Big Data's Disparate Impact

Solon Barocas* & Andrew D. Selbst**

Advocates of algorithmic techniques like data mining argue that these techniques eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data is frequently imperfect in ways that allow these algorithms to inherit the prejudices of prior decision makers. In other cases, data may simply reflect the widespread biases that persist in society at large. In still others, data mining can discover surprisingly useful regularities that are really just preexisting patterns of exclusion and inequality. Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

This Essay examines these concerns through the lens of American antidiscrimination law—more particularly, through Title

DOI: http://dx.doi.org/10.15779/Z38BG31

California Law Review, Inc. (CLR) is a California nonprofit corporation. CLR and the authors are solely responsible for the content of their publications.

* Postdoctoral Research Associate, Center for Information Technology Policy, Princeton University; Ph.D. 2014, New York University, Department of Media, Culture, and Communication. This research was supported in part by the Center for Information Technology Policy at Princeton University.

** Scholar in Residence, Electronic Privacy Information Center; Visiting Researcher, Georgetown University Law Center; Visiting Fellow, Yale Information Society Project; J.D. 2011, University of Michigan Law School. The authors would like to thank Jane Bambauer, Alvaro Bedoya, Marjory Blumenthal, Danielle Citron, James Grimmelmann, Moritz Hardt, Don Herzog, Janine Hiller, Chris Hoofnagle, Joanna Huey, Patrick Ishizuka, Michael Kirkpatrick, Aaron Konopasky, Joshua Kroll, Mark MacCarthy, Arvind Narayanan, Helen Norton, Paul Ohm, Scott Peppet, Joel Reidenberg, David Robinson, Kathy Strandburg, David Vladeck, members of the Privacy Research Group at New York University, and the participants of the 2014 Privacy Law Scholars Conference for their helpful comments. Special thanks also to Helen Nissenbaum and the Information Law Institute at New York University for giving us an interdisciplinary space to share ideas, allowing this paper to come about. Copyright © 2016 by Solon Barocas and Andrew Selbst. This Essay is available for reuse under the Creative Commons Attribution-ShareAlike 4.0 International License, http://creativecommons.org/licenses/by-sa/4.0/. The required attribution notice under the license must include the article's full citation information, e.g., "Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016)."

> *VII's prohibition of discrimination in employment. In the absence of a demonstrable intent to discriminate, the best doctrinal hope for data mining's victims would seem to lie in disparate impact doctrine. Case law and the Equal Employment Opportunity Commission's Uniform Guidelines, though, hold that a practice can be justified as a business necessity when its outcomes are predictive of future employment outcomes, and data mining is specifically designed to find such statistical correlations. Unless there is a reasonably practical way to demonstrate that these discoveries are spurious, Title VII would appear to bless its use, even though the correlations it discovers will often reflect historic patterns of prejudice, others' discrimination against members of protected groups, or flaws in the underlying data.*
>
> *Addressing the sources of this unintentional discrimination and remedying the corresponding deficiencies in the law will be difficult technically, difficult legally, and difficult politically. There are a number of practical limits to what can be accomplished computationally. For example, when discrimination occurs because the data being mined is itself a result of past intentional discrimination, there is frequently no obvious method to adjust historical data to rid it of this taint. Corrective measures that alter the results of the data mining after it is complete would tread on legally and politically disputed terrain. These challenges for reform throw into stark relief the tension between the two major theories underlying antidiscrimination law: anticlassification and antisubordination. Finding a solution to big data's disparate impact will require more than best efforts to stamp out prejudice and bias; it will require a wholesale reexamination of the meanings of "discrimination" and "fairness."*

Introduction 673
I. How Data Mining Discriminates 677
A. Defining the "Target Variable" and "Class Labels" 677
B. Training Data 680
1. Labeling Examples 681
2. Data Collection 684
C. Feature Selection 688
D. Proxies 691
E. Masking 692
II. Title VII Liability for Discriminatory Data Mining 694
A. Disparate Treatment 694
B. Disparate Impact 701
C. Masking and Problems of Proof 712
III. The Difficulty for Reforms 714
A. Internal Difficulties 715

1. Defining the Target Variable 715
2. Training Data 716
a. Labeling Examples 716
b. Data Collection 717
3. Feature Selection 719
4. Proxies 720
B. External Difficulties 723
Conclusion 729

INTRODUCTION

"Big Data" is the buzzword of the decade.[1] Advertisers want data to reach profitable consumers,[2] medical professionals to find side effects of prescription drugs,[3] supply-chain operators to optimize their delivery routes,[4] police to determine where to focus resources,[5] and social scientists to study human interactions.[6] Though useful, however, data is not a panacea. Where data is used predictively to assist decision making, it can affect the fortunes of whole classes of people in consistently unfavorable ways. Sorting and selecting for the best or most profitable candidates means generating a model with winners and losers. If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.

Although we live in the post–civil rights era, discrimination persists in American society and is stubbornly pervasive in employment, housing, credit, and consumer markets.[7] While discrimination certainly endures in part due to decision makers' prejudices, a great deal of modern-day inequality can be attributed to what sociologists call "institutional" discrimination.[8] Unconscious, implicit biases and inertia within society's institutions, rather than intentional

1. *Contra* Sanjeev Sardana, *Big Data: It's Not a Buzzword, It's a Movement*, FORBES (Nov. 20, 2013), http://www.forbes.com/sites/sanjeevsardana/2013/11/20/bigdata [https://perma.cc/9Y37-ZFT5].

2. Tanzina Vega, *New Ways Marketers Are Manipulating Data to Influence You*, N.Y. TIMES: BITS (June 19, 2013, 9:49 PM), http://bits.blogs.nytimes.com/2013/06/19/new-ways-marketers-are-manipulating-data-to-influence-you [https://perma.cc/238F-9T8X].

3. Nell Greenfieldboyce, *Big Data Peeps at Your Medical Records to Find Drug Problems*, NPR (July 21, 2014, 5:15 AM), http://www.npr.org/blogs/health/2014/07/21/332290342/big-data-peeps-at-your-medical-records-to-find-drug-problems [https://perma.cc/GMT4-ECBD].

4. *Business by Numbers*, ECONOMIST (Sept. 13, 2007), http://www.economist.com/node/9795140 [https://perma.cc/7YC2-DMYA].

5. Nadya Labi, *Misfortune Teller*, ATLANTIC (Jan.–Feb. 2012), http://www.theatlantic.com/magazine/archive/2012/01/misfortune-teller/308846 [https://perma.cc/7L72-J5L9].

6. David Lazer et al., *Computational Social Science*, 323 SCI. 721, 722 (2009).

7. Devah Pager & Hana Shepherd, *The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets*, 34 ANN. REV. SOC. 181, 182 (2008).

8. *Id.*

choices, account for a large part of the disparate effects observed.[9] Approached without care, data mining can reproduce existing patterns of discrimination, inherit the prejudice of prior decision makers, or simply reflect the widespread biases that persist in society. It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment.

Algorithms[10] could exhibit these tendencies even if they have not been manually programmed to do so, whether on purpose or by accident. Discrimination may be an artifact of the data mining process itself, rather than a result of programmers assigning certain factors inappropriate weight. Such a possibility has gone unrecognized by most scholars and policy makers, who tend to fear concealed, nefarious intentions or the overlooked effects of human bias or error in hand coding algorithms.[11] Because the discrimination at issue is unintentional, even honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions. Furthermore, because the mechanism through which data mining may disadvantage protected classes is less obvious in cases of unintentional discrimination, the injustice may be harder to identify and address.

In May 2014, the White House released a report titled *Big Data: Seizing Opportunities, Preserving Values* (Podesta Report), which hinted at the discriminatory potential of big data.[12] The report finds "that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace."[13] It suggests that there may be unintended discriminatory

9. *See* Andrew Grant-Thomas & john a. powell, *Toward a Structural Racism Framework*, 15 POVERTY & RACE 3, 4 ("'Institutional racism' was the designation given in the late 1960s to the recognition that, at very least, racism need not be individualist, essentialist or intentional.").

10. An "algorithm" is a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions. SOLON BAROCAS ET AL., DATA & CIVIL RIGHTS: TECHNOLOGY PRIMER (2014), http://www.datacivilrights.org/pubs/2014-1030/Technology.pdf [https://perma.cc/X3YX-XHNA]. Algorithms play a role in both automating the discovery of useful patterns in datasets and automating decision making that relies on these discoveries. This Essay uses the term to refer to the latter.

11. *See, e.g.*, Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 101 (2014) ("[H]ousing providers could design an algorithm to predict the [race, gender, or religion] of potential buyers or renters and advertise the properties only to those who [meet certain] profiles."); Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4 (2014) ("Because human beings program predictive algorithms, their biases and values are embedded into the software's instructions. . . ."); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1254 (2008) ("Programmers routinely change the substance of rules when translating them from human language into computer code.").

12. EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES (May 2014), http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf [https://perma.cc/ZXB4-SDL9].

13. *Id.* (introductory letter).

effects from data mining but does not detail how they might come about.[14] Because the origin of the discriminatory effects remains unexplored, the report's approach does not address the full scope of the problem.

The Podesta Report, as one might expect from the executive branch, seeks to address these effects primarily by finding new ways to enforce existing law. Regarding discrimination, the report primarily recommends that enforcement agencies, such as the Department of Justice, Federal Trade Commission, Consumer Financial Protection Bureau, and Equal Employment Opportunity Commission (EEOC), increase their technical expertise and "develop a plan for investigating and resolving violations of law in such cases."[15]

As this Essay demonstrates, however, existing law largely fails to address the discrimination that can result from data mining. The argument is grounded in Title VII because, of all American antidiscrimination jurisprudence, Title VII has a particularly well-developed set of case law and scholarship. Further, there exists a rapidly emerging field of "work-force science,"[16] for which Title VII will be the primary vehicle for regulation. Under Title VII, it turns out that some, if not most, instances of discriminatory data mining will not generate liability. While the Essay does not show this to be true outside of Title VII itself, the problem is likely not particular to Title VII. Rather, it is a feature of our current approach to antidiscrimination jurisprudence, with its focus on procedural fairness. The analysis will likely apply to other traditional areas of discrimination, such as housing or disability discrimination. Similar tendencies to disadvantage the disadvantaged will likely arise in areas that regulate legitimate economic discrimination, such as credit and insurance.

This Essay proceeds in three Parts. Part I introduces the computer science literature and proceeds through the various steps of solving a problem with data mining: defining the target variable, labeling and collecting the training data, using feature selection, and making decisions on the basis of the resulting model. Each of these steps creates possibilities for a final result that has a disproportionately adverse impact on protected classes, whether by specifying the problem to be solved in ways that affect classes differently, failing to recognize or address statistical biases, reproducing past prejudice, or considering an insufficiently rich set of factors. Even in situations where data miners are extremely careful, they can still effect discriminatory results with models that, quite unintentionally, pick out proxy variables for protected classes. Finally, Part I notes that data mining poses the additional problem of

14. *Id.* at 64 ("This combination of circumstances and technology raises difficult questions about how to ensure that discriminatory effects resulting from automated decision processes, whether intended or not, can be detected, measured, and redressed.").

15. *Id.* at 65.

16. Steve Lohr, *Big Data, Trying to Build Better Workers*, N.Y. TIMES (Apr. 20, 2013), http://www.nytimes.com/2013/04/21/technology/big-data-trying-to-build-better-workers.html [https://perma.cc/CEL2-P9XB].

giving data miners the ability to disguise intentional discrimination as accidental.

In Part II, the Essay reviews Title VII jurisprudence as it applies to data mining. Part II discusses both disparate treatment and disparate impact, examining which of the various data mining mechanisms identified in Part I will trigger liability under either Title VII theory. At first blush, either theory is viable. Disparate treatment is viable because data mining systems treat everyone differently; that is their purpose. Disparate impact is also viable because data mining can have various discriminatory effects, even without intent. But as Part II demonstrates, data mining combines some well-known problems in discrimination doctrines with new challenges particular to data mining systems, such that liability for discriminatory data mining will be hard to find. Part II concludes with a discussion of the new problems of proof that arise for intentional discrimination in this context.

Finally, Part III addresses the difficulties reformers would face in addressing the deficiencies found in Part II. These difficulties take two forms: complications internal to the logic of data mining and political and constitutional difficulties external to the problem. Internally, the different steps in a data mining problem require constant subjective and fact-bound judgments, which do not lend themselves to general legislative resolution. Worse, many of these are normative judgments in disguise, about which there is not likely to be consensus. Externally, data mining will force society to explicitly rebalance the two justifications for antidiscrimination law—rooting out intentional discrimination and equalizing the status of historically disadvantaged communities. This is because methods of proof and corrective measures will often require an explicit commitment to substantive remediation rather than merely procedural remedies. In certain cases, data mining will make it simply impossible to rectify discriminatory results without engaging with the question of what level of substantive inequality is proper or acceptable in a given context. Given current political realities and trends in constitutional doctrines, legislation enacting a remedy that results from these discussions faces an uphill battle. To be sure, data mining also has the potential to help reduce discrimination by forcing decisions onto a more reliable empirical foundation and by formalizing decision-making processes, thus limiting the opportunity for individual bias to affect important assessments.[17] In many situations, the introduction of data mining will be a boon to civil rights, even where it fails to root out discrimination altogether, and such efforts should be encouraged. Yet, understanding when and why discrimination persists in cases of data-driven decision making reveals important and sometimes troubling limits to the promise of big data, for which there are no ready solutions.

17. Tal Z. Zarsky, *Automated Prediction: Perception, Law, and Policy*, COMM. ACM, Sept. 2012, at 33–35.

I.
HOW DATA MINING DISCRIMINATES

Although commentators have ascribed myriad forms of discrimination to data mining,[18] there remains significant confusion over the precise mechanisms that render data mining discriminatory. This Part develops a taxonomy that isolates and explicates the specific technical issues that can give rise to models whose use in decision making may have a disproportionately adverse impact on protected classes. By definition, data mining is *always* a form of statistical (and therefore seemingly rational) discrimination. Indeed, the very point of data mining is to provide a rational basis upon which to distinguish between individuals and to reliably confer to the individual the qualities possessed by those who seem statistically similar. Nevertheless, data mining holds the potential to unduly discount members of legally protected classes and to place them at systematic relative disadvantage. Unlike more subjective forms of decision making, data mining's ill effects are often not traceable to human bias, conscious or unconscious. This Part describes five mechanisms by which these disproportionately adverse outcomes might occur, walking through a sequence of key steps in the overall data mining process.

A. Defining the "Target Variable" and "Class Labels"

In contrast to those traditional forms of data analysis that simply return records or summary statistics in response to a specific query, data mining attempts to locate statistical relationships in a dataset.[19] In particular, it automates the process of discovering useful patterns, revealing regularities upon which subsequent decision making can rely. The accumulated set of discovered relationships is commonly called a "model," and these models can be employed to automate the process of classifying entities or activities of interest, estimating the value of unobserved variables, or predicting future outcomes.[20] Familiar examples of such applications include spam or fraud detection, credit scoring, and insurance pricing. These examples all involve attempts to determine the status or likely outcome of cases under consideration based solely on access to *correlated* data.[21] Data mining helps identify cases of

18. Solon Barocas, *Data Mining and the Discourse on Discrimination*, PROC. DATA ETHICS WORKSHOP (2014), https://dataethics.github.io/proceedings/DataMiningandtheDiscourseOn Discrimination.pdf [https://perma.cc/D3LT-GS2X].

19. *See generally* Usama Fayyad, *The Digital Physics of Data Mining*, 44 COMM. ACM, Mar. 2001, at 62.

20. More formally, classification deals with discrete outcomes, estimation deals with continuous variables, and prediction deals with both discrete outcomes and continuous variables, but specifically for states or values *in the future*. MICHAEL J. A. BERRY & GORDON S. LINOFF, DATA MINING TECHNIQUES: FOR MARKETING, SALES, AND CUSTOMER RELATIONSHIP MANAGEMENT 8–11 (2004).

21. Pedro Domingos, *A Few Useful Things to Know About Machine Learning*, COMM. ACM, Oct. 2012, at 78–80.

spam and fraud and anticipate default and poor health by treating these states and outcomes as a function of some other set of observed characteristics.[22] In particular, by exposing so-called "machine learning" algorithms to examples of the cases of interest (previously identified instances of fraud, spam, default, and poor health), the algorithm "learns" which related attributes or activities can serve as potential proxies for those qualities or outcomes of interest.[23]

Two concepts from the machine learning and data mining literature are important here: "target variables" and "class labels." The outcomes of interest discussed above are known as target variables.[24] While the target variable defines what data miners are looking for, "class labels" divide all possible values of the target variable into mutually exclusive categories.

