HIRING BY ALGORITHM: PREDICTING AND PREVENTING DISPARATE IMPACT

Ifeoma Ajunwa,[*] **Sorelle Freidler,**[**] **Carlos Scheidegger,**[***] **Suresh Venkatasubramanian**[****]

ABSTRACT

Major advances in machine learning have encouraged corporations to rely on Big Data and algorithmic decision making with the presumption that such decisions are efficient and impartial. In this Essay, we show that protected information that is encoded in seemingly facially neutral data could be predicted with high accuracy by algorithms and employed in the decision-making process, thus resulting in a disparate impact on protected classes. We then demonstrate how it is possible to repair the data so that any algorithm trained on that data are more likely to make non-discriminatory decisions. Since this data modification is done before decisions are applied to any individuals, this process can be applied without requiring the reversal of decisions. We make the legal argument that such assessments and data modifications should be mandated as an anti-discriminatory measure. And akin to Professor Ayres' and Professor Gerarda's Fair Employment Mark, such data repair that is preventative of disparate impact would be certifiable by teams of lawyers working in tandem with software engineers and data scientists. Finally, we anticipate the business necessity defense that such data modifications could degrade the accuracy of algorithmic decision-making. While we find evidence for this trade-off, we also found that on one data set it was possible to modify the data so that despite previous decisions having had a disparate impact under the four-fifths standard, any subsequent decision-making algorithm was necessarily non-discriminatory while retaining essentially the same accuracy. Such an algorithmic "repair" could be used to refute a business necessity defense by showing that algorithms trained on modified data can still make decisions consistent with their previous outcomes.

[*] Assistant Professor of Law, University of the District of Columbia, PhD Candidate, Columbia University, Affiliate, Data and Society

[**] Assistant Professor in Computer Science, Haverford College, Fellow, Data and Society

[***] Assistant Professor in Computer Science, Arizona State University

[****] Assistant Professor in Computer Science, Utah University

TABLE OF CONTENTS

INTRODUCTION

Automation has been lauded as the yellow brick road to progress and societal harmony.[1] Within the last few decades, more than just our daily routines have become automated,[2] indeed, automation has been steadily creeping into many areas that were once thought solely reserved to human judgment and reasoning.[3] Consider the newest trend in automation – the hiring process. Whereas once, an applicant could rely on his or her interpersonal skills to make a favorable first impression on the hiring manager, these days the hiring algorithm is the initial hurdle to clear. A recent swell of start-ups — including HireVue,[4] Gild,[5] Entelo,[6] Textio,[7] Doxa,[8] Jobaline,[9] and GapJumpers[10] — are innovating new ways to

[1] "We should not be afraid of AI[Artificial Intelligence]. Instead, we should hope for the amazing amount of good it will do in the world. It will saves (sic) lives by diagnosing diseases and driving us around more safely. It will enable breakthroughs by helping us find new planets and understand Earth's climate. It will help in areas we haven't even thought of today." – *Mark Zuckerberg,* Facebook Post, January 27, 2016, 8:37AM, available at*: https://www.facebook.com/zuck/posts/10102620559534481*

[2] "While computerization has been historically confined to routine tasks involving explicit rule-based activities, algorithms for big data are now rapidly entering domains reliant upon pattern recognition and can readily substitute for labour in a wide range of non-routine cognitive tasks." Carl Benedikt Frey and Michael A. Osborne, *The Future of Employment: How Susceptible are Jobs to Computerisation?* Available at: *http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf*

[3] For example, clinical psychology was once thought of as a profession that was particularly resistant to computer automation. "But some research suggests that people are more honest in therapy sessions when they believe they are confessing their troubles to a computer, because a machine can't pass moral judgment." Derek Thompson, "A World Without Work," *The Atlantic*, July/August 2015, available at: http://www.theatlantic.com/magazine/archive/2015/07/world-without-work/395294/

[4] "Video Interviewing, Video Coaching and Predictive Analytics - Recruit and Coach the World's Best Teams". Official Company website: http://www.hirevue.com

[5] "The Gild Hiring Platform: Gild fundamentally transforms your entire talent acquisition and hiring process. By using data science, consumer-friendly technologies, and predictive analytics, Gild makes finding, engaging, and hiring the right talent simple and smart." Official Company website: www.guild.com

[6] "Build great teams. Our software helps you find, qualify, and engage with top talent." Official Company website: www.entelo.com

[7] Hire Better Candidates, Faster. Official Company website: https://textio.com

[8] "Doxa casts a light on the best places to work." http://doxascore.com

[9] "jobaline is the leading recruiting solution optimized for hourly jobs." Official Company website: www.jobaline.com

"[10] Discover talent the Voice way: employers use our technology to find untapped talent using blind auditions." https://www.gapjumpers.me. It is important to add that unlike other automated hiring programs, GapJumpers advocates "blind auditions" wherein the candidates are tested for talent without their identity or other identifying characteristics (such as school pedigree) made known to recruiters/hiring managers. Our data repair is a automate hiring. Their claim is that hiring algorithms are more effective and efficient than any human manager.[11] Some also claim that such decisions are inherently non-discriminatory,[12] we challenge this claim. For one, GapJumpers method relies heavily on a skills test.

"GapJumpers and its client create a list of skills required for the job, then design a relevant test that the applicant completes online. The first piece of information the hiring company sees is applicants' scores, and, based on those, it selects candidates to interview. Only then does it see their names and résumés."[13]

The problem with these attempts is that there is no well-established way to determine whether they work; hiding identities of candidates does little to protect against discrimination via a proxy variable (such as, for example, cultural references in the "relevant test"). It is also important to consider that hiring by algorithm may be no transient fad as traditional well-established headhunting firms like Korn Ferry start incorporating algorithms as part of business procedure.[14]

The word "algorithm" has gained prominence in research and writing in the past three decades.[15] Derived from the name of a Persian mathematician, Al-Khwarizmi,[16] the word and the process of calculation it stands for, have escaped the cloisters of the discipline of mathematics. Rather, with advancements in computing technologies, and the capacity for rapid mining of Big Data, algorithms now pervade our daily lives and

more complete method of doing this as it would remove all "societal noise" from the available data about candidates.

[11] Claire Cain Miller, "Can an Algorithm hire Better than a Human?," The *New York Times*, June 25, 2015, Available at: http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[12] Jobaline CEO Luis Salazar as quoted in "Now Algorithms Are Deciding Whom To Hire, Based On Voice," NPR All Tech Considered, March 23, 2015. Available at: http://www.npr.org/sections/alltechconsidered/2015/03/23/394827451/now-algorithms-are-deciding-whom-to-hire-based-on-voice

[13] Claire Cain Miller, "Is Blind Hiring the Best Hiring?" Feb. 25, 2016, Available at: http://www.nytimes.com/2016/02/28/magazine/is-blind-hiring-the-best-hiring.html?_r=1

[14] Sarah Green Carmichael, Hiring C-Suite Executives by Algorithm, Harvard Business Review, April 06, 2015. https://hbr.org/2015/04/hiring-c-suite-executives-by-algorithm

[15] Google InGram shows the usage of the word "algorithm" beginning in the 1800s and rapidly growing from the 1980s.

[16] Knuth, Donald (1979). *Algorithms in Modern Mathematics and Computer Science* (PDF). Springer-Verlag. ISBN 0-387-11157-3.

exert influence over many impactful decisions.[17] Consider that an algorithm decides all of the following: the answer to a search one conducts online,[18] the best romantic prospects provided by a dating website,[19] what advertisements one sees during a visit to a given website,[20] one's creditworthiness,[21] whether or not one should be considered a suspect for crime,[22] and whether or not one is qualified for a job.[23]

Although algorithms are touted as efficient[24] and impartial,[25] that they now wield much influence on the outcomes of our lives has become a source of concern.[26] Previously, much of this concern have centered on

[17] See, Neil M. Richards and Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST LAW REVIEW 393 (2014), noting that "large datasets are being mined for important predictions and often suprising insights." Id. at 393.

