvard Corp.)*, 261 F. Supp. 3d 99 (D. Mass. 2017); Anemona Hartocollis & Stephanie Saul, *Affirmative Action Battle Has a New Focus: Asian-Americans*, N.Y. TIMES (Aug. 2, 2017), https://www.nytimes.com/2017/08/02/us/affirmative-action-battle-has-a-new-focus-asian-americans.html. Related issues may also arise under the College Board's proposal to introduce an "adversity score" into the SAT scoring process; however, after much criticism, this proposal has now been withdrawn. *See* Douglas Belkin, *SAT to Give Students 'Adversity Score' to Capture Social and Economic Background*, WALL ST. J. (May 17, 2019), https://www.wsj.com/articles/sat-to-give-students-adversity-score-to-capture-social-and-economic-background-11557999000; Anemona Hartocollis, *SAT 'Adversity Score' Is Abandoned in Wake of Criticism*, N.Y. TIMES (Aug. 27, 2019), https://www.nytimes.com/2019/08/27/us/sat-adversity-score-college-board.html.

124. *See* Statement of Interest in Opposition to Defendant's Motion for Summary Judgment, *Students for Fair Admissions, Inc. v. President & Fellows of Harvard Coll. (Harvard Corp.)*, No. 1:14-cv-14176-ADB (D. Mass. 2018).

125. It is important to note that this is only one fairness definition among many, and its "fairness" in such a situation is far from certain. We use it in this example merely because it has some facial validity, and some admissions officers are likely to consider it as plausibly fair.

126. This reemphasizes the deep difficulty in delineating fairness definitions that can be universally applied. A rich literature on this difficulty is emerging (*see supra* Part II.A-B.), but much work remains to be done.

of a set policy to similarly situated individuals.[127] For example, consider an employer that has a policy to reject any applicant with five D's on her university transcript. Suppose that the employer enforced this policy for one applicant and did not enforce it for another applicant of a different race. This differential application of the policy constitutes disparate treatment. But this concern does not apply to strategy 3, where the same machine learning system or "policy" is applied to all the applicants, thereby ensuring that they receive identical treatment.

However, the uniform application of a machine learning system—even one that is blinded to race—does not necessarily insulate against disparate treatment claims because the system might inappropriately use proxies for race. A decision maker who selects applicants on the basis of race and a decision maker who selects applicants by inferring their race from their zip code are doing "exactly the same [thing], only [the latter uses] two steps rather than one. This too is a form of disparate treatment." [128] The potential use of proxies, then, brings us back to circumstantial evidence. Thus far, we have primarily focused on direct evidence of disparate treatment involving automated decision makers. This is because a machine learning system's explicit consideration of race is readily detected if race is included as a feature, hence direct evidence is likely most pertinent.[129] But if the system includes only proxies for race, rather than race itself, then a disparate treatment claim must instead rely on circumstantial evidence.

Consider the following egregious policy, where a proxy for race plays an important role in motivating the behavior that leads to differential treatment: reject all applicants from zip codes with median income below the poverty line. Application of this policy does not require access to the race of the applicants, but, given that socioeconomic status is a proxy for race, uniform application of this policy will surely lead to biased decisions. An employer that uses such a policy will almost certainly open itself up to disparate impact claims, but we note that it may also be subject to disparate treatment liability. The fact that the employer uniformly applies the same policy to the applicants is no defense because circumstantial evidence still provides a way to demonstrate disparate treatment without any evidence so blatant as a defendant stating, "I'm [taking this adverse action] because you're in

127. *See supra* note 47.

128. James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 176 (2017). This paper is also cited and briefly discussed in an earlier version of Lipton, Chouldechova & McAuley, *supra* note 96. Zachary C. Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity? [v1]*, in 32ND Conference on Neural Information Processing Systems, ARXIV, Nov. 19, 2017), https://arxiv.org/pdf/1711.07076v1.pdf.

129. *See supra* Part II.C.

a protected group."[130] Instead, the burden-shifting *McDonnell-Douglas* framework is typically used.[131] Continuing with this example, the plaintiff would argue that she was rejected unfairly by demonstrating that the defendant hired a similarly situated applicant, the only difference between them being their race. The defendant then must articulate a "clear and reasonably specific" non-discriminatory reason for acting in such a manner.[132] Finally, the burden is then placed back on the plaintiff, who must demonstrate that the defendant's proffered reason is merely pretext for her true discriminatory intent (e.g., by identifying "weaknesses, implausibilities, . . ." in the defendant's reason).[133] In this example, the plaintiff should have no problem demonstrating that the defendant's use of zip code is pretext.