The proper specification of the target variable is frequently not obvious, and the data miner's task is to define it. To start, data miners must translate some amorphous problem into a question that can be expressed in more formal terms that computers can parse. In particular, data miners must determine how to solve the problem at hand by translating it into a question about the value of some target variable. The open-endedness that characterizes this part of the process is often described as the "art" of data mining. This initial step requires a data miner to "understand[] the project objectives and requirements from a business perspective [and] then convert[] this knowledge into a data mining problem definition."[25] Through this necessarily subjective process of translation, data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes.

Problem specification is not a wholly arbitrary process, however. Data mining can only address problems that lend themselves to formalization as questions about the state or value of the target variable. Data mining works exceedingly well for dealing with fraud and spam because these cases rely on extant, binary categories. A given instance either is or is not fraud or spam, and the definitions of fraud or spam are, for the most part, uncontroversial.[26] A computer can then flag or refuse transactions or redirect emails according to

22. *Id.*

23. *Id.*

24. COMM. ON THE ANALYSIS OF MASSIVE DATA ET AL., FRONTIERS IN MASSIVE DATA ANALYSIS 101 (2013), http://www.nap.edu/catalog.php?record_id=18374 [https://perma.cc/5DNQ-UFE4]. The machine learning community refers to classification, estimation, and prediction—the techniques that we discuss in this Essay—as "supervised" learning because analysts must actively specify a target variable of interest. *Id.* at 104. Other techniques known as "unsupervised" learning do not require any such target variables and instead search for general structures in the dataset, rather than patterns specifically related to some state or outcome. *Id.* at 102. Clustering is the most common example of "unsupervised" learning, in that clustering algorithms simply reveal apparent hot spots when plotting the data in some fashion. *Id.* We limit the discussion to supervised learning because we are primarily concerned with the sorting, ranking, and predictions enabled by data mining.

25. PETE CHAPMAN ET AL., CRISP-DM 1.0: STEP-BY-STEP DATA MINING GUIDE 10 (2000).

26. *See* David J. Hand, *Classifier Technology and the Illusion of Progress*, 21 STAT. SCI. 1, 10 (2006).

well-understood distinctions.[27] In these cases, data miners can simply rely on these simple, preexisting categories to define the class labels.

Sometimes, though, defining the target variable involves the creation of *new* classes. Consider credit scoring, for instance. Although now taken for granted, the predicted likelihood of missing a certain number of loan repayments is not a self-evident answer to the question of how to successfully extend credit to consumers.[28] Unlike fraud or spam, "creditworthiness" is an artifact of the problem definition itself. There is no way to directly measure creditworthiness because the very notion of creditworthiness is a function of the particular way the credit industry has constructed the credit issuing and repayment system. That is, an individual's ability to repay some minimum amount of an outstanding debt on a monthly basis is taken to be a nonarbitrary standard by which to determine in advance and all-at-once whether he is worthy of credit.[29]

Data mining has many uses beyond spam detection, fraud detection, credit scoring, and insurance pricing. As discussed in the introduction, this Essay will focus on the use of data mining in employment decisions. Extending this discussion to employment, then, where employers turn to data mining to develop ways of improving and automating their search for good employees, they face a number of crucial choices.

Like creditworthiness, the definition of a good employee is not a given. "Good" must be defined in ways that correspond to measurable outcomes: relatively higher sales, shorter production time, or longer tenure, for example. When employers mine data for good employees, they are, in fact, looking for employees whose observable characteristics suggest that they would meet or exceed some monthly sales threshold, perform some task in less than a certain amount of time, or remain in their positions for more than a set number of weeks or months. Rather than drawing categorical distinctions along these lines, data mining could also estimate or predict the specific numerical value of sales, production time, or tenure period, enabling employers to rank rather than simply sort employees.

These may seem like eminently reasonable things for employers to want to predict, but they are, by necessity, only part of an array of possible definitions of "good." An employer may instead attempt to define the target variable in a more holistic way—by, for example, relying on the grades that prior employees have received in annual reviews, which are supposed to reflect

27. Though described as a matter of detection, this is really a classification task, where any given transaction or email can belong to one of two possible classes, respectively: fraud or not fraud, or spam or not spam.

28. *See generally* Martha Ann Poon, What Lenders See—A History of the Fair Isaac Scorecard, (2013) (unpublished Ph.D. dissertation, University of California, San Diego), http://search.proquest.com/docview/1520318884 [https://perma.cc/YD3S-B9N7].

29. Hand, *supra* note 26, at 10.

an overall assessment of performance. These target variable definitions simply inherit the formalizations involved in preexisting assessment mechanisms, which in the case of human-graded performance reviews, may be far less consistent.[30]

Thus, the definition of the target variable and its associated class labels will determine what data mining happens to find. While critics of data mining have tended to focus on inaccurate classifications (false positives and false negatives),[31] as much—if not more—danger resides in the definition of the class label itself and the subsequent labeling of examples from which rules are inferred.[32] While different choices for the target variable and class labels can seem more or less reasonable, valid concerns with discrimination enter at this stage because the different choices may have a greater or lesser adverse impact on protected classes. For example, as later Parts will explain in detail, hiring decisions made on the basis of predicted tenure are much more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity. If the turnover rate happens to be systematically higher among members of certain protected classes, hiring decisions based on predicted length of employment will result in fewer job opportunities for members of these groups, even if they would have performed as well as or better than the other applicants the company chooses to hire.

B. Training Data

As described above, data mining learns by example. Accordingly, what a model learns depends on the examples to which it has been exposed. The data that function as examples are known as "training data"—quite literally, the data that train the model to behave in a certain way. The character of the training data can have meaningful consequences for the lessons that data mining happens to learn. As computer science scholars explain, biased training data leads to discriminatory models.[33] This can mean two rather different things,

30. Joseph M. Stauffer & M. Ronald Buckley, *The Existence and Nature of Racial Bias in Supervisory Ratings*, 90 J. APPLIED PSYCHOL. 586, 588–89 (2005) (showing evidence of racial bias in performance evaluations). Nevertheless, devising new target variables can have the salutary effect of forcing decision makers to think much more concretely about the outcomes that justifiably determine whether someone is a "good" employee. The explicit enumeration demanded of data mining thus also presents an opportunity to make decision making more consistent, more accountable, and fairer overall. This, however, requires conscious effort and careful thinking, and is not a natural consequence of adopting data mining.

31. Bruce Schneier, *Data Mining for Terrorists*, SCHNEIER ON SECURITY (Mar. 9, 2006), https://www.schneier.com/blog/archives/2006/03/data_mining_for.html [https://perma.cc/ZW44-N2KR]; Oscar H. Gandy Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS & INFO. TECH. 29, 39–40 (2010); Mireille Hildebrandt & Bert-Jaap Koops, *The Challenges of Ambient Law and Legal Protection in the Profiling Era*, 73 MOD. L. REV. 428, 433–35 (2010).

32. *See infra* Part I.B.

33. Bart Custers, *Data Dilemmas in the Information Society: Introduction and Overview*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY 3, 20 (Bart Custers et al. eds., 2013).

though: (1) if data mining treats cases in which prejudice has played some role as valid examples to learn from, that rule may simply reproduce the prejudice involved in these earlier cases; or (2) if data mining draws inferences from a biased sample of the population, any decision that rests on these inferences may systematically disadvantage those who are under- or overrepresented in the dataset. Both can affect the training data in ways that lead to discrimination, but the mechanisms—improper labeling of examples and biased data collections—are sufficiently distinct that they warrant separate treatment.

1. Labeling Examples

Labeling examples is the process by which the training data is manually assigned class labels. In cases of fraud or spam, the data miners draw from examples that come prelabeled: when individual customers report fraudulent charges or mark a message as spam, they are actually labeling transactions and email for the providers of credit and webmail. Likewise, an employer using grades previously given at performance reviews is also using prelabeled examples.

In certain cases, however, there may not be any labeled data and data miners may have to figure out a way to label examples themselves. This can be a laborious process, and it is frequently fraught with peril.[34] Often the best labels for different classifications will be open to debate. On which side of the creditworthy line does someone who has missed four credit card payments fall, for example?[35] The answer is not obvious. Even where the class labels are uncontested or uncontroversial, they may present a problem because analysts will often face difficult choices in deciding which of the available labels best applies to a particular example. Certain cases may present some, but not all, criteria for inclusion in a particular class.[36] The situation might also work in reverse, where the class labels are insufficiently precise to capture meaningful differences between cases. Such imperfect matches will demand that data miners exercise judgment.

The unavoidably subjective labeling of examples will skew the resulting findings such that any decisions taken on the basis of those findings will characterize all future cases along the same lines. This is true even if such

34. Hand, *supra* note 26, at 10–11.

35. *Id.* at 10 ("The classical supervised classification paradigm also takes as fundamental the fact that the classes are well defined. That is, that there is some fixed clear external criterion, which is used to produce the class labels. In many situations, however, this is not the case. In particular, when the classes are defined by thresholding a continuous variable, there is always the possibility that the defining threshold might be changed. Once again, this situation arises in consumer credit, where it is common to define a customer as 'defaulting' if they fall three months in arrears with repayments. This definition, however, is not a qualitative one (contrast has a tumor/does not have a tumor) but is very much a quantitative one. It is entirely reasonable that alternative definitions (e.g., four months in arrears) might be more useful if economic conditions were to change.").

36. *Id.* at 11.

characterizations would seem plainly erroneous to analysts who looked more closely at the individual cases. For all their potential problems, though, the labels applied to the training data must serve as ground truth.[37] Thus, decisions based on discoveries that rest on haphazardly labeled data or data labeled in a systematically, though unintentionally, biased manner will seem valid according to the customary validation methods employed by data miners. So long as prior decisions affected by some form of prejudice serve as examples of *correctly* rendered determinations, data mining will necessarily infer rules that exhibit the same prejudice.

Consider a real-world example from a different context as to how biased data labeling can skew results. St. George's Hospital, in the United Kingdom, developed a computer program to help sort medical school applicants based on its previous admissions decisions.[38] Those admissions decisions, it turns out, had systematically disfavored racial minorities and women with credentials otherwise equal to other applicants'.[39] In drawing rules from biased prior decisions, St. George's Hospital unknowingly devised an automated process that possessed these very same prejudices. As editors at the *British Medical Journal* noted at the time, "[T]he program was not introducing new bias but merely reflecting that already in the system."[40] Were an employer to undertake a similar plan to automate its hiring decisions by inferring a rule from past decisions swayed by prejudice, the employer would likewise arrive at a decision procedure that simply reproduces the prejudice of prior decision makers. Indeed, automating the process in this way would turn the conscious prejudice or implicit bias of individuals involved in previous decision making into a formalized rule that would systematically alter the prospects of all future applicants. For example, the computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women's or historically black colleges.

Not only can data mining inherit *prior* prejudice through the mislabeling of examples, it can also reflect current prejudice through the ongoing behavior of users taken as inputs to data mining. This is what Professor Latanya Sweeney discovered in a study that found that Google queries for black-sounding names were more likely to return contextual (i.e., key-word triggered)

37. *Id.* at 12. Even when evaluating a model, the kinds of subtle mischaracterizations that happen during training will be impossible to detect because most "evaluation data" is just a small subset of the training data that has been withheld during the learning process. Any problems with the training data will be present in the evaluation data.

38. Stella Lowry & Gordon Macpherson, *A Blot on the Profession*, 296 BRIT. MED. J. 657, 657 (1988).

39. *Id.* at 657.

40. *Id.*

advertisements for arrest records than those for white-sounding names.[41] Sweeney confirmed that the companies paying for these advertisements had not set out to focus on black-sounding names; rather, the fact that black-sounding names were more likely to trigger such advertisements seemed to be an artifact of the algorithmic process that Google employs to determine which advertisements to display alongside certain queries.[42] Although it is not fully known how Google computes the so-called "quality score" according to which it ranks advertisers' bids, one important factor is the predicted likelihood, based on historical trends, that users will click on an advertisement.[43] As Sweeney points out, the process "learns over time which [advertisement] text gets the most clicks from viewers [of the advertisement]" and promotes that advertisement in its rankings accordingly.[44] Sweeney posits that this aspect of the process could result in the differential delivery of advertisements that reflect the kinds of prejudice held by those exposed to the advertisements.[45] In attempting to cater to users' preferences, Google will unintentionally reproduce the existing prejudices that inform users' choices.

A similar situation could conceivably arise on websites that recommend potential employees to employers, as LinkedIn does through its Talent Match feature.[46] If LinkedIn determines which candidates to recommend based on the demonstrated interest of employers in certain types of candidates, Talent Match will offer recommendations that reflect whatever biases employers happen to exhibit. In particular, if LinkedIn's algorithm observes that employers disfavor certain candidates who are members of a protected class, Talent Match may decrease the rate at which it recommends these candidates to employers. The recommendation engine would learn to cater to the prejudicial preferences of employers.

There is an old adage in computer science: "garbage in, garbage out." Because data mining relies on training data as ground truth, when those inputs

41. Latanya Sweeney, *Discrimination in Online Ad Delivery*, COMM. ACM, May 2013, at 44, 47 (2013).

42. *Id.* at 48, 52.

43. *Check and Understand Quality Score*, GOOGLE, https://support.google.com/adwords/answer/2454010?hl=en [https://perma.cc/A88T-GF8X] (last visited July 26, 2014).

44. Sweeney, *supra* note 41, at 52.

45. The fact that black people may be convicted of crimes at a higher rate than nonblack people does not explain why those who search for black-sounding names would be any more likely to click on advertisements that mention an arrest record than those who see the same exact advertisement when they search for white-sounding names. If the advertisement implies, in both cases, that a person of that particular name has an arrest record, as Sweeney shows, the only reason the advertisements keyed to black-sounding names should receive greater attention is if searchers confer greater significance to the fact of prior arrests when the person happens to be black. *Id.* at 53.

46. Dan Woods, *LinkedIn's Monica Rogati on "What Is a Data Scientist?,"* FORBES (Nov. 27, 2011), http://www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist [https://perma.cc/N9HT-BXU3].

are themselves skewed by bias or inattention, the resulting system will produce results that are at best unreliable and at worst discriminatory.

2. Data Collection

Decisions that depend on conclusions drawn from incorrect, partial, or nonrepresentative data may discriminate against protected classes. The individual records that a company maintains about a person might have serious mistakes,[47] the records of the entire protected class of which this person is a member might also have similar mistakes at a higher rate than other groups, and the entire set of records may fail to reflect members of protected classes in accurate proportion to others.[48] In other words, the quality and representativeness of records might vary in ways that correlate with class membership (e.g., institutions might maintain systematically less accurate, precise, timely, and complete records for certain classes of people). Even a dataset with individual records of consistently high quality can suffer from statistical biases that fail to represent different groups in accurate proportions. Much attention has focused on the harms that might befall individuals whose records in various commercial databases are error ridden.[49] Far less consideration, however, has been paid to the systematic disadvantage that members of protected classes may suffer from being miscounted and, as a result, misrepresented in the evidence base.

Recent scholarship has begun to stress this point. Jonas Lerman, for example, worries about "the nonrandom, systemic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle, and whose lives are less 'datafied' than the general population's."[50] Professor Kate Crawford has likewise warned that "[b]ecause not all data is created or even collected equally, there are 'signal problems' in big-data sets—dark zones or shadows where some citizens and communities are overlooked or

47. Data quality is a topic of lively practical and philosophical debate. *See, e.g.*, Luciano Floridi, *Information Quality*, 26 PHIL. & TECH. 1 (2013); Richard Y. Wang & Diane M. Strong, *Beyond Accuracy: What Data Quality Means to Data Consumers*, 12 J. MGMT. INFO. SYS. 5 (1996). The components of data quality have been thought to include accuracy, precision, completeness, consistency, validity, and timeliness, though this catalog of features is far from settled. *See generally* LARRY P. ENGLISH, INFORMATION QUALITY APPLIED (2009).

48. *Cf.* Zeynep Tufekci, *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*, EIGHTH INT'L AAAI CONF. WEBLOGS & SOC. MEDIA (2014), http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8062/8151 [https://perma.cc/G4G7-2VZ8].

49. *See, e.g.*, FED. TRADE COMM'N, REPORT TO CONGRESS UNDER SECTION 319 OF THE FAIR AND ACCURATE CREDIT TRANSACTIONS ACT OF 2003 A-4 (2012) (finding that nearly 20 percent of consumers had an error in one or more of their three credit reports and that 5.4 percent of consumers had errors that could result in less favorable loan terms).

50. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013).

underrepresented."[51] Errors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation.[52] Not only will the quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data.