[18] See for example, Latanya Sweeney, Discrimination in Online Ad Delivery, 56 COMM. OF THE ACM 44 (2013); detailing a study in which a search of names associated with African-Americans returned results featuring advertisements for arrest records as a result of machine learning by Google's Ad algorithm.

[19] Thorin Klosowski, "Here's How OkCupid Uses Math to Find Your Match," Lifehacker.com Available at: http://gizmodo.com/5984005/heres-how-okcupid-uses-math-to-find-your-match

[20] http://lifehacker.com/5994380/how-facebook-uses-your-data-to-target-ads-even-offline, explaining how Facebook uses your likes (in addition to those of your friends) to tailor ads or target your for specific advertisements.

[21] Frank Pasquale, BLACK BOX SOCIETY, 2015, detailing how data brokers sell the seemingly obscure personal information of American individuals to companies that determine credit worthiness.

[22] Andrew G. Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 1137 (2015), noting that, although in the past, determining who was a suspect was a more individualized process, police can now rely on large datasets to make probabilistic determinations of criminal activity.

[23] Claire Miller, Can an Algorithm hire Better than a Human?, The *New York Times*, June 25, 2015Available at: http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html "Established headhunting firms like Korn Ferry are incorporating algorithms into their work, too."

[24] *See,* Harry L. Lewis and Christos H. Papadimitriou, *The Efficiency of Algorithms*, available at: http://www.cs.princeton.edu/~arora/pubs/lewispapa.pdf

[25] *See*, for example, the initial assumptions about algorithms used for predictive policing: "After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. 'This program had absolutely nothing to do with race… but multi-variable equations,' argues Goldstein. Gillian Tett, "Mapping Crime – or Stirring Hate?" *Financial Times*. *But see*, Quentin Hardy, "Using Algorithms to Determine Character," *New York Times*, July 26, 2015, , available at: http://bits.blogs.nytimes.com/2015/07/26/using-algorithms-to-determine-character/?_r=0 (noting that algorithms used to judge character seem really to be distinguishing between social classes).

[26] *See, e.g,* Bruce Schneier, DATA AND GOLIATH: THE HIDDEN BATTLE TO COLLECT YOUR DATA AND CONTROL YOUR WORLD, (2015); Dan J. Solove, *Introduction: Privacy Self Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1881 (2013).

issues of privacy harms[27] including breaches of confidentiality,[28] but recently, scholars have started to consider the disparate impact of algorithms.[29] In the Article, *Big Data's Disparate Impact*, authors Solon Barocas and Andrew Selbst detail how data mining by algorithms may be employed deliberately or unintentionally to replicate discriminatory results that maintain an unjust status quo.[30] The authors note that data-mining algorithms can "reproduce existing patterns of discrimination, inherit the prejudice of prior decision-makers, or simply reflect the widespread biases that persist in society."[31] And all this without being explicitly designed to do so. Indeed, because of the aura of accuracy and impartiality that is imbued to algorithms employed in data-mining processes,[32] the disparate results returned could serve to exacerbate "existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment."[33]

The disparate racial impact of algorithms is exacerbated by factors that we call "societal noise." Those factors present themselves as neutral information or even as merely predictive of the outcome being sought, but we argue that those factors are, in actuality, reflective of the racial history of the United States — for example, education and housing (which are the consequence of racial discriminatory practices)[34] or credit scores (high

[27] *See, e.g.,* Paul Ohm, S*ensitive Information*, 88 S. CAL. L. REV. 1125-1196 (2015), asserting that categories of information deserve special protection because of privacy harms attached to such information.

[28] *See, e.g*, Ifeoma Ajunwa, *Genetic Testing Meets Big Data: Torts and Contract Law Issues*, 75 OHIO ST. L. J. 1225 (2014).

[29] See, Solon Barocas and Andrew Selbst, *Big Data's Disparate Impact*, 104 CAL. LAW REVIEW, Vol. 104, 3-4 (Forthcoming 2016). Available at SSRN: http://ssrn.com/abstract=2477899

[30] See, Solon Barocas and Andrew Selbst, *Big Data's Disparate Impact*, 104 CAL. LAW REVIEW, Vol. 104, 3-4 (Forthcoming 2016). Available at SSRN: http://ssrn.com/abstract=2477899 3-4

[31] See, Solon Barocas and Andrew Selbst, *Big Data's Disparate Impact*, 104 CAL. LAW REVIEW, Vol. 104, 3-4 (Forthcoming 2016). Available at SSRN: http://ssrn.com/abstract=2477899

[32] danah boyd and Kate Crawford, Six Provocations for Big Data, *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society,* September 2011, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

[33] See, Solon Barocas and Andrew Selbst, *Big Data's Disparate Impact*, 104 CAL. LAW REVIEW, Vol. 104, 4 (Forthcoming 2016). Available at SSRN: http://ssrn.com/abstract=2477899

[34] "The G.I. Bill deliberately left the distribution and implementation of federal education and housing benefits to universities, private banks, realtors, and white homeowners' associations, all of whom discriminated openly and pervasively against blacks." Juan F. Perea, *Doctrines of Delusion: How the History of the G.I. Bill and Other Inconvenient Truths Undermine the Supreme Court's Affirmative Action Jurisprudence*, 75 U. PITT. L.

scores are enabled by the inherited wealth that had been previously denied generations of African-Americans).[35]

Perhaps the most pernicious effect of the disparate impact of Big Data, however, is when it comes to employment. Recently, the New York Times[36] revealed that a new study[37] by Carnegie Mellon University found that Google search ads showed more high-paying executive jobs to people believed to be men conducting searches than to searchers believed to be women.[38] While that study revealed that algorithms may not be exempt from the biases that plague society, it, however, revealed little as to the cause of the bias, and further still, as to how to fix it.

This haplessness in regards to how to legally curtail the disparate effects of data-mining is of pressing concern given that algorithms are currently being lauded as the next revolutionary hiring tool; and, moreover, as a benevolent one with the power to solve the issues of sexism and racial bias in the workplace.[39] In this Essay, we argue that well settled legal doctrines that prohibit discrimination against job applicants on the basis of sex or race dictate an examination of how algorithms are employed in the hiring process with the specific goals of: 1) predicting whether such algorithmic decision-making could generate decisions having a disparate impact on protected classes; and 2) repairing input data in such a way as to prevent disparate impact from algorithmic decision-making.

I. DATA-MINING AND ADVERSE IMPACT

In this section, we detail the ways in which algorithms might be employed (intentionally or inadvertently) to contravene legal rules against discrimination based on a protected characteristic. We make the argument that while faulty algorithms may be blamed for these discriminatory results, we must not overlook the fact that algorithms are created and

REV. 583 (2014). Available at, http://lawecommons.luc.edu/cgi/viewcontent.cgi?article=1552&context=facpubs, See, e.g., Angela Onwuachi-Willig and Jacob Willig-Onwuachi, *A House Divided: The Invisibility of the Multiracial Family*, 44 HARV. C.R.-C.L. L. REV. __ (noting the still persistent and insidious housing discrimination and segregation in the United States).