However, applying the *McDonnell-Douglas* framework to strategy 3 yields a different result.[134] This difference is due to the way that strategy 3 uses race during training to mitigate bias, which becomes relevant in the third step of the burden-shifting framework. We emphasize that the argument here is not that strategy 3 sneakily avoids creating circumstantial evidence; rather, the argument is that strategy 3 is not engaging in disparate treatment, so an attempt to present circumstantial evidence via the *McDonnell-Douglas* framework will ultimately fail.

Suppose that the plaintiff has (ostensibly) produced evidence of a similarly situated individual (step one) and that the defendant using strategy 3 has produced some clear and reasonably specific reason for selecting one applicant over the other (step two). Producing such a reason should not be particularly difficult for the defendant. Provided the defendant has access to the system's internals, she can analyze the features (e.g., education, grades) that played a central role in the system's decisions about the applicants, including features that are proxies for race.

Recall that in the third step of the framework, the plaintiff must demonstrate that the defendant's proffered reason is false.[135] There are several ways that the plaintiff can accomplish this. First, she can demonstrate that the defendant deviated from her policy regarding the decision-making process. In the context of strategy 3, this is unlikely to be successful because the same machine learning system or "policy"

130. *See* Sheehan v. Donlen Corp., 173 F.3d 1039, 1044 (7th Cir. 1999) (holding that requiring such blatant confession would effectively "cripple enforcement of the . . . discrimination laws.").

131. *See supra* Part II.D.b.

132. Tex. Dep't of Cmty. Affairs v. Burdine, 450 U.S. 248, 258 (1980).

133. Brooks v. Cty. Comm'n of Jefferson Cty., 446 F.3d, 1160, 1163 (11th Cir. 2006).

134. It is also possible to apply the *Arlington Heights* framework (*see supra* Part II.D.b.) to strategy 3, which would produce the same result.

135. *See Supra* Part II.D.b.

is applied to all the applicants. Second, the plaintiff can produce evidence that the defendant's reason is a "*post hoc* fabrication."[136] This too is unlikely to be effective because the system's internals (e.g., its structure, features, and parameters) were all established prior to or during training, and hence settled upon before deployment. Finally, the plaintiff can identify "weaknesses, implausibilities, inconsistencies, incoherencies, or contradictions" in the defendant's reason.[137] But so long as the defendant used a reasonable definition of fairness in strategy 3, the plaintiff will struggle to identify such issues because these would all be indications of a poor definition of fairness. In other words, the plaintiff must show that the defendant's reason is "unworthy of credence," but the use of a reasonable definition of fairness in strategy 3 necessarily makes it worthy of credence, otherwise that definition would not be reasonable.[138] Therefore, the system's use of proxies for race would only function as pretext if an unreasonable definition of fairness were used in strategy 3.[139]

Of course, the above analysis relies on the plaintiff producing evidence of a similarly situated individual. In practice, it is unlikely that there exists an individual whose only difference from the plaintiff is her race. Counterfactuals do not cut so cleanly because attributes such as race are intimately tied to a constellation of other attributes. Therefore, a difference in race would likely mean a cascade of other differences, too.[140] Moreover, if two individuals are truly identical except for their race, then there can be no proxies for race available to the machine learning system, and so there cannot be any differential treatment. This means that there must be some other difference between the plaintiff and the similarly situated individual. If the features accounting for this difference are clearly relevant to employment, then this would constitute a legitimate non-discriminatory reason, even if the features are proxies for race. Hence, there would be no differential treatment, and therefore no need to appeal to affirmative action.[141]

---

136. Fuentes v. Perskie, 32 F.3d 759, 764 (3d Cir. 1994).

137. *Brooks*, 446 F.3d at 1163.

138. United States Postal Serv. Bd. of Governors v. Aikens, 460 U.S. 711, 716 (1983) (quoting Tex. Dep't. Cmty. Affairs v. Burdine, 450 U.S. 248, 256 (1980)).