As an illustrative example, Crawford points to Street Bump, an application for Boston residents that takes advantage of accelerometers built into smart phones to detect when drivers ride over potholes.[53] While Crawford praises the cleverness and cost-effectiveness of this passive approach to reporting road problems, she rightly warns that whatever information the city receives from Street Bump will be biased by the uneven distribution of smartphones across populations in different parts of the city.[54] In particular, systematic differences in smartphone ownership will very likely result in the underreporting of road problems in the poorer communities where protected groups disproportionately congregate.[55] If the city were to rely on this data to determine where it should direct its resources, it would only further underserve these communities. Indeed, the city would discriminate against those who lack the capability to report problems as effectively as wealthier residents with smartphones.[56]

A similar dynamic could easily apply in an employment context if members of protected classes are unable to report their interest in and qualification for jobs listed online as easily or effectively as others due to systematic differences in Internet access. The EEOC has established a program called "Eradicating Racism & Colorism from Employment" (E-RACE) that aims, at least in part, to prevent this sort of discrimination from occurring due

51. Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013), http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data [https://perma.cc/S9ZA-XEXH].

52. *See id.*; Lerman, *supra* note 50, at 57.

53. Crawford, *supra* note 51 (explaining that a sudden movement suggesting a broken road will automatically prompt the phone to report the location to the city).

54. *Id.*

55. *See id.*

56. This is, of course, a more general problem with representative democracy. For a host of reasons, the views and interests of the poor are relatively less well represented in the political process. *See, e.g.*, Larry M. Bartels, *Economic Inequality and Political Representation*, *in* THE UNSUSTAINABLE AMERICAN STATE 167 (Lawrence Jacobs & Desmond King eds., 2009); MARTIN GILENS, AFFLUENCE AND INFLUENCE: ECONOMIC INEQUALITY AND POLITICAL POWER IN AMERICA (2012). The worry here, as expressed by Crawford, is that, for all its apparent promise, data mining may further obfuscate or legitimize these dynamics rather than overcome them.

to an employer's desire for high-tech hiring, such as video résumés.[57] E-RACE not only attempts to lower the barriers that would disproportionately burden applicants who belong to a protected class, but also ensures that employers do not develop an inaccurate impression of the incidence of qualified and interested candidates from these communities. If employers were to rely on tallies of high-tech candidates to direct their recruiting efforts, for example, any count affected by a reporting bias could have adverse consequences for specific populations systematically underrepresented in the dataset. Employers would deny equal attention to those who reside in areas incorrectly pegged as having a relatively lower concentration of qualified candidates.

Additional and even more severe risks may reside in the systematic omission of members of protected classes from such datasets. The Street Bump and Internet job application examples only discuss decisions that depend on raw tallies, rather than datasets from which decision makers want to draw generalizations and generate predictions. But data mining is especially sensitive to statistical bias because data mining helps to discover patterns that organizations tend to treat as generalizable findings even though the analyzed data only includes a partial sample from a circumscribed period. To ensure that data mining reveals patterns that hold true for more than the particular sample under analysis, the sample must be proportionally representative of the entire population, even though the sample, by definition, does not include every case.[58]

If a sample includes a disproportionate representation of a particular class (more or less than its actual incidence in the overall population), the results of an analysis of that sample may skew in favor of or against the over- or underrepresented class. While the representativeness of the data is often simply assumed, this assumption is rarely justified and is "perhaps more often incorrect than correct."[59] Data gathered for routine business purposes tend to lack the rigor of social scientific data collection.[60] As Lerman points out, "Businesses may ignore or undervalue the preferences and behaviors of

57. *Why Do We Need E-RACE?*, EQUAL EMPLOY. OPPORTUNITY COMM'N, http://www1.eeoc.gov/eeoc/initiatives/e-race/why_e-race.cfm [https://perma.cc/S3GY-2MD6] (last visited Mar. 1, 2013). Due to the so-called "digital divide," communities underserved by residential Internet access rely heavily on mobile phones for connectivity and thus often have trouble even uploading and updating traditional résumés. Kathryn Zickuhr & Aaron Smith, *Digital Differences*, PEW RES. CTR. (Apr. 13, 2012), http://www.pewinternet.org/2012/04/13/digital-differences [https://perma.cc/S545-42GY] ("Among smartphone owners, young adults, minorities, those with no college experience, and those with lower household income levels are more likely than other groups to say that their phone is their main source of internet access.").

58. Data mining scholars have devised ways to address this known problem, but applying these techniques is far from trivial. *See* Sinno Jialin Pan & Qiang Yang, *A Survey on Transfer Learning*, 22 IEEE TRANSACTIONS ON KNOWLEDGE & DATA ENG'G 1345, 1354–56 (2010).

59. Hand, *supra* note 26, at 7.

60. David Lazer, *Big Data and Cloning Headless Frogs*, COMPLEXITY & SOC. NETWORKS BLOG (Feb. 16, 2014), https://web.archive.org/web/20140711164511/http://blogs.iq.harvard.edu/netgov/2014/02/big_data_and_cloning_headless.html [https://perma.cc/TQ9A-TP2Z].

consumers who do not shop in ways that big data tools can easily capture, aggregate, and analyze."[61]

In the employment context, even where a company performs an analysis of the data from its entire population of employees—avoiding the apparent problem of even having to select a sample—the organization must assume that its future applicant pool will have the same degree of variance as its current employee base. An organization's tendency, however, to perform such analyses in order to *change* the composition of their employee base should put the validity of this assumption into immediate doubt. The potential effect of this assumption is the future mistreatment of individuals predicted to behave in accordance with the skewed findings derived from the biased sample. Worse, these results may lead to decision procedures that limit the future contact an organization will have with specific groups, skewing still further the sample upon which subsequent analyses will be performed.[62] Limiting contact with specific populations on the basis of unsound generalizations may deny members of these populations the opportunity to prove that they buck the apparent trend.

Overrepresentation in a dataset can also lead to disproportionately high adverse outcomes for members of protected classes. Consider an example from the workplace: managers may devote disproportionate attention to monitoring the activities of employees who belong to a protected class and consequently observe mistakes and transgressions at systematically higher rates than others, in part because these managers fail to subject others who behave similarly to the same degree of scrutiny. Not only does this provide managers with justification for their prejudicial suspicions, but it also generates evidence that overstates the relative incidence of offenses by members of these groups. Where subsequent managers who hold no such prejudicial suspicions cannot observe everyone equally, they may rely on this evidence to make predictions about where to focus their attention in the future and thus further increase the disproportionate scrutiny that they place on protected classes.

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

---

61. Lerman, *supra* note 50, at 59.

62. Practitioners, particularly those involved in credit scoring, are well aware that they do not know how the person purposefully passed over would have behaved if he had been given the opportunity. Practitioners have developed methods to correct for this bias (which, in the case of credit scoring, they refer to as reject inference). *See, e.g.*, Jonathan Crook & John Banasik, *Does Reject Inference Really Improve the Performance of Application Scoring Models?*, 28 J. BANKING & FIN. 857 (2004).

C. Feature Selection

Through a process called "feature selection," organizations—and the data miners that work for them—make choices about what attributes they observe and subsequently fold into their analyses.[63] These decisions can also have serious implications for the treatment of protected classes if those factors that better account for pertinent statistical variation among members of a protected class are not well represented in the set of selected features.[64] Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve.

This problem arises because data are necessarily reductive representations of an infinitely more specific real-world object or phenomenon.[65] These representations may fail to capture enough detail to allow for the discovery of crucial points of contrast. Increasing the resolution and range of the analysis may still fail to capture the mechanisms that account for different outcomes because such mechanisms may not lend themselves to exhaustive or effective representation in the data, if such representations even exist. As Professors Toon Calders and Indrė Žliobaitė explain, "[I]t is often impossible to collect all the attributes of a subject or take all the environmental factors into account with a model."[66] While these limitations lend credence to the argument that a dataset can never fully encompass the full complexity of the individuals it seeks to represent, they do not reveal the inherent inadequacy of representation as such.

At issue, really, are the coarseness and comprehensiveness of the criteria that permit statistical discrimination and the uneven rates at which different groups happen to be subject to erroneous determinations. Crucially, these erroneous and potentially adverse outcomes are artifacts of statistical reasoning rather than prejudice on the part of decision makers or bias in the composition of the dataset. As Professor Frederick Schauer explains, decision makers that rely on statistically sound but nonuniversal generalizations "are being simultaneously rational and unfair" because certain individuals are "actuarially saddled" by statistically sound inferences that are nevertheless inaccurate.[67]

---

63. FEATURE EXTRACTION, CONSTRUCTION AND SELECTION 71–72 (Huan Liu & Hiroshi Motoda eds., 1998).

64. Toon Calders & Indrė Žliobaitė, *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 33, at 43, 46 ("[T]he selection of attributes by which people are described in [a] database may be incomplete.").

65. Annamarie Carusi, *Data as Representation: Beyond Anonymity in E-Research Ethics*, 1 INT'L J. INTERNET RES. ETHICS 37, 48–61 (2008).

66. Calders & Žliobaitė, *supra* note 64, at 47.

67. FREDERICK SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 3–7 (2006). Insurance offers the most obvious example of this: the rate that a person pays for car insurance, for

Obtaining information that is sufficiently rich to permit precise distinctions can be expensive. Even marginal improvements in accuracy may come at significant practical costs and may justify a less granular and encompassing analysis.[68]

To take an obvious example from the employment context, hiring decisions that consider academic credentials tend to assign enormous weight to the reputation of the college or university from which an applicant has graduated, even though such reputations may communicate very little about the applicant's job-related skills and competencies.[69] If equally competent members of protected classes happen to graduate from these colleges or universities at disproportionately low rates, decisions that turn on the credentials conferred by these schools, rather than some set of more specific qualities that more accurately sort individuals, will incorrectly and systematically discount these individuals. Even if employers have a rational incentive to look beyond credentials and focus on criteria that allow for more precise and more accurate determinations, they may continue to favor credentials because they communicate pertinent information at no cost to the employer.[70]

Similar dynamics seem to account for the practice known as "redlining,"[71] in which financial institutions employ especially general criteria to draw distinctions between subpopulations (i.e., the neighborhood in which individuals happen to reside), despite the fact that such distinctions fail to capture significant variation within each subpopulation that would result in a different assessment for certain members of these groups. While redlining in America is well known to have had its basis in racial animus and prejudice,[72] decision makers operating in this manner may attempt to justify their behavior by pointing to the cost efficiency of relying on easily accessible information. In other words, decision makers can argue that they are willing to tolerate higher rates of erroneous determinations for certain groups because the benefits

---

instance, is determined by the way other people with similar characteristics happen to drive, even if the person is a better driver than those who resemble him on the statistically pertinent dimensions.

68. Kasper Lippert-Rasmussen, *"We Are All Different": Statistical Discrimination and the Right to Be Treated as an Individual*, 15 J. ETHICS 47, 54 (2011) ("[O]btaining information is costly, so it is morally justified, all things considered, to treat people on the basis of statistical generalizations even though one knows that, in effect, this will mean that one will treat some people in ways, for better or worse, that they do not deserve to be treated."); *see also* Brian Dalessandro, Claudia Perlich & Troy Raeder, *Bigger Is Better, but at What Cost?: Estimating the Economic Value of Incremental Data Assets*, 2 BIG DATA 87 (2014).

69. *See* Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. TIMES (Apr. 28, 2013), http://www.nytimes.com/2013/04/28/technology/how-big-data-is-playing-recruiter-for-specialized-workers.html [https://perma.cc/DC7A-W2B5].

70. As one commentator has put it in contemplating data-driven hiring, "Big Data has its own bias. . . . You measure what you can measure." *Id.*

71. *See generally* DAVID M. P. FREUND, COLORED PROPERTY: STATE POLICY AND WHITE RACIAL POLITICS IN SUBURBAN AMERICA (2010).

72. *Id.*

derived from more granular data—and thus better accuracy—do not justify the costs. Of course, it may be no coincidence that such cost-benefit analyses seem to justify treating groups composed disproportionately of members of protected classes to systematically less accurate determinations.[73] Redlining is illegal because it can systematically discount entire areas composed primarily of members of a protected class, despite the presence of some qualified candidates.[74]

Cases of so-called rational racism are really just a special instance of this more general phenomenon—one in which race happens to be taken into consideration explicitly. In such cases, decision makers take membership in a protected class into account, even if they hold no prejudicial views, because such membership seems to communicate relevant information that would be difficult or impossible to obtain otherwise. Accordingly, the persistence of distasteful forms of discrimination may be the result of a lack of information, rather than a continued taste for discrimination.[75] Professor Lior Strahilevitz has argued, for instance, that when employers lack access to criminal records, they may consider race in assessing an applicant's likelihood of having a criminal record because there are statistical differences in the rates at which members of different racial groups have been convicted of crimes.[76] In other words, employers fall back on more immediately available and coarse features when they cannot access more specific or verified information.[77] Of course, as Strahilevitz points out, race is a highly imperfect basis upon which to predict an individual's criminal record, despite whatever differences may exist in the rates at which members of different racial groups have been convicted of crimes, because it is too coarse as an indicator.[78]

---

73. While animus was likely the main motivating factor for redlining, the stated rationales were economic and about housing value. *See* DOUGLAS S. MASSEY & NANCY A. DENTON, AMERICAN APARTHEID: SEGREGATION AND THE MAKING OF THE UNDERCLASS 51–52 (1993). Redlining persists today and may actually be motivated by profit, but it has the same deleterious effects. *See* Rachel L. Swarns, *Biased Lending Evolves, and Blacks Face Trouble Getting Mortgages*, N.Y. TIMES (Oct. 30 2015), http://www.nytimes.com/2015/10/31/nyregion/hudson-city-bank-settlement.html [https://perma.cc/P4YX-NTT9].

74. *See* Nationwide Mut. Ins. Co. v. Cisneros, 52 F.3d 1351, 1359 (6th Cir. 1995) (holding that the Fair Housing Act prohibited redlining in order "to eliminate the discriminatory business practices which might prevent a person economically able to do so from purchasing a house regardless of his race"); NAACP v. Am. Family Mut. Ins. Co., 978 F.2d 287, 300 (7th Cir. 1992).

75. *See generally* Andrea Romei & Salvatore Ruggieri, *Discrimination Data Analysis: A Multi-Disciplinary Bibliography*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 33, at 109, 120.

76. Lior Jacob Strahilevitz, *Privacy Versus Antidiscrimination*, 75 U. CHI. L. REV. 363, 364 (2008).

77. *Id.* This argument assumes that criminal records are relevant to employment, which is often not true. *See infra* text accompanying note 175.

78. Strahilevitz, *supra* note 76, at 364; *see also infra* Part II.A. The law holds that decision makers should refrain from considering membership in a protected class even if statistical evidence seems to support certain inferences on that basis. The prohibition does not depend on whether decision

D. Proxies

Cases of decision making that do not artificially introduce discriminatory effects into the data mining process may nevertheless result in systematically less favorable determinations for members of protected classes. This is possible when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job—as formalized in some fashion—may also sort individuals according to class membership.

In certain cases, there may be an obvious reason for this. Just as "mining from historical data may . . . discover traditional prejudices that are endemic in reality (i.e., taste-based discrimination)," so, too, may data mining "discover patterns of lower performances, skills or capacities of protected-by-law groups."[79] These discoveries not only reveal the simple fact of inequality, but they also reveal that these are inequalities in which members of protected classes are frequently in the relatively less favorable position. This has rather obvious implications: if features held at a lower rate by members of protected groups nevertheless possess relevance in rendering legitimate decisions, such decisions will necessarily result in systematically less favorable determinations for these individuals. For example, by conferring greater attention and opportunities to employees that they predict will prove most competent at some task, employers may find that they subject members of protected groups to consistently disadvantageous treatment because the criteria that determine the attractiveness of employees happen to be held at systematically lower rates by members of these groups.[80]

Decision makers do not necessarily intend this disparate impact because they hold prejudicial beliefs; rather, their reasonable priorities as profit seekers unintentionally recapitulate the inequality that happens to exist in society. Furthermore, this may occur even if proscribed criteria have been removed from the dataset, the data are free from latent prejudice or bias, the features are especially granular and diverse, and the only goal is to maximize classificatory or predictive accuracy. The problem stems from what researchers call "redundant encodings," cases in which membership in a protected class happens to be encoded in other data.[81] This occurs when a particular piece of data or certain values for that piece of data are highly correlated with

---

makers can gain (easy or cheap) access to alternative criteria that hold greater predictive value. *See* Grutter v. Bollinger, 539 U.S. 306, 326 (2003).