[35] Sarah Ludwig, Credit Scores in America Perpetuate Racial Injustice: Here's How, *The Guardian,* available at: http://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why

[36] Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found. http://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

[37] http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdfHere

[38] http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdfHere

[39] https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct

maintained by humans. The responsibility for interrogating whether algorithms are returning truly accurate decisions rather than merely mimicking societal biases ultimately rests on the shoulders of the human architects of algorithms.

A. *The Fault in the Machine*

Part of the problem with algorithmic decision-making is that while the results might raise suspicions, it is far more difficult to prove malicious intent and, more onerous still, to pinpoint the nefarious act or acts[40]. Consider the case of Debra Wolverton, who was laid off from her retail sales job in June 2013. That day, she immediately inquired for work at some businesses on her way home in Austin, Texas and she was instructed to complete an online application.[41] Even though Wolverton completed numerous online job applications, she rarely got a callback. Today, Wolverton remains without full-time work. Wolverton holds the belief that her resume was often dismissed by computer programs that cull jobless applicants who are older or who have been out of work a long time, but she has no way of proving it.[42]

Professionals who create hiring algorithms for companies support Wolverton's claims. Sheeroy Desai, co-founder and chief executive of Gild, developer of hiring software notes: "Every company vets its own way, by schools or companies on résumés…it can be predictive, but the problem is it is biased. They're dismissing tons and tons of qualified people."[43] Despite this awareness, hiring by algorithm seems to be the

[40] Frank Pasquale, *The Black Box Society: The Secret Algorithms that Control Money and Information*. Harvard University Press, January 2015. See also infra, noting that while a human might be able to understand (and also question) each step followed to reach a decision, algorithms are neither as transparent in the steps taken to reach a decision, nor do they possess the capacity to question the fairness of each step.

[41] Jeffrey Stinson, *Hiring Bias Against the Unemployed: Should There Be a Law?,* PewTrusts (Aug. 24, 2014) *http://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2014/08/25/hiring-bias-against-the-unemployed-should-there-be-a-law*.

[42] Jeffrey Stinson, *Hiring Bias Against the Unemployed: Should There Be a Law?,* PewTrusts (Aug. 24, 2014) *http://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2014/08/25/hiring-bias-against-the-unemployed-should-there-be-a-law*. If Wolverton's suspicions were substantiated, the companies involved would be liable for age discrimination in contravention of the Age Discrimination in Employment Act (ADEA).

[43] Claire Cain Miller, *Can an Algorithm Hire Better Than a Human?*, NYT (June 25, 2015) *http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html?_r=0*. Consider that vetting for job candidates by prestigious schools (with their expensive tuition, past history of racial bias and legacy admissions) would have a

wave of the future. Several start-up companies are designing algorithms to replace recruitment managers and to automate various hiring procedures, and those algorithms are marketed on the claim that such algorithms can perform more effectively and efficiently than people can.[44] Even established headhunting firms like Korn Ferry have started incorporating algorithms to carry out their job duties.[45]

Note that we do not claim that creators of algorithms conceal a malevolent intent to discriminate against protected classes. Rather, like Frankenstein's monster, algorithms are wont to escape their masters' clutches and behave in unpredictable ways. Yet, try as they might to disavow their creation, the creators of algorithms owe a duty to create situations in which the good behavior of their algorithms are reasonably assured.

Let us consider a basic example of a task that can be solved through data mining. Take a human-resources department at a company looking to hire new employees. There are a limited number of job openings to be filled out, and so some selection process must take place: some applicants will be hired and others will be turned down. The goal of the HR department is to fill out as many openings it can, while maximizing the odds that the hired employees will be successful at their positions. The institution might use one of two broad strategies. The traditional strategy is to employ *hiring officers*, and a modern strategy is to use *data mining.*[46]

In the first scenario, the hiring officer will take, in one way or another, applicant qualifications into account in their decision process. The important aspect to note in this scenario is that the officer's judgment is subjective: although the criteria for hiring might have been determined elsewhere, the decision on how to apply these criteria and the responsibility for doing so lies with the officer. In this scenario, the potential for discrimination exists and is well documented, either through explicit disparate treatment[47] or disparate impact.[48]

disproportionate impact on racial minorities and those from the lower socio-economic strata.

[44] Gild's marketing materials ask: "Does your Hiring Platform have Smart Sourcing? Candidate Recommendations? CRM? ATS? 100 Million Profiles? Business Intelligence? Availability Ratings? Data Refresh? Ours does." http://www.gild.com. Entelo's materials ask: "Drowning in resumes? Check out our new inbound recruiting solution" http://www.entelo.com. Last accessed Feb 28 2016.

[45] Sarah Green Carmichael, *Hiring C-Suite Executives by Algorithm*. https://hbr.org/2015/04/hiring-c-suite-executives-by-algorithm

[46] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

[47] See as an example of direct discrimination, the disparate treatment as seen in the racially restrictive covenants of the 1920s *Understanding Fair Housing*, U.S.

Now consider a contrast with data mining. Here, the financial institution creates a large collection of past examples of *pairs* of job applications and employee evaluations. This collection will serve as *training data.*[49] Data mining refers to using computer programs that use the training data to construct decision rules. The hope of the data-mining enterprise is that there is enough information in the training data such that the rules that were automatically derived *generalize well*: informally, the rule should work well for *future* applications, in the sense that hiring officers should try and hire those applicants who will succeed, and try not to hire those who will not.[50] Roughly speaking, any given data-mining procedure works by finding the *best* set of decision rules among a large set of candidate rules. As we will see below–and crucially–different considerations about what to consider "best" can yield rules with drastically different behavior. In addition–and just as crucially–we will see, just like human activity does, that automated algorithms present the potential for discrimination.

When it comes time to apply these rules during hiring, they must have access to information about the job applicant. But since they will be used before we can know whether the applicant would have been successful or not, a decision rule must decide whether or not to offer a position *with foresight*: without having access to the eventual result of its decision and based only on its knowledge of past example employees. In general, information about past examples is divided into two broad classes: *attributes* and *outcomes*. Attributes constitute information that would have been knowable ahead of making the decision. In the case of the hypothetical HR institution, the attributes might include applicant information such as credit history, income level, educational attainment, and sex. Outcomes, on the other hand, constitute information about what happened after the decision. This can mean simply whether the employee receives favorable performance reviews after being hired.

In this scenario, the decision of whether or not to hire an applicant appears to no longer lie with the HR officer–all of the information used by the data mining procedure comes from the training data–but the matter of assigning responsibility is different than that of determining if discrimination has happened.

Commission on Civil Rights, Clearinghouse Publication 42, February 1973 (http://www.law.umaryland.edu/marshall/usccr/documents/cr11042.pdf)

[48] Jack M. Guttentag and Susan M. Wachter, *Redlining and Public Policy* (New York: New York University, 1980), wherein redlining of neighborhoods had a negative impact on the minority populations that lived there.

[49] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[50] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

We can already see one way in which data mining can go wrong: the choice of training examples we provide to a data-fitting procedure can have a profound difference on the produced decision rules.[51] In addition, the set of attributes that we expose to a data mining procedure can also influence the produced model.[52] If data mining is given data that is skewed one way, then, the produced rules cannot help but replicate the same skew with the training data. If some of the problems with the training data include reproducing society's structural biases, there exists a large risk, then, for the algorithm to return results that discriminate against protected classes.

For example, earlier we mentioned the study[53] that found that Google search ads showed more high-paying executive jobs to people believed to be men conducting searches than to searchers believed to be women.[54] One explanation for this result, other than deliberate sexism on the part of Google engineers, is that the algorithm used to conduct searches was trained on data that is reflective of the currently existing structural sexism of the C-suite and that the data reflects that more men are employed in higher paying jobs than women. The algorithm did not (and could not) interrogate the reasons *why* this was so, or if this *ought to* be so, rather, it concluded that, *it is so* and proceeded to return results to match that conclusion by showing higher paid jobs to men rather than women.