139. This defense can also be leveraged by strategy 2, but there is little reason to do so because strategy 2 still expressly classifies applicants on the basis of their race, thereby providing direct evidence of disparate treatment, whereas strategy 3 avoids this problem.

140. *See* Issa Kohler-Hausmann, *Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination*, 113 NW. U. L. REV. 1163 (2019).

141. We note that there is an edge case in which the features accounting for the other difference between the plaintiff and the similarly situated individual are not relevant to employment and are proxies for race. This situation would appear to be, in the words of Grimmelmann and Westreich, disparate treatment "in two steps rather than one." *See* Grimmelmann, *supra* note 128. In this situation, the defendant would have to appeal to affirmative action, meaning that strategy 3 would have no (legal) advantage over strategy 2. This means that taking a "kitchen sink" approach to including features (viz. including features that may not be relevant to employment) when using strategy 3 could result in disparate treatment.

Therefore, the success or failure of strategy 3 at avoiding disparate treatment claims hangs on the particular definition of fairness that is used. This means that strategy 3 is not quite a panacea because there is much disagreement about fairness definitions in the literature and no easy answers.[142] Some definitions may work better in criminal justice than in employment or university admissions. Additionally, there are questions regarding how courts will react to the various definitions that have been proposed, and which definition(s) will be deemed acceptable in which circumstances. Courts have successfully established fairness definitions in the past, such as the EEOC's four-fifth's rule which asserts that "[a] selection rate for any race, sex, or ethnic group which is less than four-fifths . . . of the rate for the group with the highest rate will generally be regarded . . . as evidence of adverse impact."[143] This concern over the fairness definitions used in the context of machine learning systems is a novel issue still to be navigated. That said, our analysis hopefully shows that if a reasonable definition can be agreed upon, strategy 3 does not appear to violate Title VII, without needing to appeal to affirmative action.

V. CONCLUSION: WHAT NOW?

"Considering the social context when designing technical solutions will lead to better—and more fair—sociotechnical systems."
Andrew D. Selbst et al.[144]

Disparate treatment poses a difficult problem for mitigating bias in the context of machine learning. Although strategy 1 appears to comply with Title VII, it runs a risk of yielding a biased system. Meanwhile, strategy 2 is more effective at mitigating bias but violates Title VII by explicitly considering race during deployment. However, strategy 3 is just right. Strategy 3 effectively mitigates bias from a machine learning perspective and also appears to avoid disparate treatment claims because it does not use racial classifications for decision-making. Moreover, its consideration of race during training can be analogized to legally accepted auditing procedures, with the main difference being that strategy 3 considers the sensitive attribute *ex ante* rather than *ex post*. However, the impact of using strategy 3 depends greatly

142. *See supra* Part II.A-B. for discussion of the challenges inherent to selecting an agreed-upon fairness definition.

143. Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D) (2015) (which also asserts that meeting or failing to meet the four-fifth standard is not dispositive, as "[s]maller differences in selection rate may nevertheless constitute adverse impact [and] . . . [g]reater differences in selection rate may not constitute adverse impact."). Also, note that this rule does not violate the Equal Protection Clause by setting a quota, so it is unlikely that the fairness definitions frequently used in machine learning would be deemed quota-filling either.

144. Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, 15 ACM Conference on Fairness, Accountability, and Transparency (FAT*) (2019).

on the specific definition of fairness that is used. An unreasonable or poor definition could result in strategy 3 failing to mitigate bias, thereby opening up the possibility of disparate treatment claims.

Although we focus on discrimination in employment and Title VII of the Civil Rights Act of 1964, this is only one example illustrating a more general tension between laws designed to regulate human behavior and the problem of stretching them to apply to machines. We propose strategy 3 as a way to balance this problematic tension. We note that strategy 3 involves innovative work in machine learning (viz. the development of disparate learning processes) and creative legal analysis (viz. analogizing strategy 3 to legally accepted auditing procedures like those conducted by the EEOC). We contend that this multipronged approach is necessary not just for the success of strategy 3 in particular, but for the success of any solution to the general problem of stretching human laws to apply to machines. This is because any such solution must be sociotechnical in nature, drawing on work in both machine learning and law.[145]

145. For a discussion of the importance of these solutions being interdisciplinary, *see* Selbst, *supra* note 144.