79. Romei & Ruggieri, *supra* note 75, at 121.

80. Faisal Kamiran, Toon Calders & Mykola Pechenizkiy, *Techniques for Discrimination-Free Predictive Models*, *in* DISCRIMINATION AND PRIVACY IN THE INFORMATION SOCIETY, *supra* note 33, at 223–24.

81. Cynthia Dwork et al., *Fairness Through Awareness*, 3 PROC. INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214 app. at 226 (2012) ("Catalog of Evils").

membership in specific protected classes. Data's significant statistical relevance to the decision at hand helps explain why data mining can result in seemingly discriminatory models even when its only objective is to ensure the greatest possible accuracy for its determinations. If there is a disparate distribution of an attribute, a more precise form of data mining will be more likely to capture that distribution. Better data and more features will simply come closer to exposing the exact extent of inequality.

E. Masking

Data mining could also breathe new life into traditional forms of intentional discrimination because decision makers with prejudicial views can mask their intentions by exploiting each of the mechanisms enumerated above. Stated simply, any form of discrimination that happens unintentionally can also be orchestrated intentionally. For instance, decision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes.[82] They could likewise attempt to preserve the known effects of prejudice in prior decision making by insisting that such decisions constitute a reliable and impartial set of examples from which to induce a decision-making rule. And decision makers could intentionally rely on features that only permit coarse-grained distinction making—distinctions that result in avoidably higher rates of erroneous determinations for members of a protected class. In denying themselves finer-grained detail, decision makers would be able to justify writing off entire groups composed disproportionately of members of protected classes. A form of digital redlining, this decision masks efforts to engage in intentional discrimination by abstracting to a level of analysis that fails to capture lower level variations. As a result, certain members of protected classes might not be seen as attractive candidates. Here, prejudice rather than some legitimate business reason (such as cost) motivates decision makers to intentionally restrict the particularity of their decision making to a level that can only paint in avoidably broad strokes. This condemns entire groups, composed disproportionately of members of protected classes, to systematically less favorable treatment.

Because data mining holds the potential to infer otherwise unseen attributes, including those traditionally deemed sensitive,[83] it can indirectly determine individuals' membership in protected classes and unduly discount, penalize, or exclude such people accordingly. In other words, data mining could grant decision makers the ability to distinguish and disadvantage members of protected classes even if those decision makers do not have access to explicit information about individuals' class membership. Data mining could

82. *See id.* (discussing the "[s]elf-fulfilling prophecy").

83. *See* Solon Barocas, Leaps and Bounds: Toward a Normative Theory of Inferential Privacy 9 (Nov. 11, 2015) (in-progress and unpublished manuscript) (on file with authors).

instead help to pinpoint reliable proxies for such membership and thus place institutions in the position to automatically sort individuals into their respective class without ever having to learn these facts directly.[84] The most immediate implication is that institutions could employ data mining to circumvent the barriers, both practical and legal, that have helped to withhold individuals' protected class membership from consideration.

Additionally, data mining could provide cover for intentional discrimination of this sort because the process conceals the fact that decision makers determined and considered the individual's class membership. The worry, then, is not simply that data mining introduces novel ways for decision makers to satisfy their taste for illegal discrimination; rather, the worry is that it may mask actual cases of such discrimination.[85] Although scholars, policy makers, and lawyers have long been aware of the dangers of masking,[86] data mining significantly enhances the ability to conceal acts of intentional discrimination by finding ever more remote and complex proxies for proscribed criteria.[87]

Intentional discrimination and its masking have so far garnered disproportionate attention in discussions of data mining,[88] often to the exclusion of issues arising from the many forms of unintentional discrimination described above. While data mining certainly introduces novel ways to discriminate intentionally and to conceal those intentions, most cases of employment discrimination are already sufficiently difficult to prove; employers motivated by conscious prejudice would have little to gain by pursuing these complex and costly mechanisms to further mask their intentions.[89] When it comes to data mining, unintentional discrimination is the more pressing concern because it is likely to be far more common and easier to overlook.

---

84. *Id.* at 9–13.

85. Data miners who wish to discriminate can do so using relevant or irrelevant criteria. Either way the intent would make the action "masking." If an employer masked using highly relevant data, litigation arising from it likely would be tried under a "mixed-motive" framework, which asks whether the same action would have been taken without the intent to discriminate. *See infra* Part II.A.

86. *See, e.g.*, Custers, *supra* note 33, at 9–10.

87. *See* Barocas, *supra* note 83.

88. *See, e.g.*, Alistair Croll, *Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It*, SOLVE FOR INTERESTING (July 31, 2012, 12:40 PM), http://solveforinteresting.com/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it [https://perma.cc/BS8S-6T7S]. This post generated significant online chatter immediately upon publication and has become one of the canonical texts in the current debate. It has also prompted a number of responses from scholars. *See, e.g.*, Anders Sandberg, *Asking the Right Questions: Big Data and Civil Rights*, PRAC. ETHICS (Aug. 16, 2012), http://blog.practicalethics.ox.ac.uk/2012/08/asking-the-right-questions-big-data-and-civil-rights [https://perma.cc/NC36-NBZN].

89. *See* Linda Hamilton Krieger, *The Content of Our Categories: A Cognitive Bias Approach to Discrimination and Equal Employment Opportunity*, 47 STAN. L. REV. 1161, 1177 (1995).

II.
TITLE VII LIABILITY FOR DISCRIMINATORY DATA MINING

Current antidiscrimination law is not well equipped to address the cases of discrimination stemming from the problems described in Part I. This Part considers how Title VII might apply to these cases. Other antidiscrimination laws, such as the Americans with Disabilities Act, will exhibit differences in specific operation, but the main thrust of antidiscrimination law is fairly consistent across regimes, and Title VII serves as an illustrative example.[90]

An employer sued under Title VII may be found liable for employment discrimination under one of two theories of liability: disparate treatment and disparate impact.[91] Disparate treatment comprises two different strains of discrimination: (1) formal disparate treatment of similarly situated people and (2) intent to discriminate.[92] Disparate impact refers to policies or practices that are facially neutral but have a disproportionately adverse impact on protected classes.[93] Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result.[94]

Liability under Title VII for discriminatory data mining will depend on the particular mechanism by which the inequitable outcomes are generated. This Part explores the disparate treatment and disparate impact doctrines and analyzes which mechanisms could generate liability under each theory.

A. Disparate Treatment

Disparate treatment recognizes liability for both explicit formal classification and intentional discrimination.[95] Formal discrimination, in which membership in a protected class is used as an input to the model, corresponds to an employer classifying employees or potential hires according to membership in a protected class and differentiating them on that basis. Formal

90. The biggest difference between the Americans with Disabilities Act and Title VII is the requirement that an employer make "reasonable accommodations" for disabilities. 42 U.S.C. § 12112(b)(5) (2012). But some scholars have argued that even this difference is illusory and that accommodations law is functionally similar to Title VII, though worded differently. *See* Samuel R. Bagenstos, *"Rational Discrimination," Accommodation, and the Politics of (Disability) Civil Rights*, 89 VA. L. REV. 825, 833 & n.15 (2003) (comparing accommodations law to disparate treatment); Christine Jolls, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 652 (2001) (comparing accommodations law to disparate impact).

91. *See* 42 U.S.C. § 2000e; Ricci v. DeStefano, 557 U.S. 557, 577 (2009).

92. Richard A. Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1351 n.56 (2010) (explaining that, for historical reasons, disparate treatment became essentially "not-disparate-impact" and now we rarely notice the two different embedded theories).

93. *See* Griggs v. Duke Power Co., 401 U.S. 424, 430 (1971).

94. 42 U.S.C. § 2000e-2(k).

95. *Id.* § 2000e-2(a), (k); *see* Primus, *supra* note 92, at 1350–51 n.56.

discrimination covers both the straightforward denial of opportunities based on protected class membership and the use of rational racism.[96] In traditional contexts, rational racism is considered rational because there are cases in which its users believe it is an accurate, if coarse-grained, proxy—or at least the best available one in a given situation.[97] In the world of data mining, though, that need not be the case. Even if membership in a protected class were specified as an input, the eventual model that emerges could see it as the least significant feature. In that case, there would be no discriminatory effect, but there would be a disparate treatment violation, because considering membership in a protected class as a potential proxy is a legal classificatory harm in itself.[98]

Formal liability does not correspond to any particular discrimination mechanism within data mining; it can occur equally well in any of them. Because classification itself can be a legal harm, irrespective of the effect,[99] the same should be true of using protected class as an input to a system for which the entire purpose is to build a classificatory model.[100] The irony is that the use of protected class as an input is usually irrelevant to the outcome in terms of discriminatory effect, at least given a large enough number of input features. The target variable will, in reality, be correlated to the membership in a protected class somewhere between 0 percent and 100 percent. If the trait is perfectly uncorrelated, including membership in the protected class as an input will not change the output, and there will be no discriminatory effect.[101] On the other end of the spectrum, where membership in the protected class is perfectly predictive of the target variable, the fact will be redundantly encoded in the other data. The only way using membership in the protected class as an explicit feature will change the outcome is if the information is otherwise not rich enough to detect such membership. Membership in the protected class will prove relevant to the exact extent it is already redundantly encoded. Given a rich enough set of features, the chance that such membership is redundantly encoded approaches certainty. Thus, a data mining model with a large number of variables will determine the extent to which membership in a protected class is relevant to the sought-after trait whether or not that information is an input. Formal discrimination therefore should have no bearing whatsoever on the

96. Michelle R. Gomez, *The Next Generation of Disparate Treatment: A Merger of Law and Social Science*, 32 REV. LITIG. 553, 562 (2013).

97. Strahilevitz, *supra* note 76, at 365–67.

98. Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 504 (2003).

99. *See* Jed Rubenfeld, *Affirmative Action*, 107 YALE L.J. 427, 433 (1997) (discussing "[c]lassificationism"); Primus, *supra* note 98, at 504, 567–68 (discussing expressive harms).

100. Membership in a protected class is still a permissible input to a holistic determination when the focus is diversity, but where classification is the goal, such as here, it is not. *See* Grutter v. Bollinger, 539 U.S. 306, 325 (2003) (noting that "diversity is a compelling state interest" that can survive strict scrutiny).

101. That is, not counting any expressive harm that might come from classification by protected class.

outcome of the model. Additionally, by analyzing the data, an employer could probabilistically determine an employee's membership in that same protected class, if the employer did indeed want to know.

To analyze intentional discrimination other than mere formal discrimination, a brief description of disparate treatment doctrine is necessary. A Title VII disparate treatment case will generally proceed under either the *McDonnell-Douglas* burden-shifting scheme or the *Price-Waterhouse* "mixed motive" regime.[102] Under the *McDonnell-Douglas* framework, the plaintiff who has suffered an adverse employment action has the initial responsibility to establish a prima facie case of discrimination by demonstrating that a similarly situated person who is not a member of a protected class would not have suffered the same fate.[103] This can be shown with circumstantial evidence of discriminatory intent, such as disparaging remarks made by the employer or procedural irregularities in promotion or hiring; only very rarely will an employer openly admit to discriminatory conduct. If the plaintiff successfully demonstrates that the adverse action treated protected class members differently, then the burden shifts to the defendant-employer to offer a legitimate, nondiscriminatory basis for the decision. The defendant need not prove the reason is true; his is only a burden of production.[104] Once the defendant has offered a nondiscriminatory alternative, the ultimate burden of persuasion falls to the plaintiff to demonstrate that the proffered reason is pretextual.[105]

In the data mining context, liability for masking is clear as a theoretical matter, no matter which mechanism for discrimination is employed. The fact that it is accomplished algorithmically does not make it less of a disparate treatment violation, as the entire idea of masking is pretextual. In fact, in the traditional, non–data mining context, the word masking has occasionally been used to refer to pretext.[106] Like in any disparate treatment case, however, proof will be difficult to come by, something even truer for masking.[107]

---

102. McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973); Price Waterhouse v. Hopkins, 490 U.S. 228 (1989).

103. This is similar to the computer science definition of discrimination. Calders & Žliobaitė, *supra* note 64, at 49. ("A classifier discriminates with respect to a sensitive attribute, e.g. gender, if for two persons which only differ by their gender (and maybe some characteristics irrelevant for the classification problem at hand) that classifier predicts different labels.").

104. St. Mary's Honor Ctr. v. Hicks, 509 U.S. 502, 507 (1993).

105. *Id.*

106. *See* Keyes v. Sec'y of the Navy, 853 F.2d 1016, 1026 (1st Cir. 1988) (explaining that it is the plaintiff's burden to show that the proffered reasons for hiring an alternative were "pretexts aimed at masking sex or race discrimination"); Custers, *supra* note 33, at 9–10; Megan Whitehill, *Better Safe than Subjective: The Problematic Intersection of Prehire Social Networking Checks and Title VII Employment Discrimination*, 85 TEMP. L. REV. 229, 250 (2012) (referring to "[m]asking [p]retext" in the third stage of *McDonnell-Douglas* framework).

107. *See supra* Part I.E. This is a familiar problem to antidiscrimination law, and it is often cited as one of the rationales for disparate impact liability in the first place—to "smoke out" intentional invidious discrimination. *See infra* Part III.B.

The *McDonnell-Douglas* framework operates on a presumption that if the rationale that the employer has given is found to be untrue, the employer must be hiding his "true" discriminatory motive.[108] Because the focus of the *McDonnell-Douglas* framework is on pretext and cover-up, it can only address conscious, willful discrimination.[109] Under the *McDonnell-Douglas* framework, a court must find either that the employer *intended* to discriminate or did not discriminate at all.[110] Thus, unintentional discrimination will not lead to liability.

A Title VII disparate treatment case can also be tried under the mixed-motive framework, first recognized in *Price Waterhouse v. Hopkins*[111] and most recently modified by *Desert Palace, Inc. v. Costa*.[112] In the mixed-motive framework, a plaintiff need not demonstrate that the employer's nondiscriminatory rationale was pretextual, but merely that discrimination was a "motivating factor" in the adverse employment action.[113] As a practical matter, this means that the plaintiff must show that the same action would not have been taken absent the discriminatory motive.[114] As several commentators

108. McDonnell Douglas Corp. v. Green, 411 U.S. 792, 805 (1973) (The plaintiff "must be given a full and fair opportunity to demonstrate by competent evidence that the presumptively valid reasons for his rejection were in fact a coverup for a racially discriminatory decision"). While, as a theoretical matter, the plaintiff must prove that the employer's reason was a pretext for discrimination specifically, the Supreme Court has held that a jury can reasonably find that the fact that an employer had only a pretextual reason to fall back on is itself circumstantial evidence of discrimination. *Hicks*, 509 U.S. at 511 ("The factfinder's disbelief of the reasons put forward by the defendant (particularly if disbelief is accompanied by a suspicion of mendacity) may, together with the elements of the prima facie case, suffice to show intentional discrimination.").

109. *See* Tristin K. Green, *Discrimination in Workplace Dynamics: Toward a Structural Account of Disparate Treatment Theory*, 38 HARV. C.R.-C.L. L. REV. 91, 114 (2003) ("Presuming that individuals know the real reason for their actions, the pretext model of disparate treatment provides that an employer can be held to have discriminated when the plaintiff establishes a minimal prima facie case and shows that the reason given for the adverse decision is unworthy of credence."); Susan Sturm, *Second Generation Employment Discrimination: A Structural Approach*, 101 COLUM. L. REV. 458, 458 (2001); *see also* Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 ALA. L. REV. 741, 749–50 (2005) (critiquing the courts' requirement of proving employer "dishonesty," but suggesting that, absent this requirement, Title VII could handle unconscious discrimination without altering the law).

110. Krieger, *supra* note 89, at 1170.

111. 490 U.S. 228 (1989).

112. 539 U.S. 90 (2003).

113. 42 U.S.C. § 2000e-2(m) (2012); *Desert Palace*, 539 U.S. at 101 ("In order to obtain [a mixed-motive jury instruction], a plaintiff need only present sufficient evidence for a reasonable jury to conclude, by a preponderance of the evidence, that 'race, color, religion, sex, or national origin was a motivating factor for any employment practice.'"). The efficacy of data mining is fundamentally dependent on the quality of the data from which it attempts to draw useful lessons. If these data capture the prejudicial or biased behavior of prior decision makers, data mining will learn from the bad example that these decisions set. If the data fail to serve as a good sample of a protected group, data mining will draw faulty lessons that could serve as a discriminatory basis for future decision making.