There is another way in which data mining can go wrong. This specific mode of failure comes from an issue alluded to earlier: the choice that analysts have in how to define the "best" model. Consider a thought experiment in which a data mining procedure is used to classify people. The way in which this classification is carried out is unimportant for the thought experiment, but consider a procedure that could achieve a "correct" classification in 98 out of every 100 cases. Although by itself this might look like a good outcome, consider that 2% of the U.S. population is Native American,[55] and the data-mining algorithm could achieve "98% accuracy" by misclassifying every single Native American person. That is, the rule we chose could achieve "98% accuracy" by performing every single classification of a Native American person

[51] Avrim L. Blum and Pat Langley. "Selection of relevant features and examples in machine learning." *Artificial intelligence* 97.1 (1997): 245-271.

[52] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." *The Journal of Machine Learning Research* 3 (2003): 1157-1182.

[53] http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf

[54] http://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf

[55] Humes, Karen, Nicholas A. Jones, and Roberto R. Ramirez. *Overview of race and Hispanic origin, 2010*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2011.

incorrectly. In other words, naive measures of accuracy do not take into account the effect of the decision rule on specific groups of people.

In our research[56] (and other authors have done so as well[57]), we propose using a different measure of accuracy, one that does take into account the effect of a data-mining rule on different groups.[58] What we propose is to only consider data mining algorithms that assess their accuracy as measured separately on legally protected classes. We call these measures *class-conditioned measures*.

For example, Title VII forbids racial discrimination.[59] Still using our hypothetical classifier, we can interpret a disparity of 100% accuracy on the non Native American subpopulation contrasted with 0% accuracy on Native American subpopulation as discrimination. This is evidence, then, that when taking discrimination into account, the data-mining procedure itself needs to be repurposed: the definition of *best model* must change. Consider now one simple change. Instead of choosing the rule considering the accuracy over the entire population, imagine that the data-mining procedure were to 1) measure the model's accuracy for the Native American subpopulation and the non Native American subpopulation, and 2) use the arithmetic mean of these two values. Let us now examine the rule that we previously thought was best. Even if the rule predicts outcomes of non Native American applicants with 100% accuracy, the fact that it predicts Native American applicants with 0% accuracy means that the model's overall class-conditioned accuracy is only 50%.

With this new measure in mind, consider now an alternative rule, one that can predict with 80% accuracy on either subpopulation above. Under the old measure, this data-mining rule would not be considered the better of the two; under the new measure, it would. Intuitively, what accounts for the difference is that a high class-conditioned accuracy means that the model's performance must have been at least somewhat good on most of the sub-populations of interest.

In the previous example, we used the arithmetic mean between the model's accuracy in each possible value of the legally protected class. The

[56] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[57] For a good survey, see Romei, Andrea, and Salvatore Ruggieri. "A multidisciplinary survey on discrimination analysis." The Knowledge Engineering Review 29.05 (2014): 582-638. Section 10 contains the computer science-specific work.

[58] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Section 4.1, p. 261-262.

[59] Title VII, Civil Rights Act of 1964, (Pub. L. 88-352)

EEOC's rule, on the other hand, uses the *ratio* between these values.[60] Specifically, if the probability of a positive outcome for the protected class is less than 80% of the probability for the unprotected class, the procedure is said to exhibit disparate impact.[61] The four-fifths rule, like class-conditioned error measures, also ensures that decisions on minority classes will not receive low importance simply because the protected class is outnumbered by the majority.

Determining disparate impact in data mining, then, is a relatively simple matter of evaluating existing data-mining algorithms using the EEOC rule. Of course, practically speaking, the problem we encounter in trying to ascertain disparate impact in existing data-mining systems is that it might be hard to legally obtain the full outcomes of these systems, and companies might have an incentive not to make their proprietary algorithms visible[62]. On the other hand, if it were possible to ensure that disparate impact cannot happen in a data-mining algorithm, then companies would be able to protect themselves from disparate impact claims, without having to disclose their proprietary information. This is the crux of our technical proposal: a technique that makes it difficult for data-mining procedures to create disparate impact.

[60] "The agencies have adopted a rule of thumb under which they will generally consider a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5ths) or eighty percent (80%) of the selection rate for the group with the highest selection rate as a substantially different rate of selection." U.S. Equal Employment Commission, *Adoption of Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*, Federal Register, March 2nd, 1979, available at: http://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html

[61] U.S. Equal Employment Commission, *Adoption of Questions and Answers To Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures*, Federal Register, March 2nd, 1979, available at: http://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html This "4/5ths" or "80%" rule of thumb is not intended as a legal definition, but is a practical means of keeping the attention of the enforcement agencies on serious discrepancies in rates of hiring, promotion and other selection decisions. For example, if the hiring rate for whites other than Hispanics is 60%, for American Indians 45%, for Hispanics 48%, and for Blacks 51%, and each of these groups constitutes more than 2% of the labor force in the relevant labor area (see Question 16), a comparison should be made of the selection rate for each group with that of the highest group (whites). These comparisons show the following impact ratios: American Indians 45/60 or 75%; Hispanics 48/60 or 80%; and Blacks 51/60 or 85%. Applying the 4/ 5ths or 80% rule of thumb, on the basis of the above information alone, adverse impact is indicated for American Indians but not for Hispanics or Blacks.

[62] For example, Nicole Wong in her role as Google Inc's Associate General Counsel, has stated that "Google avidly protects every aspect of its search technology from disclosure". See https://googleblog.blogspot.com/2006/02/response-to-doj-motion.html

B. *Exposing the Mechanical Turk*[63]

An important feature of algorithms is that they tend to obscure the role of the human – the final result is attributed solely to the machine. Consider that the proponents of machine learning tend to downplay or deny the role of the human mastermind. For example, "The Mechanical Turk" also known as "the chess Turk" was a chess-playing machine constructed in the late 18th century.[64] Although the Mechanical Turk was presented as an automaton chess-playing machine that was capable of beating the best human players, the secret of the machine was that it contained a human man, concealed inside its chambers.[65] The hidden chess master controlled the machine while the seemingly automated machine beat notable statesmen like Napoleon Bonaparte and Benjamin Franklin at chess.[66] Thus, the Mechanical Turk operated on obfuscation and subterfuge and sought to reserve the glory of the win to the machine.[67]

Modern day algorithms operate in ways similar to the Mechanical Turk in that the human decisions behind the creation of algorithms operated by businesses are generally considered trade secrets that are jealously guarded and protected from government oversight.[68] But it is important to recognize that while algorithms remove the final decision from a human entity, humans must still make the initial decisions as to what data to train the algorithm on and as to what factors are deemed relevant or irrelevant.[69] Even more importantly, the decisions for what

[63] The technical, algorithmic solutions suggested in this section are based heavily on previous work by three of the authors: Section 4, "Computational Fairness," of Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[64] Tom Standage (2002-04-01). *The Turk: The Life and Times of the Famous 19th Century Chess-Playing Machine*. Walker. ISBN 978-0-8027-1391-9.

[65] Ricky Jay, "The Automaton Chess Player, the Invisible Girl, and the Telephone", *Jay's Journal of Anomalies*, vol. 4 no. 4, 2000.

[66] Tom Standage (2002-04-01). *The Turk: The Life and Times of the Famous 19th Century Chess-Playing Machine*. Walker. ISBN 978-0-8027-1391-9.