114. Charles A. Sullivan, *Disparate Impact: Looking Past the* Desert Palace *Mirage*, 47 WM. & MARY L. REV. 911, 914–16, 916 n.20 (2005); *see also* Krieger, *supra* note 89, at 1170–72; D. Don Welch, *Removing Discriminatory Barriers: Basing Disparate Treatment Analysis on Motive Rather than Intent*, 60 S. CAL. L. REV. 733, 740 (1987).

have pointed out, motive and intent are not necessarily synonymous.[115] Motive can be read more broadly to include unconscious discrimination, including anything that influences a person to act, such as emotions or desires.[116] Nonetheless, courts have conflated the meanings of motive and intent such that the phrase "motive or intent" has come to refer only to conscious choices.[117] Thus, while most individual decision making probably belongs in a mixed-motive framework, as each decision a person makes comprises a complicated mix of motivations,[118] the mixed-motive framework will be no better than the pretext framework at addressing bias that occurs absent conscious intent.[119]

Except for masking, discriminatory data mining is by stipulation unintentional. Unintentional disparate treatment is not a problem that is new to data mining. A vast scholarly literature has developed regarding the law's treatment of unconscious, implicit bias.[120] Such treatment can occur when an employer has internalized some racial stereotype and applies it or, without realizing it, monitors an employee more closely until the employer finds a violation.[121] The employee is clearly treated differently, but it is not intentional, and the employer is unaware of it. As Professor Samuel Bagenstos summarized, at this point, "it may be difficult, if not impossible, for a court to go back and reconstruct the numerous biased evaluations and perceptions that ultimately resulted in an adverse employment decision."[122] Within the scholarly literature, there is "[s]urprising unanimity" that the law does not adequately address unconscious disparate treatment.[123]

---

115. Krieger, *supra* note 89, at 1243; Sullivan, *supra* note 114, at 915.

116. Krieger, *supra* note 89, at 1243; Sullivan, *supra* note 114, at 915 n.18 (quoting *Motive*, OXFORD ENGLISH DICTIONARY (1st ed. 1933)).

117. Sullivan, *supra* note 114, at 914–16, 916 n.20.

118. Amy L. Wax, *Discrimination as Accident*, 74 IND. L.J. 1129, 1149 & n.21 (1999); Krieger, *supra* note 89, at 1223. In fact, after the Supreme Court decided *Desert Palace*, many scholars thought that it *had* effectively overruled the *McDonnell-Douglas* framework, forcing all disparate treatment cases into a mixed-motive framework. *See, e.g.*, Sullivan, *supra* note 114, at 933–36 (discussing the then-emerging scholarly consensus). This has not played out so far, with courts and scholars split on the matter. *See, e.g.*, Kendall D. Isaac, *Is It "A" or Is It "The"? Deciphering the Motivating-Factor Standard in Employment Discrimination and Retaliation Cases*, 1 TEX. A&M L. REV. 55, 74 (2013) ("*McDonnell Douglas* has never been overruled and remains widely utilized."); Barrett S. Moore, *Shifting the Burden: Genuine Disputes and Employment Discrimination Standards of Proof*, 35 U. ARK. LITTLE ROCK L. REV. 113, 123–29, 128 n.146 (2012) (noting a circuit split on the issue).

119. *See* Krieger, *supra* note 89, at 1182–83.

120. *See, e.g.*, Christine Jolls & Cass R. Sunstein, *The Law of Implicit Bias*, 94 CALIF. L. REV. 969, 978 n.45 (2006) (collecting sources); Linda Hamilton Krieger & Susan T. Fiske, *Behavioral Realism in Employment Discrimination Law: Implicit Bias and Disparate Treatment*, 94 CALIF. L. REV. 997, 1003 n.21 (2006) (collecting sources).

121. This example can be ported directly to data mining as overrepresentation in data collection. *See supra* Part I.B.2.

122. Samuel R. Bagenstos, *The Structural Turn and the Limits of Antidiscrimination Law*, 94 CALIF. L. REV. 1, 9 (2006).

123. Sullivan, *supra* note 114, at 1000. There is, however, no general agreement on whether the law should treat such discrimination as disparate treatment or disparate impact. *Compare* Krieger, *supra* note 89, at 1231 (explaining that because the bias causes employers to *treat* people differently, it

There are a few possible ways to analogize discriminatory data mining to unintentional disparate treatment in the traditional context, based on where one believes the "treatment" lies. Either the disparate treatment occurs at the decision to apply a predictive model that will treat members of a protected class differently, or it occurs when the disparate result of the model is used in the ultimate hiring decision. In the first scenario, the intent at issue is the decision to apply a predictive model with known disproportionate impact on protected classes. In the second, the disparate treatment occurs if, after the employer sees the disparate result, he proceeds anyway. If the employer continues *because* he liked the discrimination produced in either scenario, then intent is clear. If not, then this just devolves into a standard disparate impact scenario, with liability based on effect. Under disparate impact theory, deciding to follow through on a test with discriminatory effect does not suddenly render it disparate *treatment*.[124]

Another option is to imagine the *model* as the decision maker exhibiting implicit bias. That is, because of biases hidden to the predictive model such as nonrepresentative data or mislabeled examples, the model reaches a discriminatory result. This analogy turns every mechanism except proxy discrimination into the equivalent of implicit bias exhibited by individual decision makers. The effect of bias is one factor among the many different factors that go into the model-driven decision, just like in an individual's adverse employment decision.[125] Would a more expansive definition of motive fix this scenario?

Because the doctrine focuses on *human* decision makers as discriminators, the answer is no. Even if disparate treatment doctrine could capture unintentional discrimination, it would only address such discrimination stemming from human bias. For example, the person who came up with the idea for Street Bump ultimately devised a system that suffers from reporting bias,[126] but it was not because he or she was implicitly employing some racial stereotype. Rather, it was simply inattentiveness to problems with the sampling frame. This is not to say that his or her own bias had nothing to do with it—the person likely owned a smartphone and thus did not think about the people who do not—but no one would say that it was even implicit bias against protected

---

should be considered a disparate treatment violation), *with* Sullivan, *supra* note 114, at 969–71 (arguing that the purpose of disparate impact is a catch-all provision to address those types of bias that disparate treatment cannot reach). This disagreement is important and even more pronounced in the case of data mining. *See infra* Part III. For now, we assume each case can be analyzed separately.

124. In fact, after *Ricci v. DeStefano*, 557 U.S. 557 (2009), deciding *not* to apply such a test after noticing the discriminatory effect may give rise to a disparate treatment claim in the other direction.

125. Bagenstos, *supra* note 122, at 9; Krieger, *supra* note 89, at 1185–86 ("Not only disparate treatment analysis, but the entire normative structure of Title VII's injunction 'not to discriminate,' rests on the assumption that decisionmakers possess 'transparency of mind'—that they are aware of the reasons why they are about to make, or have made, a particular employment decision.").

126. *See supra* note 51 and accompanying text.

classes that motivated the decision, even under the expansive definition of the word "motive."[127]

The only possible analogy relevant to disparate treatment, then, is to those data mining mechanisms of unintentional discrimination that reflect a real person's bias—something like LinkedIn's Talent Match recommendation engine, which relies on potentially prejudiced human assessments of employees.[128] As a general rule, an employer may not avoid disparate treatment liability by encoding third-party preferences as a rationale for a hiring decision.[129] But, once again, to be found liable under current doctrine, the employer would likely both have to know that this is the specific failure mechanism of the model and choose it based on this fact.

There is one other interesting question regarding disparate treatment doctrine: whether the intent standard includes knowledge. This is not a problem that arises often when a human is making a single employment determination. Assuming disparate treatment occurs in a given case, it is generally either intended or unconscious. What would it mean to have an employer *know* that he was treating an employee differently, but still take the action he had always planned to take without *intent* to treat the employee differently? It seems like an impossible line to draw.[130]

With data mining, though, unlike unconscious bias, it is possible to audit the resulting model and inform an employer that she will be treating individuals differently before she does so. If an employer *intends* to employ the model, but *knows* it will produce a disparate impact, does she intend to discriminate? This is a more realistic parsing of intent and knowledge than in the case of an individual, nonsystematic employment decision. Neither pretext nor motive exists here, and throughout civil and criminal law, "knowledge" and "intent" are considered distinct states of mind, so there would likely be no liability. On the other hand, courts may use knowledge of discrimination as evidence to find intent.[131] And while the statute's language only covers intentional discrimination,[132] a broad definition of intent could include knowledge or

127. Of course, the very presumption of a design's neutrality is itself a bias that may work against certain people. *See* Langdon Winner, *Do Artifacts Have Politics?*, 109 DAEDALUS 121, 125 (1980). But, as this is a second-order effect, we need not address it here.

128. *See* Woods, *supra* note 46.

129. *See* 29 C.F.R. § 1604.2(a)(1)(iii) (2015) (stating the EEOC's position that "the preferences of coworkers, the employer, clients or customers" cannot be used to justify disparate treatment); *see also* Fernandez v. Wynn Oil Co., 653 F.2d 1273, 1276–77 (9th Cir. 1981); Diaz v. Pan Am. World Airways, Inc., 442 F.2d 385, 389 (5th Cir. 1971).

130. *See* Krieger, *supra* note 89, at 1185 (discussing disparate treatment's "assumption of decisionmaker self-awareness").

131. Columbus Bd. of Educ. v. Penick, 443 U.S. 449, 464 (1979) ("[A]ctions having foreseeable and anticipated disparate impact are relevant evidence to prove the ultimate fact, forbidden purpose."); Pers. Adm'r of Mass. v. Feeney, 442 U.S. 256, 279 n.25 (1979) ("[W]hen the adverse consequences of a law upon an identifiable group are . . . inevitable . . . , a strong inference that the adverse effects were desired can reasonably be drawn.").

132. 42 U.S.C. § 2000e-2(h) (2012).

substantial certainty of the result.[133] Because the situation has not come up often, the extent of the "intent" required is as yet unknown.[134]

In sum, aside from rational racism and masking (with some difficulties), disparate treatment doctrine does not appear to do much to regulate discriminatory data mining.

B. Disparate Impact

Where there is no discriminatory intent, disparate impact doctrine should be better suited to finding liability for discrimination in data mining. In a disparate impact case, a plaintiff must show that a particular facially neutral employment practice causes a disparate impact with respect to a protected class.[135] If shown, the defendant-employer may "demonstrate that the challenged practice is job related for the position in question and consistent with business necessity."[136] If the defendant makes a successful showing to that effect, the plaintiff may still win by showing that the employer could have used an "alternative employment practice" with less discriminatory results.[137]

The statute is unclear as to the required showing for essentially every single element of a disparate impact claim. First, it is unclear how much disparate impact is needed to make out a prima facie case.[138] The EEOC, charged with enforcing Title VII's mandate, has created the so-called "four-fifths rule" as a presumption of adverse impact: "A selection rate for any race, sex, or ethnic group which is less than four-fifths . . . of the rate for the group

133. *See* Julia Kobick, Note, *Discriminatory Intent Reconsidered: Folk Concepts of Intentionality and Equal Protection Jurisprudence*, 45 HARV. C.R.-C.L. L. REV. 517, 551 (2010) (arguing that courts should regularly consider knowledge and foreseeability of disparate impact as an intended effect); *cf.* RESTATEMENT (SECOND) OF TORTS § 8A cmt. b (AM. LAW INST. 1965) ("Intent is not . . . limited to consequences which are desired. If the actor knows that the consequences are certain, or substantially certain, to result from his act, and still goes ahead, he is treated by the law as if he had in fact desired to produce the result.").

134. Determining that a model is discriminatory is also like trying and failing to validate a test under disparate impact doctrine. *See infra* Part II.B. If a test fails validation, the employer using it would know that he is discriminating if he applies it, but that does not imply that he is subject to disparate treatment liability. Nonetheless, validation is part of the business necessity defense, and that defense is not available against disparate treatment claims. Thus, the analysis does not necessarily have the same result. 42 U.S.C. § 2000e-2(k)(2). One commentator has argued that including knowledge as a state of mind leading to disparate treatment liability would effectively collapse disparate impact and disparate treatment by conflating intent and effect. Jessie Allen, *A Possible Remedy for Unthinking Discrimination*, 61 BROOK. L. REV. 1299, 1314 (1995). But others still have noted that with respect to knowledge, a claim is still about the *treatment* of an individual, not the incidental disparate impact of a neutral policy. *See* Carin Ann Clauss, *Comparable Worth—The Theory, Its Legal Foundation, and the Feasibility of Implementation*, 20 U. MICH. J.L. REFORM 7, 62 (1986).

135. 42 U.S.C. § 2000e-2(k)(1)(A).

136. *Id.*

137. *Id.*

138. The statute does not define the requirement and Supreme Court has never addressed the issue. *See, e.g.*, Sullivan, *supra* note 114, at 954 & n.153. For a brief discussion of the different approaches to establishing disparate impact, see Pamela L. Perry, *Two Faces of Disparate Impact Discrimination*, 59 FORDHAM L. REV. 523, 570–74 (1991).

with the highest rate will generally be regarded . . . as evidence of adverse impact."[139] The Uniform Guidelines on Employment Selection Procedures (Guidelines) also state, however, that smaller differences can constitute adverse impact and greater differences may not, depending on circumstances. Thus, the four-fifths rule is truly just a guideline.[140] For the purposes of this Part, it is worthwhile to just assume that the discriminatory effects are prominent enough to establish disparate impact as an initial matter.[141]

The next step in the litigation is the "business necessity" defense. This defense is, in a very real sense, the crux of disparate impact analysis, weighing Title VII's competing goals of limiting the effects of discrimination while allowing employers discretion to advance important business goals. *Griggs v. Duke Power Co.*[142]—the decision establishing the business necessity defense alongside disparate impact doctrine itself—articulated the defense in several different ways:

> A challenged employment practice must be "shown to be related to job performance," have a "manifest relationship to the employment in question," be "demonstrably a reasonable measure of job performance," bear some "relationship to job-performance ability," and/or "must measure the person for the job and not the person in the abstract."[143]

The Supreme Court was not clear on what, if any, difference existed between job-relatedness and business necessity, at one point seeming to use the terms interchangeably: "The touchstone is business necessity. If an employment practice which operates to exclude Negroes cannot be shown to be related to job performance, the practice is prohibited."[144] The focus of the Court was clearly on future job performance, and the term "job-related" has come to mean a practice that is predictive of job performance.[145] Because the definitions of job-relatedness and business necessity have never been clear, courts defer when applying the doctrine and finding the appropriate balance.[146]

Originally, the business necessity defense seemed to apply narrowly. In *Griggs*, Duke Power had instituted new hiring requirements including a high school diploma and success on a "general intelligence" test for previously

---

139. Uniform Guidelines on Employment Selection Procedures, 29 C.F.R. § 1607.4(D) (2015) [hereinafter Guidelines].

140. *Id.*

141. We will return to this when discussing the need to grapple with substantive fairness. *See infra* Part III.B.

142. 401 U.S. 424 (1971).

143. Linda Lye, Comment, *Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defense*, 19 BERKELEY J. EMP. & LAB. L. 315, 321 (1998) (footnotes omitted) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431–36 (1971)).

144. *Griggs*, 401 U.S. at 431; *see also* Lye, *supra* note 143, at 320.

145. Lye, *supra* note 143, at 355 & n.206.

146. *Id.* at 319–20, 348–53; Amy L. Wax, *Disparate Impact Realism*, 53 WM. & MARY L. REV. 621, 633–34 (2011).

white-only divisions. Duke Power did not institute such requirements in divisions where it had previously hired black employees.[147] The Court ruled that the new requirements were not a business necessity because "employees who have not completed high school or taken the tests have continued to perform satisfactorily and make progress in departments for which the high school and test criteria are now used."[148] Furthermore, the requirements were implemented without any study of their future effect.[149] The Court also rejected the argument that the requirements would improve the "overall quality of the workforce."[150]

By 1979, the Court began treating business necessity as a much looser standard.[151] In *New York City Transit Authority v. Beazer*,[152] the transit authority had implemented a rule barring drug users from employment, including current users of methadone, otherwise known as *recovering* heroin addicts. In dicta, the Court stated that a "narcotics rule," which "significantly serves" the "legitimate employment goals of safety and efficiency," was "assuredly" job related.[153] This was the entire analysis of the business necessity defense in the case. Moreover, the rationale was acceptable as applied to the entire transit authority, even where only 25 percent of the jobs were labeled as "safety sensitive."[154] Ten years later, the Court made the business necessity doctrine even more defendant-friendly in *Wards Cove Packing Co. v. Atonio*.[155] After *Wards Cove*, the business necessity defense required a court to engage in "a reasoned review of the employer's justification for his use of the challenged practice. . . . [T]here is no requirement that the challenged practice be 'essential' or 'indispensable' to the employer's business for it to pass muster . . . ."[156] The Court also reallocated the burden to plaintiffs to prove that business necessity was lacking and even referred to the defense as a "business justification" rather than a business necessity.[157] The *Wards Cove* Court went so far that Congress directly addressed the decision in the Civil Rights Act of 1991 (1991 Act), which codified disparate impact and reset the standards to the day before *Wards Cove* was decided.[158]

Because the substantive standards for job-relatedness or business necessity were uncertain before *Wards Cove*, however, the confusion persisted

147. *Griggs*, 401 U.S. at 427–28.
148. *Id.* at 431–32.
149. *Id.* at 432.
150. *Id.* at 431.
151. *See* Nicole J. DeSario, *Reconceptualizing Meritocracy: The Decline of Disparate Impact Discrimination Law*, 38 HARV. C.R.-C.L. L. REV. 479, 495–96 (2003); Lye, *supra* note 143, at 328.
152. 440 U.S. 568 (1979).
153. *Id.* at 587 & n.31.
154. *Id.*
155. 490 U.S. 642 (1989).
156. *Id.* at 659.
157. *Id.*
158. 42 U.S.C. § 2000e-2(k)(1)(C) (2012).

even after the 1991 Act was passed.[159] At the time, both sides—civil rights groups and the Bush administration, proponents of a rigorous and more lenient business necessity defense respectively—declared victory.[160]

Since then, courts have recognized that business necessity lies somewhere in the middle of two extremes.[161] Some courts require that the hiring criteria bear a "manifest relationship"[162] to the employment in question or that they be "significantly correlated" to job performance.[163] The Third Circuit was briefly an outlier, holding "that hiring criteria must effectively measure the 'minimum qualifications for successful performance of the job'" in order to meet the strict business necessity standard.[164] This tougher standard would, as a practical matter, ban general aptitude tests with any disparate impact because a particular cutoff score cannot be shown to distinguish between those able and completely unable to do the work.[165] For example, other unmeasured skills and abilities could theoretically compensate for the lower score on an aptitude test, rendering a certain minimum score not "necessary" if it does not measure minimum qualifications.[166] In a subsequent case, however, the Third Circuit recognized that Title VII does not require an employer to choose someone "less qualified" (as opposed to unqualified) in the name of nondiscrimination and noted that aptitude tests can be legitimate hiring tools if they accurately measure a person's qualifications.[167] The court concluded:

---

159. Legislative history was no help either. The sole piece of legislative history is an "interpretive memorandum" that specifies that the standards were to revert to before *Wards Cove*, coupled with an explicit instruction in the Act to ignore any other legislative history regarding business necessity. Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases*, 30 GA. L. REV. 387, 392–93 (1996).

160. Andrew C. Spiropoulos, *Defining the Business Necessity Defense to the Disparate Impact Cause of Action: Finding the Golden Mean*, 74 N.C. L. REV. 1479, 1484 (1996).

161. Though courts generally state the standard to reflect this middle position, the Supreme Court's latest word on disparate impact—in which the Court reaffirmed the doctrine generally and held that it applied in the Fair Housing Act—included the decidedly defendant-friendly observation that "private policies are not contrary to the disparate-impact requirement unless they are 'artificial, arbitrary, and unnecessary barriers.'" Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Project, Inc.,135 S. Ct. 2507, 2512 (2015) (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971)).

162. *See, e.g.*, Gallagher v. Magner, 619 F.3d 823, 834 (8th Cir. 2010); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 265 (4th Cir. 2005).

163. Gulino v. N.Y. State Educ. Dep't, 460 F.3d 361, 383 (2d Cir. 2006) (noting that hiring criteria are "significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated" (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 431 (1975))).

164. El v. Se. Pa. Transp. Auth., 479 F.3d 232, 242 (3d Cir. 2007) (quoting Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478, 481 (3d Cir. 1999)).

165. Michael T. Kirkpatrick, *Employment Testing: Trends and Tactics*, 10 EMP. RTS. & EMP. POL'Y J. 623, 633 (2006).

166. *Id.* Note, though, that this is similar to arguing that there is a less discriminatory alternative employment practice. This argument, then, would place the burden of the alternative employment practice prong on the defendant, contravening the burden-shifting scheme in the statute. *See infra* notes 170–74 and accompanying text.

167. *El*, 479 F.3d at 242.

> Putting these standards together, then, we require that employers show that a discriminatory hiring policy accurately—but not perfectly—ascertains an applicant's ability to perform successfully the job in question. In addition, Title VII allows the employer to hire the applicant most likely to perform the job successfully over others less likely to do so.[168]

Thus, all circuits seem to accept varying levels of job-relatedness rather than strict business necessity.[169]

The last piece of the disparate impact test is the "alternative employment practice" prong. Shortly after *Griggs*, the Supreme Court decided *Albemarle Paper Co. v. Moody*, holding in part that "[i]f an employer does then meet the burden of proving that its tests are 'job related,' it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'"[170] This burden-shifting scheme was codified in the 1991 Act as the "alternative employment practice" requirement.[171] Congress did not define the phrase, and its substantive meaning

---

168. *Id.*

169. Interestingly, it seems that many courts read identical business necessity language in the Americans with Disabilities Act to refer to a minimum qualification standard. *See, e.g.*, Sullivan v. River Valley Sch. Dist., 197 F.3d 804, 811 (6th Cir. 1999) ("[T]here must be significant evidence that could cause a reasonable person to inquire as to whether an employee is still capable of performing his job. An employee's behavior cannot be merely annoying or inefficient to justify an examination; rather, there must be genuine reason to doubt whether that employee can 'perform job-related functions.'" (quoting 42 U.S.C. § 12112(d)(4)(B))). Presumably, this is because disability, when compared to race or sex, more immediately raises questions regarding a person's ability to perform a job. Ironically, however, this means that disparate impact will be *more* tolerated where it is less likely to be obviously justified. Christine Jolls has in fact argued that disparate impact is, to a degree, functionally equivalent to accommodations law. Jolls, *supra* note 90, at 652.

170. 422 U.S. 405, 425 (1975) (quoting McDonnell Douglas Corp. v. Green, 411 U.S. 792, 801 (1973)).

171. 42 U.S.C. § 2000e-2(k)(1)(A) (2012). The "alternative employment practice" test has not always been treated as a separate step. *See, e.g.*, Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 659 (1989) (treating the alternative employment practice test as part of the "business justification" phase); Dothard v. Rawlinson, 433 U.S. 321, 332 (1977) (treating the alternative employment practice test as a narrow tailoring requirement for the business necessity defense). The *Albemarle* Court, though creating a surrebuttal and thus empowering plaintiffs, seemed to regard the purpose of disparate impact as merely smoking out pretexts for intentional discrimination. 422 U.S. at 425; *see also* Primus, *supra* note 98, at 537. If the *Albemarle* Court's approach is correct, treating the alternative employment practice requirement as a narrow tailoring requirement does make sense, much as the narrow tailoring requirement of strict scrutiny in equal protection serves the function of smoking out invidious purpose. City of Richmond v. J.A. Croson Co., 488 U.S. 469, 493 (1989); Rubenfeld, *supra* note 99, at 428.

Every circuit to address the question, though, has held that the 1991 Act returned the doctrine to the *Albemarle* burden-shifting scheme. Jones v. City of Boston, 752 F.3d 38, 54 (1st Cir. 2014); Howe v. City of Akron, 723 F.3d 651, 658 (6th Cir. 2013); Tabor v. Hilti, Inc., 703 F.3d 1206, 1220 (10th Cir. 2013); Puffer v. Allstate Ins. Co., 675 F.3d 709, 717 (7th Cir. 2012); Gallagher v. Magner, 619 F.3d 823, 833 (8th Cir. 2010); Gulino v. N.Y. State Educ. Dep't, 460 F.3d 361, 382 (2d Cir. 2006); Int'l Bhd. of Elec. Workers Local Unions Nos. 605 & 985 v. Miss. Power & Light Co., 442 F.3d 313, 318 (5th Cir. 2006); Anderson v. Westinghouse Savannah River Co., 406 F.3d 248, 277

remains uncertain. *Wards Cove* was the first case to use the specific phrase, so Congress's instruction to reset the law to the pre–*Wards Cove* standard is particularly perplexing.[172] The best interpretation is most likely *Albemarle*'s reference to "other tests or selection devices, without a similarly undesirable racial effect."[173] But this interpretation is slightly odd because in *Albemarle*, business necessity was still somewhat strict, and it is hard to imagine a business practice that is "necessary" while there exists a less discriminatory alternative that is just as effective.[174] If business necessity or job-relatedness is a less stringent requirement, though, then the presence of the alternative employment practice requirement does at least give it some teeth.

Now return to data mining. For now, assume a court does not apply the strict business necessity standard but has some variation of "job related" in mind (as all federal appellate courts do today).[175] The threshold issue is clearly whether the sought-after trait—the target variable—is job related, regardless of the machinery used to predict it. If the target variable is not sufficiently job related, a business necessity defense would fail, regardless of the fact that the decision was made by algorithm. Thus, disparate impact liability can be found for improper care in target variable definition. For example, it would be difficult for an employer to justify an adverse determination based on the appearance of an advertisement suggesting a criminal record alongside the search results for a candidate's name. Sweeney found such a search to have a disparate impact,[176] and the EEOC and several federal courts have interpreted Title VII to prohibit discrimination on the sole basis of criminal record, unless there is a specific reason the particular conviction is related to the job.[177] This

(4th Cir. 2005); Ass'n of Mexican-Am. Educators v. California, 231 F.3d 572, 584 (9th Cir. 2000); EEOC v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); Lanning v. Se. Pa. Transp. Auth., 181 F.3d 478, 485 (3d Cir. 1999). The D.C. Circuit has not explicitly observed that a burden-shifting framework exists.

172. Sullivan, *supra* note 114, at 964; Michael J. Zimmer, *Individual Disparate Impact Law: On the Plain Meaning of the 1991 Civil Rights Act*, 30 LOY. U. CHI. L.J. 473, 485 (1999).

173. *Albemarle*, 422 U.S. at 425; *accord, e.g.*, *Jones*, 752 F.3d at 53 (citing *Albemarle* to find meaning in the 1991 Act's text); *Allen v. City of Chicago*, 351 F.3d 306, 312 (7th Cir. 2003) (same, but with a "see also" signal).

174. William R. Corbett, *Fixing Employment Discrimination Law*, 62 SMU L. REV. 81, 92 (2009).

175. The difference would be whether mining for a single job-related trait, rather than a holistic ranking of "good employees," is permissible at all. *See infra* text accompanying notes 197–99.

176. Sweeney, *supra* note 41, at 51.

177. *See* El v. Se. Pa. Transp. Auth., 479 F.3d 232, 243 (3d Cir. 2007) (finding that though the criminal record policy had a disparate impact, it satisfied business necessity in that case); Green v. Mo. Pac. R.R., 523 F.2d 1290, 1298 (8th Cir. 1975); McCain v. United States, No. 2:14-cv-92, 2015 WL 1221257, at *17 (D. Vt. Mar. 17, 2015); EQUAL EMP'T OPPORTUNITY COMM'N, CONSIDERATION OF ARREST AND CONVICTION RECORDS IN EMPLOYMENT DECISIONS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (2012), http://www.eeoc.gov/laws/guidance/upload/arrest_conviction.pdf [https://perma.cc/JY47-2HVT]; *see also* Univ. of Tex. Sw. Med. Ctr. v. Nassar, 133 S. Ct. 2517, 2540 (2013) ("The position set out in the EEOC's guidance and compliance manual merits respect."); Michael Connett, Comment, *Employer Discrimination Against Individuals with a Criminal Record: The Unfulfilled Role of State Fair Employment Agencies*, 83 TEMP. L. REV. 1007, 1017 & nn.82–83

is true independent of the fact that the disparity is an artifact of third-party bias; all that matters is whether the target variable is job related. In the end, though, because determining that a business practice is not job related actually requires a normative determination that it is instead discriminatory, courts tend to accept most common business practices for which an employer has a plausible story.[178]

Once a target variable is established as job related, the first question is whether the model is predictive of that trait. The nature of data mining suggests that this will be the case. Data mining is designed entirely to predict future outcomes, and, if seeking a job-related trait, future job performance. One commentator lamented that "[f]ederal case law has shifted from a prospective view of meritocracy to a retrospective view, thereby weakening disparate impact law."[179] The author meant that, in *Griggs*, the Court recognized that education and other external factors were unequal and therefore discounted a measure of meritocracy that looked to past achievements, in favor of comparing the likelihood of future ones. But by the time the Court had decided *Wards Cove*, it had shifted to a model of retrospective meritocracy that presumed the legitimacy of past credentials, thus upholding the status quo.[180] While data mining must take the past—represented by the training data—as given, it generates predictions about workplace success that are much more accurate than predictions based on those past credentials that disparate impact doctrine has come to accept.[181] In a hypothetical perfect case of data mining, the available information would be rich enough that reliance on the past information would fully predict future performance. Thus, robust data mining would likely satisfy even the *Griggs* Court's standard that the models are looking toward future job performance, not merely past credentials.

The second question asks whether the model adequately predicts what it is supposed to predict. In the traditional context, this question arises in the case of general aptitude tests that might end up measuring unrelated elements of cultural awareness rather than intelligence.[182] This is where the different data

(2011) (citing EQUAL EMP'T OPPORTUNITY COMM'N, POLICY STATEMENT ON THE ISSUE OF CONVICTION RECORDS UNDER TITLE VII OF THE CIVIL RIGHTS ACT OF 1964 (1987), http://www.eeoc.gov/policy/docs/convict1.html [https://perma.cc/PY24-V8V7]). *But see, e.g.*, Manley v. Invesco, 555 Fed. App'x 344, 348 (5th Cir. 2014) (per curiam) ("Persons with criminal records are not a protected class under Title VII.").

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

179. DeSario, *supra* note 151, at 481.

180. *Id.* at 493; *see also infra* Conclusion.

181. *See* Don Peck, *They're Watching You at Work*, ATLANTIC (Nov. 20, 2013), http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681 [https://perma.cc/JFP8-CZKC] (discussing Google's choice to abandon traditional hiring metrics because they are not good predictors of performance).

182. *See, e.g.*, Griggs v. Duke Power Co., 420 F.2d 1225, 1239 n.6 (4th Cir. 1970), *rev'd*, 401 U.S. 424 (1971) ("Since for generations blacks have been afforded inadequate educational opportunities and have been culturally segregated from white society, it is no more surprising that their

mining mechanisms for discriminatory effects matter. Part I posited that proxy discrimination optimizes correctly. So if it evidences a disparate impact, it reflects unequal distribution of relevant traits in the real world. Therefore, proxy discrimination will be as good a job predictor as possible given the current shape of society. Models trained on biased samples and mislabeled examples, on the other hand, will result in correspondingly skewed assessments rather than reflect real-world disparities. The same effect may be present in models that rely on insufficiently rich or insufficiently granular datasets: by designation they do not reflect reality. These models might or might not be considered job related, depending on whether the errors distort the outcomes enough that the models are no longer good predictors of job performance.

The Guidelines have set forth validation procedures intended to create a job-relatedness standard. Quantifiable tests that have a disparate impact must be validated according to the procedures in the Guidelines if possible; otherwise, a presumption arises that they are not job related.[183] Under the Guidelines, a showing of validity takes one of three forms: criterion-related, content, or construct.[184] Criterion-related validity "consist[s] of empirical data demonstrating that the selection procedure is predictive of or significantly correlated with important elements of job performance."[185] The "relationship between performance on the procedure and performance on the criterion measure is statistically significant at the 0.05 level of significance. . . ."[186] Content validity refers to testing skills or abilities that generally are or have been learned on the job, though not those that could be acquired in a "brief orientation."[187] Construct validity refers to a test designed to measure some innate human trait such as honesty. A user of a construct "should show by empirical evidence that the selection procedure is validly related to the construct and that the construct is validly related to the performance of critical or important work behavior(s)."[188]

As a statistical predictive measure, a data mining model could be validated by either criterion-related or construct validity, depending on the trait being sought. Either way, there must be statistical significance showing that the result of the model correlates to the trait (which was already determined to be an important element of job performance). This is an exceedingly low bar for data mining because data mining's predictions necessarily rest on demonstrated

---

performance on 'intelligence' tests is significantly different than whites' than it is that fewer blacks have high school diplomas.").

183. 29 C.F.R. §§ 1607.3, 1607.5 (2015). The Guidelines also cite two categories of practices that are unsuitable for validation: informal, unscored practices and technical infeasibility. *Id.* § 1607.6(B). For the latter case, the Guidelines state that the selection procedure still should be justified somehow or another option should be chosen.