[67] See, Frank Pasquale, BLACK BOX SOCIETY, 2015 (arguing that algorithms operate on obfuscation). Conversely, Amazon's Mechanical Turk program does the opposite. The program allows businesses or individual clients to assign human intelligence tasks, that is, tasks that are difficult or impossible for machines to complete (like sorting photographs, writing product descriptions, completing surveys, etc.) to humans. Amazon explicitly bans the use of automated bots to complete such tasks.

[68] Pasquale, Frank. "Restoring Transparency to Automated Authority." *Journal on Telecommunications and High Technology Law, Vol. 9, No. 235, 2011.*

[69] Even when automated feature selection methods are used, the final decision to use or not use the results, as well as the choice of feature selection method and any fine-tuning of its parameters, are choices made by humans. For more on feature selection see, e.g.,

data is important in the training data – decisions that are then matched as closely as possible by the algorithm -- were also made by humans.[70]

Now, with an understanding of the many ways that discrimination can manifest in machine-learned decisions, we turn to the question of how such discrimination can be avoided *in advance*. In order to make such determinations and assuming that the algorithmic decisions will be made by machine learning algorithms trained on data, in order to use the method described here we require access to the training data.[71] This training data should include all information that will be given to the machine learning algorithm in order to make its decision, e.g., a job applicant's GPA, their previous work experience, etc., as well as the protected class status of each individual and the resulting decision (e.g., "hire" or "no hire").[72] From this training data, and *without access to the algorithm*, it is possible to determine if the algorithm *could* make discriminatory decisions under the four-fifths rule.[73]

Before we describe this testing procedure, let's discuss the restrictions that it operates under. First, it assumes no access to the decision-making algorithm that is trained based on the data.[74] There are multiple important reasons for this assumption. The algorithm might be proprietary or, even if access is granted, so large or obtuse so as to render any manual examination of it impossible.[75] More importantly, however, even if examination of the algorithm were possible, without the training data it

James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: Springer, 2013.

[70] See, e.g., the way that hiring startup Jobaline verifies their technique by using the ratings that people listening give voice snippets of job candidates: Li, Ying, Jose D. Contreras, and Luis J. Salazar. "Predicting Voice Elicited Emotions." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[71] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See in Section 4, p. 261-263.

[72] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Section 4, p. 261-263.

[73] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Theorem 4.1, p. 262.

[74] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Trust model in Section 4, p. 261.

[75] Diakopoulos, Nicholas. "Algorithmic Accountability: On the Investigation of Black Boxes." Report by the Tow Center for Digital Journalism. http://towcenter.org/research/algorithmic-accountability-on-the-investigation-of-black-boxes-2

might appear neutral even if it is not.[76] Thus, it is preferable to test a decision-making algorithm via its training data.

Second, the procedure only determines the possibility of discriminatory outcome, not the certainty of it.[77] Thus, a certification that there is no disparate impact means that the data is safe to use without fear of disparate impact in the resulting decisions, even for any algorithm trained on this data. However, a certification of disparate impact may not result in a discriminatory outcome. While this might seem to present the possibility of over-reacting by labeling all training data as possibly discriminatory, coupled with the "repair" method presenting in the following section this procedure instead allows us to identify data that might pose a problem and then prevent the discriminatory decisions from happening.

Now we can discuss the procedure to identify potentially discriminatory training data. As described in the previous section, discriminatory effect is introduced to machine-learned decisions when attributes that are correlated with protected class status, but not explicitly linked, are used as a proxy to determine the outcome.[78] This same observation can be used in order to determine the possibility of discriminatory effect via this simple procedure.[79] The procedure to determine whether a data set has disparate impact for a given protected class involves an experiment to predict the protected class status from other attributes of the data set. The experiment works this way[80]: 1) if the prediction of the protected class from the remaining attributes of the data set has a large amount of error, any resulting decision will be non-discriminatory. 2) On the other hand, if the prediction of the protected class from the remaining attributes of the data set is highly accurate, a discriminatory decision could result.

[76] See the previous section for an examination of the ways these biases could appear, as well as the thorough treatment of this subject in Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." *Available at SSRN 2477899* (2014).

[77] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Section 4.2, p. 262-263.

[78] See also, Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." *Available at SSRN 2477899* (2014).

[79] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Algorithm in Section 4.2, p. 262-263.

[80] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Theorem 4.1, p. 262.

The proof follows this simple thought experiment[81]: Given a prediction of someone's protected class status, write it down. If the person is predicted to be a member of a historically advantaged group, give them a positive outcome (i.e., hire them). Otherwise, the person is a member of a historically disadvantaged group; give them a negative outcome (don't hire them). By making these maximally biased decisions based on the best predictions you have, the resulting decisions are the most discriminatory they could be.

This testing procedure provides for an interesting method of enforcement because it uses the results of a prediction algorithm.[82] As machine learning algorithms become more powerful, we might worry that methods of discrimination would become harder to detect. Yet by using these same machine-learning algorithms as part of the detection process,[83] we avoid being left in the dark as to the discriminatory effects of the algorithms.

The thought experiment also illuminates an important subtlety. We describe predicting an individual's protected class status and then *writing it down*. However, from an algorithmic perspective, the second step is not necessary.[84] A decision can be made based on the prediction without ever writing it down or explicitly storing it to memory. Had the decision been written down and then purposefully used as described to make a decision that might have been clear disparate treatment. Yet, as the law stands, this same procedure when the prediction is not written down is not disparate treatment. Disparate impact affords the only recourse.

C. Duty to Correct Algorithmic Bias

Corporations and organizations bear a legal duty to correct algorithmic bias; and this duty is not mitigated by a lack of intent to discriminate or even a lack of awareness that an algorithm is producing biased results. Ignorance is never bliss when it comes to the law. It is well settled law that a facially neutral practice, which is found to be discriminatory in effect, is

[81] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the definition of a purely biased mapping in Section 4.1, p. 262.

[82] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Algorithm in Section 4.2, p. 262-263.

[83] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Algorithm in Section 4.2, p. 262-263.

[84] See the discussion of Proxies in Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." *Available at SSRN 2477899* (2014).

in contravention of the law.[85] While the history of antidiscrimination law reveals that the court previously mandated a showing of animus,[86] even for laws that have a disproportionate impact,[87] the law has evolved to recognize disparate impact as discriminatory harm, even absent a showing that the purpose of the challenged practice was to discriminate.[88] Whereas, previously, a plaintiff who alleged not purposeful discrimination, but rather "disparate impact" and/or disproportionate

[85] *See* Griggs v. Duke Power Co., 401 U.S. 424 (1971) (which established that pursuant to Title VII, an employer must provide a business necessity justification for the use of a test that has a disparate impact on a protected class. Holding also concludes that with Title VII, Congress meant to correct discriminatory impact and not solely overt discrimination.). In 1991, Congress amended Title VII to codify into law the "disparate impact test" established by *Griggs*. *But see* Ricci v. DeStefano, 557 U.S. 557 (2009) (in which Justice Ginsburg's dissent raises the issue of whether the holding in *Ricci* has effectively overruled *Griggs* and in which Justice Scalia implies in his concurring opinion that Title VII's disparate impact provision is unconstitutional).

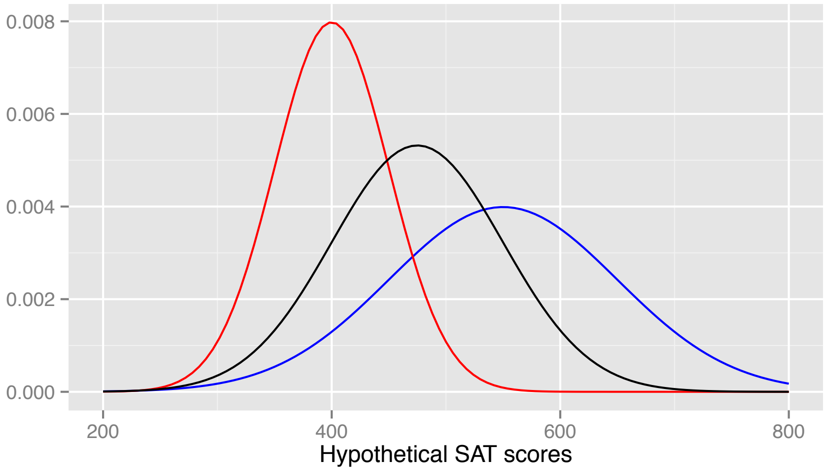
[86] *See* Washington v. Davis, 426 U.S. 229 (1976) (finding that laws that have a racially discriminatory impact but which do not have a racially discriminatory purpose are not unconstitutional). *See* Village of Arlington Heights v. Metro. Hous. Dev. Corp., 429 U.S. 252 (1977) (establishing the disparate impact test wherein the challenging party bears the burden of demonstrating that the law in question: 1) affects a protected class in greater proportion, and 2) was created with the intent or purpose to discriminate against the protected class). *See also* McClesky v. Kemp, 481 U.S. 279 (1987) (holding that racially discriminatory impact of death penalty as shown by comprehensive study is not enough to overturn verdict without a showing of a racially discriminatory purpose.

[87] See for example, *Yick Wo v. Hopkins.*[87] In that case, the court struck down a San Francisco ordinance that sought to curtail the operation of laundries in wooden buildings, and which disproportionately negatively affected people of Chinese descent as 95% of the city's 320 laundries were operated in wooden buildings and two-thirds of those wooden laundry buildings were owned by Chinese immigrants.[87] Of course, it must be noted that the reach of *Yick Wo'*s precedent was limited. Even after *Yick Wo* in 1886, the Court in *Plessy v. Ferguson*[87] upheld laws that discriminated against African Americans by asserting a "separate but equal" standard that allowed for legal segregation until that standard was overturned by the *Brown v. Board of Education*[87] case in 1954.

[88] Most antidiscrimination laws now have disparate impact clauses. *See e.g.,* Title VII, specifically, Title VII expressly prohibits employers from using any "particular employment practice that causes a disparate impact on the basis of race, color, religion, sex, or national origin." 42 U.S.C. § 2000e-2(k)(1)(A)(i).The Americans with Disabilities Act (ADA) also has a disparate impact clause detailing that the phrase "discriminate against a qualified individual on the basis of disability" includes neutral policies and practices "that have the effect of discrimination on the basis of disability." *See* 42 U.S.C. §§ 12112(a), (b)(3)(A). While the Age Discrimination in Employment Act (ADEA) does not have an explicit a disparate impact clause, the court has read disparate impact as one of its proscriptions. *See* Smith v. City of Jackson, Miss., 544 U.S. 228, 233–40 (2005) (interpreting 29 U.S.C. § 623(a)(2) of the ADEA by virtue of identical text in Title VII and the Supreme Court's interpretation of the Title VII provision in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971)

impact based on grounds other than race,[89] confronted a much more onerous burden of proof in court, [90] this has rapidly changed.[91]

II. SOLUTIONS TO DISCRIMINATION BY ALGORITHM



[89] *See* Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256 (1979) (a gender neutral law with an exclusionary impact on women is not unconstitutional absent a showing of discriminatory purpose. The court applied a two part test derived from the *Arlington* case: "When a statute gender-neutral on its face is challenged on the ground that its effects upon women are disproportionably adverse, a twofold inquiry is thus appropriate. The first question is whether the statutory classification is indeed neutral in the sense that it is not gender-based. If the classification itself, covert or overt, is not based upon gender, the second question is whether the adverse effect reflects invidious gender-based discrimination . . . ."); *But see* Griggs v. Duke Power Co., 401 U.S. 424 (1971) (which established that pursuant to Title VII, an employer must provide a business necessity justification for the use of a test that has a disparate impact on a protected class. Holding also concludes that with Title VII, Congress meant to correct discriminatory impact and not solely overt discrimination.). In 1991, Congress amended Title VII to codify into law the "disparate impact test" established by *Griggs*.

[90] *See e.g.* McClesky v. Kemp, 481 U.S. 279 (1987) (holding that racially discriminatory impact of death penalty as shown by comprehensive study is not enough to overturn verdict without a showing of a racially discriminatory purpose

[91] The Supreme Court recently read a disparate impact cause of action for the Fair Housing Act in *Texas Department of Housing v. Inclusive Communities Project*. Kali Borkoski, *Evening round-up: Texas Department of Housing v. Inclusive Communities Project*, SCOTUSblog (Jun. 25, 2015, 7:50 PM), http://www.scotusblog.com/2015/06/evening-round-up-texas-department-of-housing-v-inclusive-communities-project/

A. *Ex Machina: A Technological Solution*[92]

We demonstrate how it is possible to modify the data so that algorithms trained on the data are more likely to make non-discriminatory decisions under the four-fifths disparate impact rule.[93] Before describing this process in detail, we want to emphasize the goals and motivations of this technique. First, it is a technological solution to the issues raised by the hidden Turk, that is, the unseen human influence on the training data. By this, we mean that our goal is to remove the ability, exploited by the thought experiment described in our earlier section, to predict the protected class status of an individual. The process described here is designed to allow the resulting data to pass this experimental test by making it hard to accurately guess an individual's protected class status and, therefore, hard to discriminate against them by using it.[94] In other words, the goal is to *remove* any information related to the protected class status from the data. This is essentially a more complete method of arriving at "blind hiring," the hiring method that startups like GapJumpers are claiming to offer and which some experts believe will reverse structural biases in the hiring process.[95]

Second, while we will describe this *removal* or *repair* of the data as being enacted based on the protected class status of an individual, the technique will have the same effect on the data whether the class information used is the protected class status or related, strongly correlated, information such as socio-economic status.[96] For example, the University of Texas "Top 10% Rule" admissions plan automatically grants admission to the top ten percent of every Texas high school's class.[97] This process is

[92] The technical, algorithmic solutions suggested in this section are based heavily on previous work by three of the authors: Section 5, "Removing Disparate Impact," of Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[93] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 5, p. 263-265. Figure from p. 263.

[94] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Section 5, see especially Theorem 5.1 on p. 264.

[95] Claire Cain Miller, Is Blind Hiring the Best Hiring? Feb. 25, 2016, *The NY Times,* Available at: http://www.nytimes.com/2016/02/28/magazine/is-blind-hiring-the-best-hiring.html?_r=0

[96] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the introductory discussion at the beginning of Section 5, p. 263.

[97] Texas House Bill 588, 75th Legislature. Education Code.

similar to the repair procedure we will describe in that, since high schools in Texas are segregated by race and socio-economic status,[98] the high school attended acts as a proxy for race as protected class status. Thus, as we will explain in more detail later, the process can be used in a way that is facially neutral.