184. *Id.* § 1607.5(B).

185. *Id.*

186. *Id.* § 1607.14(B)(5).

187. *Id.* §§ 1607.5(F), 1607.14(C).

188. *Id.* § 1607.14(D)(3).

statistical relationships. Data mining will likely only be used if it is actually predictive of *something*, so the business necessity defense solely comes down to whether the trait sought is important enough to job performance to justify its use in any context.

Even assuming the Guidelines' validation requirement is a hurdle for data mining, some courts ignore the Guidelines' recommendation that an unvalidated procedure be rejected, preferring to rely on "common sense" or finding a "manifest relationship" between the criteria and successful job performance.[189] Moreover, it is possible that the Supreme Court inadvertently overruled the Guidelines in 2009. In *Ricci v. Destefano*, a case that will be discussed in greater detail in Part III.B, the Court found no genuine dispute that the tests at issue met the job-related and business necessity standards[190] despite not having been validated under the Guidelines and despite the employer *actively denying* that they could be validated.[191] While the business necessity defense was not directly at issue in *Ricci*, "[o]n the spectrum between heavier and lighter burdens of justification, the Court came down decidedly in favor of a lighter burden."[192]

Thus, there is good reason to believe that any or all of the data mining models predicated on legitimately job-related traits pass muster under the business necessity defense. Models trained on biased samples, mislabeled examples, and limited features, however, might trigger liability under the alternative employment practice prong. If a plaintiff can show that an alternative, less discriminatory practice that accomplishes the same goals exists and that the employer "refuses" to use it, the employer can be found liable. In this case, a plaintiff could argue that the obvious alternative employment practice would be to fix the problems with the models.

Fixing the models, however, is not a trivial task. For example, in the LinkedIn hypothetical, where the demonstrated interest in different kinds of employees reflects employers' prejudice, LinkedIn is the party that determines the algorithm by which the discrimination occurs (in this case, based on reacting to third-party preferences). If an employer were to act on the recommendations suggested by the LinkedIn recommendation engine, there

189. Wax, *supra* note 146, at 633–34.

190. David A. Drachsler, *Assessing the Practical Repercussions of* Ricci, AM. CONST. SOC'Y BLOG (July 27, 2009), http://www.acslaw.org/node/13829 [https://perma.cc/AH9G-B3GN] (observing that the Court in *Ricci v. DeStefano* found no genuine dispute that the unvalidated tests at issue met the job-related and business necessity standards despite the Guidelines creating a presumption of invalidity for unvalidated tests that are discriminatory).

191. New Haven's primary argument was that it had to withdraw the tests or it would have faced Title VII liability. *See* Mark S. Brodin, Ricci v. DeStefano*: The New Haven Firefighters Case & the Triumph of White Privilege*, 20 S. CAL. REV. L. & SOC. JUST. 161, 178 n.128 (2011) ("New Haven forcefully argued throughout the litigation that the exams were 'flawed' and may not have identified the most qualified candidates for the supervisory positions.").

192. George Rutherglen, Ricci v. Destefano*: Affirmative Action and the Lessons of Adversity*, 2009 SUP. CT. REV. 83, 107.

would not be much he could do to make it less reflective of third-party prejudice, aside from calling LinkedIn and asking nicely. Thus, it could not really be said that the employer "refuses" to use an alternative employment practice. The employer could either use the third-party tool or not. Similarly, it might be possible to fix an app like Street Bump that suffers from reporting bias, but the employer would need access to the raw input data in order to do so.[193] In the case of insufficiently rich or granular features, the employer would need to collect more data in order to make the model more discerning. But collecting more data can be time consuming and costly,[194] if not impossible for legal or technical reasons.

Moreover, the under- and overrepresentation of members of protected classes in data is not always evident, nor is the mechanism by which such under- or overrepresentation occurs. The idea that the representation of different social groups in the dataset can be brought into proportions that better match those in the real world presumes that analysts have some independent mechanism for determining these proportions. Thus, there are several hurdles to finding disparate impact liability for models employing data that under- or overrepresents members of protected classes. The plaintiff must prove that the employer created or has access to the model, can discover that there is discriminatory effect, and can discover the particular mechanism by which that effect operates. The same can be said for models with insufficiently rich feature sets. Clearly there are times when more features would improve an otherwise discriminatory outcome. But it is, almost by definition, hard to know which features are going to make the model more or less discriminatory. Indeed, it is often impossible to know which features are missing because data miners do not operate with causal relationships in mind. So while theoretically a less discriminatory alternative would almost always exist, proving it would be difficult.

There is yet another hurdle. Neither Congress nor courts have specified what it means for an employer to "refuse" to adopt the less discriminatory procedure. Scholars have suggested that perhaps the employer cannot be held liable until it has considered the alternative and rejected it.[195] Thus, if the employer has run an expensive data collection and analysis operation without ever being made aware of its any discriminatory tendencies, and the employer cannot afford to re-run the entire operation, is the employer "refusing" to use a less discriminatory alternative, or does one simply not exist? How much would the error correction have to cost an employer before it is not seen as a refusal to use the procedure?[196] Should the statute actually be interpreted to mean that an

193. *See infra* Part III.B.1.

194. *See generally* Dalessandro, Perlich & Raeder, *supra* note 68.

195. Sullivan, *supra* note 114, at 964; Zimmer, *supra* note 172, at 505–06.

196. For a discussion of courts using cost as a rationale here, see Ernest F. Lidge III, *Financial Costs as a Defense to an Employment Discrimination Claim*, 58 ARK. L. REV. 1, 32–37 (2005).

employer "unreasonably refuses" to use an alternative employment practice? These are all difficult questions, but suffice it to say, the prospect of winning a data mining discrimination case on alternative employment practice grounds seems slim.

The third and final consideration regarding disparate impact liability for data mining is whether a court or Congress might reinvigorate strict business necessity.[197] In that case, things look a little better for plaintiffs bringing disparate impact claims. Where an employer models job tenure,[198] for example, a court may be inclined to hold that it is job related because the model is a "legitimate, non-discriminatory business objective."[199] But it is clearly not necessary to the job. The same reasoning applies to mining for any single trait that is job related—the practice of data mining is not focused on discovering make-or-break skills. Unless the employer can show that below the cut score, employees cannot do the work, then the strict business necessity defense will fail. Thus, disparate impact that occurs as an artifact of the problem-specification stage can potentially be addressed by strict business necessity.

This reasoning is undermined, though, where employers do not mine for a single trait, but automate their decision process by modeling job performance on a holistic measure of what makes good employees. If employers determine traits of a good employee by simple ratings, and use data mining to appropriately divine good employees' characteristics among several different variables, then the argument that the model does not account for certain skills that could compensate for the employee's failings loses its force. Taken to an extreme, an 8,000-feature holistic determination of a "good employee" would still not be strictly "necessary." Holding a business to such a standard, however, would simply be forbidding that business from ranking candidates if any disparate impact results. Thus, while the strict business necessity defense could prevent myopic employers from creating disparate impacts by their choice of target variable, it would still not address forms of data mining that model general job performance rather than predict specific traits.

Disparate impact doctrine was created to address unintentional discrimination. But it strikes a delicate balance between allowing businesses the leeway to make legitimate business judgments and preventing "artificial, arbitrary, and unnecessary" discrimination.[200] Successful data mining operations will often both predict future job performance and have some

197. This would likely require Congressional action because strict business necessity essentially transfers the burden to prove a lack of an alternative employment practice to the defense. By implication, if a practice is "necessary," there cannot be alternatives. The statute, as it reads now, clearly states that the plaintiff has the burden for that prong. 42 U.S.C. § 2000e-2(k)(1)(A)(ii) (2012).

198. This is an increasingly common practice in low-wage, high-turnover jobs. *See* Peck, *supra* note 181.

199. Equal Emp't Opportunity Comm'n v. Joe's Stone Crab, Inc., 220 F.3d 1263, 1275 (11th Cir. 2000); *see also* Gallagher v. Magner, 619 F.3d 823, 834 (8th Cir. 2010).

200. Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

disparate impact. Unless the plaintiff can find an alternative employment practice to realistically point to, a tie goes to the employer.

C. Masking and Problems of Proof

Masking poses separate problems for finding Title VII liability. As discussed earlier, there is no theoretical problem with finding liability for masking.[201] It is a disparate treatment violation as clear as any. But like traditional forms of intentional discrimination, it suffers from difficulties of proof. While finding intent from stray remarks or other circumstantial evidence is challenging in any scenario, masking presents additional complications for detection.

Data mining allows employers who wish to discriminate on the basis of a protected class to disclaim any knowledge of the protected class in the first instance while simultaneously inferring such details from the data. An employer may want to discriminate by using proxies for protected classes, such as in the case of redlining.[202] Due to housing segregation, neighborhood is a good proxy for race and can be used to redline candidates without reference to race.[203] This is a relatively unsophisticated example, however. It is possible that some combination of musical tastes,[204] stored "likes" on Facebook,[205] and network of friends[206] will reliably predict membership in protected classes. An employer can use these traits to discriminate by setting up future models to sort by these items and then disclaim any knowledge of such proxy manipulation.

More generally, as discussed in Part I, any of the mechanisms by which unintentional discrimination can occur can also be employed intentionally. The example described above is intentional discrimination by proxy, but it is also possible to intentionally bias the data collection process, purposefully mislabel examples, or deliberately use an insufficiently rich set of features,[207] though some of these would probably require a great deal of sophistication. These methods of intentional discrimination will look, for all intents and purposes, identical to the unintentional discrimination that can result from data mining. Therefore, detecting discrimination in the first instance will require the same techniques as detecting unintentional discrimination, namely a disparate impact analysis. Further, assuming there is no circumstantial evidence like an employer's stray remarks with which to prove intent, a plaintiff might attempt

201. *See supra* text accompanying notes 106–07.

202. *See supra* Part I.E.

203. *See* MASSEY & DENTON, *supra* note 73, at 51–52.

204. Croll, *supra* note 88.

205. Michal Kosinski, David Stillwell & Thore Graepel, *Private Traits and Attributes Are Predictable from Digital Records of Human Behavior*, 110 PROC. NAT'L ACAD. SCI. 5802 (2013).

206. Carter Jernigan & Behram F.T. Mistree, *Gaydar: Facebook Friendships Expose Sexual Orientation*, FIRST MONDAY (Oct. 5, 2009), http://firstmonday.org/article/view/2611/2302 [https://perma.cc/G36G-S26X].

207. *See* Dwork et al., *supra* note 81, app. at 226 ("Catalog of Evils").

to prove intent by demonstrating that the employer is using less representative data, poorer examples, or fewer and less granular features than he might otherwise use were he interested in the best possible candidate. That is, one could show that the neutral employment practice is a pretext by demonstrating that there is a more predictive alternative.

This looks like disparate impact analysis. A plaintiff proving masked intentional discrimination asks the same question as in the "alternative employment practice" prong: whether there were more relevant measures the employer could have used.[208] But the business necessity defense is not available in a disparate treatment case,[209] so alternative employment practice is not the appropriate analysis. Scholars have noted, though, that the line between disparate treatment and disparate impact in traditional Title VII cases is not always clear,[210] and sometimes employer actions can be legitimately categorized as either or both.[211] As Professor George Rutherglen has pointed out, "Concrete issues of proof, more than any abstract theory, reveal the fundamental similarity between claims of intentional discrimination and those of disparate impact. The evidence submitted to prove one kind of claim invariably can be used to support the other."[212] Rutherglen's point is exactly what must happen in the data mining context: disparate treatment and disparate impact become essentially the same thing from an evidentiary perspective.

To the extent that disparate impact and treatment are, in reality, different theories, they are often confused for each other. Plaintiffs will raise both types of claims as a catch-all because they cannot be sure on which theory they might win, so both theories will be in play in a given case.[213] As a result, courts often seek evidence of state of mind in disparate impact cases[214] and objective, statistical evidence in disparate treatment cases.[215] Assuming the two theories are not functionally the same, using the same evidence for disparate treatment and disparate impact will only lead to more confusion and, as a result, more uncertainty within the courts. Thus, despite its clear nature as a theoretical violation, it is less clear that a plaintiff will be able to win a masking disparate treatment case.

A final point is that traditionally, employers who do *not* want to discriminate go to great lengths to avoid raising the prospect that they have

208. *Cf.* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (creating an alternative employment practice prong for the purpose of rooting out pretext).

209. 42 U.S.C. § 2000e-2(k)(2) (2012).

210. George Rutherglen, *Disparate Impact, Discrimination, and the Essentially Contested Concept of Equality*, 74 FORDHAM L. REV. 2313, 2313 (2006); Stacy E. Seicshnaydre, *Is the Road to Disparate Impact Paved with Good Intentions?: Stuck on State of Mind in Antidiscrimination Law*, 42 WAKE FOREST L. REV. 1141, 1142–43 (2007).

211. Rutherglen, *supra* note 210, at 2320–21.

212. *Id.* at 2320.

213. Seicshnaydre, *supra* note 210, at 1147–48.

214. *Id.* at 1153–63.

215. Rutherglen, *supra* note 210, at 2321–22.

violated the law. Thus they tend to avoid collecting information about attributes that reveal an individual's membership in a protected class. Employers even pay third parties to collect relatively easy-to-find information on job applicants, such as professional honors and awards, as well as compromising photos, videos, or membership in online groups, so that the third party can send back a version of the report that "remove[s] references to a person's religion, race, marital status, disability and other information protected under federal employment laws."[216] This allows employers to honestly disclaim any knowledge of the protected information. Nonetheless, if an employer seeks to discriminate according to protected classes, she would be able to infer class membership from the data. Thus, employers' old defense to suspicion of discrimination—that they did not even see the information—is no longer adequate to separate would-be intentional discriminators from employers that do not intend to discriminate.

III.
THE DIFFICULTY FOR REFORMS

While each of the mechanisms for discrimination in data mining presents difficulties for Title VII as currently written, there are also certain obstacles to reforming Title VII to address the resulting problems. Computer scientists and others are working on technical remedies,[217] so to say that there are problems with legal remedies does not suggest that the problems with discrimination in data mining cannot be solved at all. Nonetheless, this Part focuses on the legal aspects. As it illustrates, even assuming that the political will to reform Title VII exists, potential legal solutions are not straightforward.

This Part discusses two types of difficulties with reforming Title VII. First, there are issues internal to the data mining process that make legal reform difficult. For example, the subjectivity in defining a "good employee" is unavoidable, but, at the same time, some answers are clearly less discriminatory than others.[218] How does one draw that line? Can employers gain access to the additional data necessary to correct for collection bias? How much will it cost them to find it? How do we identify the "correct" baseline historical data to avoid reproducing past prejudice or the "correct" level of detail and granularity in a dataset? Before laws can be reformed, policy-level answers to these basic technical, philosophical, and economic questions need to be addressed at least to some degree.

---

216. Jennifer Preston, *Social Media History Becomes a New Job Hurdle*, N.Y. TIMES (July 20, 2011), http://www.nytimes.com/2011/07/21/technology/social-media-history-becomes-a-new-job-hurdle.html [https://perma.cc/NZ8U-M296].

217. For a list of the wide-ranging research underway in computer science, see generally Resources, FAT ML, http://www.fatml.org/resources.html [https://perma.cc/T2QW-ARHX].

218. *See supra* Part I.A.

Second, reform will face political and constitutional constraints external to the logic of data mining that will affect how Title VII can be permissibly reformed to address it. Not all of the mechanisms for discrimination seem to be amenable to procedural remedies. If that holds true, only after-the-fact reweighting of results may be able to compensate for the discriminatory outcomes. This is not a matter of missing legislation; it is a matter of practical reality. Unfortunately, while in many cases no procedural remedy will be sufficient, any attempt to design a legislative or judicial remedy premised on reallocation of employment outcomes will not survive long in the current political or constitutional climate, as it raises the specter of affirmative action. Politically, 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.[219]

A. Internal Difficulties

1. Defining the Target Variable

Settling on a target variable is a necessarily subjective exercise.[220] Disputes over the superiority of competing definitions are often insoluble because the target variables are themselves incommensurable. There are, of course, easier cases, where prejudice or carelessness leads to definitions that subject members of protected classes to avoidably high rates of adverse determinations. But most cases are likely to involve genuine business disagreements over ideal definitions, with each having a potentially greater or lesser impact on protected classes. There is no stable ground upon which to judge the relative merits of definitions because they often reflect competing ideas about the very nature of the problem at issue.[221] As Professor Oscar Gandy has argued, "[C]ertain kind[s] of biases are inherent in the selection of the goals or objective functions that automated systems will [be] designed to support."[222] There is no escape from this situation; a target variable *must* reflect judgments about what really is the problem at issue in making hiring decisions. For certain employers, it might be rather obvious that the problem is one of reducing the administrative costs associated with turnover and training; for others, it might be improving sales; for still others, it might be increasing

219. *See* Lyle Denniston, *Argument Analysis: Now, Three Options on College Affirmative Action*, SCOTUSBLOG (Dec. 9, 2015, 2:47 PM), http://www.scotusblog.com/2015/12/argument-analysis-now-three-options-on-college-affirmative-action [https://perma.cc/XF75-N82F] (analysis of oral argument in Fisher v. Univ. of Tex., 758 F.3d 633 (5th Cir. 2014), *cert. granted*, 135 S. Ct. 2888, (June 29, 2015)); *see also* Fisher v. Univ. of Tex., 133 S. Ct. 2411, 2419 (2013) ("[A]ny official action that treats a person differently on account of his race or ethnic origin is inherently suspect." (internal citation omitted)).