Finally, we will call the process a *repair* of the data since the goal is to remove the systemic bias inherent in the data. We think of this systemic bias as being noise that detracts from the decision-making goal that the data is being used to answer, noise that should be removed (while information relevant to the decision-making process is left intact) so that better decisions can be made. Importantly, unlike quota systems or point systems, if there is no systemic bias present in the data, the repair will do nothing.[99]

Now we can explain our proposed repair of the data. The process operates on the data per-attribute, so if the data contains information about applicants' SAT scores, GPA, and years of experience, this process will occur once for each of those attributes.[100] It begins by considering the distribution of the attribute when conditioned on the applicants' protected class status (or the proxy variable being used to represent protected class status).[101] To illustrate, we will use SAT scores as an example attribute and sex as an example protected class.[102] Then the process repairing the SAT scores begins by finding the group of scores achieved by men and the group of scores achieved by women. These scores are divided into quantiles so that, e.g., there is a group of men who are known to have the top five percent of men's SAT scores and a group of women who have the top five percent of women's SAT scores. These quantiles are grouped so that the top scores for the men and the top scores for the women are in the same group. The median score of this group is then applied to everyone in

[98] Marta Tienda and Sunny Niu. 2004. "Capitalizing on Segregation, Pretending Neutrality: College Admissions and the Texas Top 10% Law."

[99] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 5, p. 263 (the Algorithm): when the distributions conditioned on the protected class status are the same, the median of those distributions will also be identical.

[100] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 5, p. 263.

[101] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. p. 263.

[102] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Example taken from Figure 1, Section 5, p. 263.

the group. See, for example, the hypothetical SAT scores figure above. Suppose that the blue curve represents the distribution of women's scores and the red curve shows men's scores, then the black curve shows the resulting (combined) distribution of scores.

In general, our procedure repairs the data attribute by attribute, by ensuring that the per-attribute distributions, when conditioned on the protected class status, are the same. We can mathematically establish[103] that under this repair procedure, disparate impact with respect to the protected attribute that can be inferred from a single other attribute is removed. In cases where discrimination with respect to the protected attribute is hidden across multiple attributes, the repair may not be able to provably eliminate the entirety of the disparate impact,[104] though experimental results have shown that in practice, the repair successfully eliminates all discriminating signals.[105]

Let us consider the effects of this repair process. The top scoring women and men receive the same SAT scores, so, e.g., the top scoring man will never be moved ahead of the top scoring woman. Note also that the repair process was applied relative to the given scores (per attribute), so if there was no difference in the distributions of scores between men and women, there will be no effect on the data. In other words, the repair guarantees that there will be no disparate impact (under the four-fifths rule) against *any* group, not just the historically disadvantaged group.[106] Thus, the procedure is valid within the Civil Rights Act of 1991.[107]

Another important point is that the repair is only applied to the attributes used to make the final decision.[108] It is not applied to the given outcomes, the yes or no decisions, on which a decision-making procedure

[103] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Theorem 5.1, p. 264.

[104] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. On p. 263, the repair procedure is described in terms of a single attribute Y, and Theorem 5.1 only applies to this case.

[105] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Section 6, Figure 3, p. 267. Note that in this figure, all data sets are repaired to have no disparate impact (shown as a DI value of 1 on the figure).

[106] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. Theorem 5.1, p. 264.

[107] Civil Rights Act of 1991 - Pub. L. 102-166

[108] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the Algorithm, Section 5, p. 263.

will be learned. The procedure under which an applicant will receive a yes or no decision will be determined based on the repaired data. Thus, this process only obscures the information related to protected class – it forces the learning algorithm to use other information related to the original outcome to learn a pattern that can be applied to future applicants. Additionally, since this data modification is done before decisions are applied to any individuals, this process can be applied without requiring the reversal of decisions. It does not run afoul of the *Ricci v. DeStefano* decision.[109]

Let us go back to some of the goals and motivations of this technique with this procedural understanding. First, we mentioned that our goal is to remove the "societal noise" inherent in the data while leaving intact the information relevant to the classification task. By conditioning on the protected class status and maintaining the full distribution of data (see, e.g., the black curve in the Hypothetical SAT Scores example) in the resulting repaired information, we aim to ensure that information identifying someone as a top scorer is maintained.[110] Why do we consider the removed information to be "societal noise?" It is important to understand that any potential separation in distributions between, e.g., sexes or races, is due to societal factors that are heavily correlated with race and sex. For example, researchers have shown that SAT scores are biased against African American test takers in comparison to white test takers *of the same ability*.[111] Thus, if raw SAT scores are used to determine the outcome of an automated decision process, the results will include this racially biased noise. In the case of race, this includes historical racial discrimination resulting in, for example, substandard educational facilities as filtered through the property tax that sustains school districts.[112] In the case of sex, this sexism, both covert and overt which results in assumption about the intellectual abilities of women and the resulting stereotype threat that tends to have a real effect on the

[109] 557 U.S. 557 (2009). The facts of the case show that the employers invalidated a test that would determine who is promoted after the test had already been administered.

[110] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the definition of a repair that strongly preserves rank in Section 5, p. 263.

[111] Santelices, Maria Veronica, and Mark Wilson. "Unfair treatment? The case of Freedle, the SAT, and the standardization approach to differential item functioning." *Harvard Educational Review* 80.1 (2010): 106-134.

[112] Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANNU. REV. SOCIOL. 181 (2008).

outcomes of standardized tests taken by women.[113] Thus, we believe these differences in scores are not "due to" race and sex, but are rather "societal noise" that is irrelevant to the decision-making goal and should be removed from the data.

Second, this repair can be performed in a facially neutral way. The question the procedure is really asking of the person applying it is: "What information is irrelevant to the decision being made?" In the University of Texas Top 10% Rule admissions process described earlier,[114] the irrelevant information is the high school the student went to. One interpretation of the motivation for this rule is that, since students don't have control over which public high school they are assigned to attend, this information should not be used to determine whether they are admitted to the University of Texas. Similarly, the user of the repair procedure considers if sex, race, or socio-economic status should validly be used to determine the outcome of the decision-making procedure for hiring. If these factors are believed to be irrelevant to job performance, they should be chosen as the feature to repair the data with respect to when determining who to hire.

Finally, we emphasize that the method for determining the final decision is created *after* the repair is done.[115] Why is this important? The repair is not a "job qualifying" procedure; merely having undergone the repair does not guarantee the applicant a job. The determination of the job qualifying procedure happens *after* the repair. The repair eliminates the noise of factors not relevant to the job qualifications and then the choice of how to determine what factors are relevant to the job happens afterwards based on the data that remains. Thus, this data repair complies with the decisions in both *Fisher v. Texas*[116] and *Ricci v. DeStefano*[117] as the protected category is not a criteria used for the final hiring decision.

[113] Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35,* 4-28..

[114] Texas House Bill 588, 75th Legislature. Education Code.

[115] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 5, p. 263 – 264, where the Algorithm works to change D, the training data for the machine learning algorithm, before the algorithm is trained.

[116] 507 U.S. ____(2013), following the precedent of *Grutter v. Bollinger* (20013), allowing that diversity in student admissions is a valid goal for universities to pursue and outlining the guidelines to pursue that objective.

[117] 557 U.S. 557 (2009). The facts of the case show that the employers invalidated a test that would determine who is hired. Conversely, the data repair does not determine who is hired, it merely allows for a more expansive pool of candidates, in keeping with the recognized American objective of equal opportunity in employment.

B. *Ex Fida Bona: A Legal Solution*

But we do not merely offer a technological solution to the issue of "runaway discriminatory algorithms," rather, like previous scholars we recognize the need for "the establishment of ethical principles and best practices that guide government agencies, corporate actors,"[118] and other organizations. We argue that it is good business practice for corporations and other organizations to routinely evaluate the algorithms they rely on for their hiring and retention decisions for disparate impact concerns. This means that there is a role for qualified labor and employment lawyers who are also versed in data science to advice or work in-house with the human resources departments of large corporations and the engineers of hiring algorithms to ensure that the companies are not inadvertently contravening the spirit of Title VII and other civil rights laws that have been promulgated to grant true equal opportunity for employment to all American citizens.

In fact, we consider this such good business practice, that we argue that such self-study and publicized reports by large business corporations should be mandated as a matter of law as part of an anti-discriminatory measure aimed at ensuring that hiring policies, as aided by advancements in computing, are not inadvertently excluding from hiring consideration otherwise well qualified members of protected classes. Such data repair could also be expanded to allow the chance for the employment of classes of disadvantaged individuals that are not yet considered protected classes, such as the formerly incarcerated, veterans, the long-time unemployed, and mothers, as these are classes that are easily culled without discretion by algorithms set up to check for conviction status or periods of absence in the work history.

III. ANSWERING BUSINESS CONCERNS

In this section, we acknowledge and address business concerns that "data repair" might impact the accuracy of algorithmic decision-making and also the concern that mandating self-audits of hiring decisions that have already been delegated to a computer program negates the efficiency gains of automation in business.

[118] Neil M. Richards and Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST LAW REVIEW 393 (2014)

A. *Maintaining Accuracy in Decision-Making*[119]

We anticipate the business necessity defense that such data modifications could degrade the accuracy of algorithmic decision-making. While we find this trade-off to be true in general,[120] on one data set it was possible to modify the decisions made from well under the four-fifths disparate impact threshold to parity in decisions while the accuracy remained essentially stable.[121] Such an algorithmic "repair" could be used to refute a business necessity defense by showing that algorithms trained on modified data can still make good decisions.

Suppose that a credit granting agency would like to determine whether an adult makes more or less than $50,000 per year based on census information about them such as their educational level, marital status, occupation, and capital gain amount. One might expect that sex would be highly predictive in this task, since women make 74 cents for every dollar that men make.[122] Indeed, if a Gaussian Naïve Bayes classifier is used to predict the income level of adults, with data taken from the 1994 U.S. Census,[123] the results have disparate impact with respect to sex under the 80% rule.[124] In fact, the results are well under 80% at 57%.[125] The accuracy, or percentage of adults for whom the classifier correctly predicts whether the income is more or less than $50,000 per year, is 79.6% on this data when sex is not used explicitly, but is allowed to be used implicitly via a proxy variable, to make the decision.[126]

[119] The experimental analysis in this section is based heavily on previous work by three of the authors: Section 6, "Experiments," of Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[120] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the German credit data set and the Ricci data set in Section 6, p. 265-267.

[121] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the Adult Income data set in Section 6, p. 265-267.

[122] http://www.payscale.com/data-packages/gender-pay-gap

[123] http://archive.ics.uci.edu/ml/datasets/Adult

[124] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See the Adult Income data set in Section 6, p. 265-267.

[125] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 6, Figure 2, p. 266, the GNB Adult Income data point furthest to the left.

[126] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data*

If the agency is not allowed to use sex to make this decision, they might worry that the accuracy of their classification will go down. But using the repair method described earlier, on this data set that is essentially untrue. Again using a Gaussian Naïve Bayes classifier, the resulting accuracy decreases only to 79.1%.[127] The results are then well above the 80% threshold, at 99.7%.[128]

Thus, in this case, with essentially no loss in accuracy (only a half a percentage point[129]), it's possible to take a decision from discriminatory (well below the 80% threshold) to not (well above the 80% threshold). While there are many data sets for which this is not true (see, e.g., the decrease in accuracy when attempting to predict whether someone has good credit while not taking age into account in a set of German credit data[130]), in decision-making scenarios where it is possible to make non-discriminatory, accurate decisions, we claim that business necessity defenses should not hold.

B. *How Businesses Can Fulfill The Duty to Protect*

Corporations cannot afford to blissfully abdicate control of sensitive decisions in hiring to algorithms trained via Big Data without proper oversight and monitoring of the disparate impact of the decisions that are being returned by that process. Professors Ian Ayres and Jennifer Gerarda Brown have developed a framework that could be taken by corporations to certify discrimination-free workplaces that comply with ENDA.[131] The professors have created a certifying mark (FE), which they call "the Fair

Mining. ACM, 2015. See Section 6, Figure 5, p. 268, the GNB Adult Income data point furthest to the left.

[127] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 6, Figure 5, p. 268, the GNB Adult Income data point furthest to the right.

[128] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See Section 6, Figure 2, p. 266, the GNB Adult Income data point furthest to the right.

[129] As described above, the original accuracy was 79.6% and the accuracy on the repaired data set is 79.1%, so the loss in accuracy is 0.5%. Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

[130] Feldman, Michael, et al. "Certifying and removing disparate impact." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015. See German Credit data in Section 6, p. 265 – 267.

[131] Ian Ayres and Jennifer Gerarda Brown, *Mark(et)ing Nondiscrimination: Privatizing ENDA with a Certification Mark*, 104 MICHIGAN LAW REVIEW 1639 (2006)

Employment Mark."[132] Ayres and Gerarda describe their regulatory framework as follows:

"By signing the licensing agreement with us, an employer gains the right (but not the obligation) to use the mark and in return promises to abide by the word-for-word strictures of ENDA. Displaying the mark signals to knowing consumers and employees that the company manufacturing the product or providing the service has committed itself not to discriminate on the basis of sexual orientation."[133]

Similarly, there is a need for auditory organizations or businesses comprising lawyers and software engineers/data scientists who would audit the hiring algorithms employed by corporations and organizations, who could conduct a data repair to ensure that the algorithms the companies are using do not result in disparate impact, and who could then subsequently certify corporations as being free from algorithmic disparate impact.

Furthermore, while much can be said about the ethical benefits of a diverse workforce, particularly in regards to reducing economic inequality and its negative effects and reflecting the truth of equal participation in a liberal economy, we do not think it too crass to note that diversity is also a benefit to business.[134] Our proposed solution goes beyond merely being an antidiscrimination tool, rather, we believe that for companies, it could serve as a self-imposed audit of the potential for innovation and a repair to achieve higher creativity, better decision-making, and greater innovation.

CONCLUSION

Proponents of algorithms have favorably likened its workings to that of an oracle. For those adherents, the algorithm is all knowing and will infallibly provide the answers the intrepid pilgrim seeks. This represents a simplistic version of the opaque nature of an oracle. Consider the ur-Oracle, the Oracle of Delphi. The Oracle spoke veraciously but it was truth that was wrapped in layers, spun in riddle, and with many streams of

[132] Ian Ayres and Jennifer Gerarda Brown, *Mark(et)ing Nondiscrimination: Privatizing ENDA with a Certification Mark*, 104 MICHIGAN LAW REVIEW 1639, 1643 (2006)

[133] Ian Ayres and Jennifer Gerarda Brown, *Mark(et)ing Nondiscrimination: Privatizing ENDA with a Certification Mark*, 104 MICHIGAN LAW REVIEW 1639, 1643 (2006)

[134] Sheen S. Levine, Evan P. Apfelbaum, Mark Bernard, Valerie L. Bartelt, Edward J. Zajac, and David Stark. "Ethnic diversity deflates price bubbles" *PNAS* 2014 111 (52) 18524-18529; published ahead of print November 17, 2014, doi:10.1073/pnas.1407301111 (sociological research showing that diverse teams make better decisions and are more innovative)

interpretation. The pilgrim who did not take the time to fully interrogate the Oracle did so at her peril, departing with a seemingly simple answer that was highly vulnerable to misinterpretation. The same is true of algorithms. Although these computerized mathematical processes possess utility for corporations and organizations in the automation of hiring processes, we must continue to interrogate them to ensure that the answers we obtain and how we are interpreting those answers represent the whole truth and is in furtherance of our shared goal of a just and equal society.