220. *See supra* Part I.A.

221. *See* David J. Hand, *Deconstructing Statistical Questions*, 157 J. ROYAL STAT. SOC'Y. SERIES A (STAT. SOC'Y) 317, 318–20 (1994).

222. Gandy, *supra* note 31, at 39.

innovation. Any argument for the superiority of one target variable over the other will simply make appeals to competing and incommensurate values.

For these same reasons, however, defining the target variable also offers an opportunity for creative thinking about the potentially infinite number of ways of making sound hiring decisions. Data miners can experiment with multiple definitions that each seem to serve the same goal, even if these fall short of what they themselves consider ideal. In principle, employers should rely on proxies that are maximally proximate to the actual skills demanded of the job. While there should be a tight nexus between the sought-after features and these skills, this may not be possible for practical and economic reasons. This leaves data miners in a position to dream up many different nonideal ways to make hiring decisions that may have a greater or less adverse impact on protected classes.

The Second Circuit considered such an approach in *Hayden v. County of Nassau*.[223] In *Hayden*, the county's goal was to find a police entrance exam that was "valid, yet minimized the adverse impact on minority applicants."[224] The county thus administered an exam with twenty-five parts that could be scored independently. By design, a statistically valid result could be achieved by one of several configurations that counted only a portion of the test sections, without requiring all of them.[225] The county ended up using nine of the sections as a compromise, after rejecting one configuration that was more advantageous to minority applicants but less statistically sound.[226] This is a clear example of defining a problem in such a way that it becomes possible to reduce the disparate impact without compromising the accuracy of the assessment mechanism.

2. Training Data

a. Labeling Examples

Any solution to the problems presented by labeling must be a compromise between a rule that forbids employers from relying on past discrimination and one that allows them to base hiring decisions on historical examples of good employees. In theory, a rule that forbids employers from modeling decisions based on historical examples tainted by prejudice would address the problem of improper labeling. But if the only examples an employer has to draw on are those of past employees who had been subject to discrimination, all learned rules will recapitulate this discrimination.

Title VII has always had to balance its mandate to eliminate discrimination in the workplace with employers' legitimate discretion. For

223. 180 F.3d 42, 47 (2d Cir. 1999).
224. *Id.*
225. *Id.*
226. *Id.*

example, one of the most common selection procedures that explicitly reproduced past discrimination was seniority.[227] Seniority was, and is still often, a legitimate metric for promotion and is especially important in collective bargaining. After the passage of Title VII, however, seniority was also often used to keep black people from advancing to better jobs because they had not been hired until Title VII forced employers to hire them.[228] Despite this obvious problem with seniority, Title VII contains an explicit carve-out for "bona fide seniority or merit system[s]."[229] As a result, the Supreme Court has held that "absent a discriminatory purpose, the operation of a seniority system cannot be an unlawful employment practice even if the system has some discriminatory consequences."[230] Given the inherent tension between ensuring that past discrimination is not reproduced in future decisions and permitting employers legitimate discretion, it should be unsurprising that, when translated to data mining, the problem is not amenable to a clear solution.

In fact, this difficulty is even more central to data mining. Data miners who attempt to remove the influence of prejudice on prior decisions by recoding or relabeling examples may find that they cannot easily resolve what the nonprejudicial determination would have been. As Calders and Žliobaitė point out, "[T]he notion of what is the correct label is fuzzy."[231] Employers are unlikely to have perfectly objective and exhaustive standards for hiring; indeed, part of the hiring process is purposefully subjective. At the same time, employers are unlikely to have discriminated so completely in the past that the only explanation for rejecting an applicant was membership in protected classes. This leaves data miners tasked with correcting for prior prejudice with the impossible challenge of determining what the correct subjective employment decision would have been absent prejudice. Undoing the imprint of prejudice on the data may demand a complete rerendering of the biased decisions rather than simply adjusting those decisions according to some fixed statistical measure.

b. Data Collection

Although there are some cases with obviously skewed datasets that are relatively easy to identify and correct, often the source and degree of the bias will not be immediately apparent.[232] Street Bump suffered from a visually

227. Selmi, *supra* note 178, at 715.

228. *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 450 (1975) (Burger, J., concurring) ("The basis of Albemarle's liability was that its seniority system perpetuated the effects of past discrimination . . . .").

229. 42 U.S.C. § 2000e-2(h) (2012).

230. Trans World Airlines, Inc. v. Hardison, 432 U.S. 63, 82 (1977).

231. Calders & Žliobaitė, *supra* note 64, at 48.

232. For example, establishing whether and to what extent crime statistics misrepresent the relative proportion of offenses committed by different social groups is not an easy task. Especially challenging are those crimes that are more likely to go under- or unreported if not directly observed by

evident bias when the data was plotted on a map. Boston's Office of New Urban Mechanics was therefore able to partner with "a range of academics to take into account issues of equitable access and digital divides."[233] In many cases, however, an analyst can only determine the extent of—and correct for—unintentional discrimination that results from reporting, sampling, and selection biases if the analyst has access to information that somehow reveals misrepresentations of protected classes in the dataset. Often, there may be no practical alternative method for collecting information that would even reveal the existence of a bias.

Any attempt to correct for collection bias immediately confronts the problem of whether or not the employer recognizes the specific type of bias that is producing disparate results. Then, in order to correct for it, an employer must have access to the underlying data and often an ability to collect more. Where more data is clearly not accessible, data miners can proactively compensate for some of the bias by oversampling underrepresented communities.[234]

If the employer fails to be proactive or tries and fails to detect the bias that causes the disparate impact, liability is an open question. As discussed in Part II.B, liability partly depends on how liberally a court interprets the requirement that an employer "refuses" to use an alternative scheme.[235] Even a liberal interpretation, though, would require evidence of the particular type of discrimination at issue, coupled with evidence that such an alternative scheme exists. Thus, finding liability seems unlikely. Worse, where such showing is possible, there may be no easy or obvious way to remedy the situation.

To address collection bias directly, an employer or an auditor must have access to the underlying data and the ability to adjust the model. Congress could require this directly of any employer using data mining techniques. Some employers are investing in their own data now and could potentially meet such requirements.[236] But employers also seem happy to rely on models developed and administered by third parties, who may have a far greater set of examples and far richer data than any individual company.[237] Furthermore, due to economies of scale that are especially important in data analysis, one can imagine that third parties specializing in work-force science will be able to offer employers this service much less expensively than they could manage it

---

the police. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE (2007).

233. Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data [https://perma.cc/9A7V-3UVD]. Such techniques would also address the concerns raised in Lerman, *supra* note 50.

234. Faisal Kamiran & Toon Calders, *Data Preprocessing Techniques for Classification Without Discrimination*, 33 KNOWLEDGE & INFO. SYS. 1, 3 (2011).

235. *See supra* Part II.B.

236. Peck, *supra* note 181.

237. *See* Richtel, *supra* note 69.

themselves. If Congress attempted to demand that employers have access to the data, it would face strong resistance from the ever-growing data analysis industry, whose business depends on the proprietary nature of the amassed information. More likely, Congress could require audits by a third party like the EEOC or a private auditor, in order to protect trade secrets, but this still seems a tall task. Ultimately, because proactive oversampling and retroactive data correction are at least possible, collection bias has the most promising prospects for a workable remedy of any of the identified data mining mechanisms.

3. Feature Selection

Even in the absence of prejudice or bias, determining the proper degree of precision in the distinctions drawn through data mining can be extremely difficult. Under formal disparate treatment, this is straightforward: any decision that expressly classifies by membership in a protected class is one that draws distinctions on illegitimate grounds. It is far less clear, however, what constitutes legitimate statistical discrimination when individuation does not rely on proscribed criteria. In these cases, the perceived legitimacy seems to depend on a number of factors: (1) whether the errors seem avoidable because (2) gaining access to additional or more granular data would be trivial or (3) would not involve costs that (4) outweigh the benefits. This seems to suggest that the task of evaluating the legitimacy of feature selection can be reduced to a rather straightforward cost-benefit analysis. Companies would have an obligation to pursue ever more—and more granular—data until the costs of gathering that data exceed the benefits conferred by the marginal improvements in accuracy.

Unfortunately, as is often the case with cost-benefit analyses, this approach fails to consider how different actors will perceive the value of the supposed benefits as well as the costs associated with errors. The obvious version of this criticism is that "actuarially saddled" victims of inaccurate determinations may find cold comfort in the fact that certain decisions are rendered more reliably overall when decision makers employ data mining.[238] A more sophisticated version of this criticism focuses on the way such errors assign costs and benefits to different actors at systematically different rates. A model with any error rate that continues to turn a profit may be acceptable to decision makers at a company, no matter the costs or inconvenience to specific customers.[239] Even when companies are subject to market pressures that would

238. SCHAUER, *supra* note 67, at 5. As Schauer explains, perfectly particularized decisions are, of course, a logical impossibility. Accepting this inherent limitation introduces a different sort of procedural concern: occasional errors might be tolerable if they are easy to detect and rectify, which is why, among other things, the perceived legitimacy of decisions often also depends on due process. *See id.* at 172; *see also* Citron, *supra* note 11.

239. Gandy, *supra* note 31, at 36.

force them to compete by lowering these error rates, the companies may find that there is simply no reason to invest in efforts that do so if the errors happen to fall disproportionately on especially unprofitable groups of consumers. Furthermore, assessing data mining as a matter of balancing costs and benefits leaves no room to consider morally salient disparities in the degree to which the costs are borne by different social groups. This raises the prospect that there might be systematic differences in the rates at which members of protected classes are subject to erroneous determinations.[240] Condemning these groups to bear the disproportionate burden of erroneous determinations would strike many as highly objectionable, despite greater accuracy in decision making for the majority group.[241] Indeed, simply accepting these cost differences as a given would subject those already in less favorable circumstances to less accurate determinations.

Even if companies assume the responsibility for ensuring that members of protected classes do not fall victim to erroneous determinations at systematically higher rates, they could find that increasing the resolution and range of their analyses still fails to capture the causal relationships that account for different outcomes because those relationships are not easily represented in data.[242] In such cases, rather than reducing the error rate for those in protected classes, data miners could structure their analyses to minimize the difference in error rates between groups. This solution may involve some unattractive tradeoffs, however. In reducing the disparate impact of errors, it may increase the overall amount of errors. In other words, generating a model that is equally unfair to protected and unprotected classes might increase the overall amount of unfairness.

4. Proxies

Computer scientists have been unsure how to deal with redundant encodings in datasets. Simply withholding these variables from the data mining exercise often removes criteria that hold demonstrable and justifiable relevance to the decision at hand. As Calders and Žliobaitė note, "[I]t is problematic [to remove a correlated attribute] if the attribute to be removed also carries some objective information about the label [quality of interest]."[243] Part of the problem seems to be that there is no obvious way to determine *how* correlated a relevant attribute must be with class membership to be worrisome. Nor is there a self-evident way to determine when an attribute is sufficiently relevant to justify its consideration, despite its high correlation with class membership. As

240. Moritz Hardt, *How Big Data Is Unfair*, MEDIUM (Sept. 26, 2014), https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de [https://perma.cc/YN44-M4DQ].

241. *See, e.g.*, Gandy, *supra* note 31, at 39.

242. *See supra* note 64 and accompanying text.

243. Calders & Žliobaitė, *supra* note 64, at 54.

Professors Devin Pope and Justin Sydnor explain, "[V]ariables are likely neither solely predictive nor purely proxies for omitted characteristics."[244]

But there is a bigger problem here: attempting to ensure fairly rendered decisions by excising highly correlated criteria only makes sense if the disparate impact happens to be an *avoidable* artifact of a particular way of rendering decisions. And yet, even when denied access to these highly correlated criteria, data mining may suggest alternative methods for rendering decisions that still result in the same disparate impact. Focusing on isolated data points may be a mistake because class membership can be encoded in more than one specific and highly correlated criterion. Indeed, it is very likely that class membership is reflected across a number of interrelated data points.[245] But such outcomes might instead demonstrate something more unsettling: that *other* relevant criteria, whatever they are, happen to be possessed at different rates by members of protected classes. This explains why, for instance, champions of predictive policing have responded to critics by arguing that "[i]f you wanted to remove everything correlated with race, you couldn't use anything. That's the reality of life in America."[246] Making accurate determinations means considering factors that are somehow correlated with proscribed features.

Computer scientists have even shown that "[r]emoving all such correlated attributes before training does remove discrimination, but with a high cost in classifier accuracy."[247] This reveals a rather uncomfortable truth: the current distribution of relevant attributes—attributes that can and should be taken into consideration in apportioning opportunities fairly—is demonstrably correlated with sensitive attributes because the sensitive attributes have meaningfully conditioned what relevant attributes individuals happen to possess.[248] As such, attempts to ensure procedural fairness by excluding certain criteria from consideration may conflict with the imperative to ensure accurate determinations. The only way to ensure that decisions do not systematically disadvantage members of protected classes is to reduce the overall accuracy of

244. Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 206 (2011).

245. *Supra* discussion accompanying note 101.

246. Labi, *supra* note 5 (quoting Ellen Kurtz, Director of Research for Philadelphia's Adult Probation and Parole Department).

247. Toon Calders & Sicco Verwer, Presentation at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases: Three Naïve Bayes Approaches for Discrimination-Free Classification 9 (2010), http://wwwis.win.tue.nl/~tcalders/dadm/lib/exe/fetch.php?media=ecmlpkdd_2010_discrimination.pdf [http://perma.cc/9V72-2NVM].

248. In a sense, computer scientists have unwittingly furnished the kind of evidence that social scientists routinely seek: the particular contours of inequality. *See, e.g.*, SOCIAL INEQUALITY (Kathryn M. Neckerman ed., 2004).

all determinations. As Dwork et al. remark, these results "demonstrate a quanti[t]ative trade-off between fairness and utility."[249]

In certain contexts, data miners will never be able to fully disentangle legitimate and proscribed criteria. For example, the workforce optimization consultancy, Evolv, discovered that "[d]istance between home and work . . . is strongly associated with employee engagement and retention."[250] Despite the strength of this finding, Evolv "never factor[s] [it] into the score given each applicant . . . because different neighborhoods and towns can have different racial profiles, which means that scoring distance from work could violate equal-employment-opportunity standards."[251] Scholars have taken these cases as a sign that the "major challenge is how to find out which part of information carried by a sensitive (or correlated) attribute is sensitive and which is objective."[252] While researchers are well aware that this may not be easy to resolve, let alone formalize into a computable problem, there is a bigger challenge from a legal perspective: any such undertaking would necessarily wade into the highly charged debate over the degree to which the relatively less favorable position of protected classes warrants the protection of antidiscrimination law in the first instance.

The problems that render data mining discriminatory are very rarely amenable to obvious, complete, or welcome resolution. When it comes to setting a target variable and feature selection, policy cannot lay out a clear path to improvement; reducing the disparate impact will necessitate open-ended exploration without any way of knowing when analysts have exhausted the possibility for improvement. Likewise, policies that compel institutions to correct tainted datasets or biased samples will make impossible demands of analysts. In most cases, they will not be able to determine what the objective determination should have been or independently observe the makeup of the entire population. Dealing with both of these problems will ultimately fall to analysts' considered judgment. Solutions that reduce the accuracy of decisions to minimize the disparate impact caused by coarse features and unintentional proxies will force analysts to make difficult and legally contestable trade-offs. General policies will struggle to offer the specific guidance necessary to determine the appropriate application of these imperfect solutions. And even when companies voluntarily adopt such strategies, these internal difficulties will likely allow a disparate impact to persist.

---

249. Dwork et al., *supra* note 81, at 215; *cf.* Wax, *supra* note 146, at 711 (noting intractable problems due to a "validity-diversity tradeoff" in employment metrics).

250. Peck, *supra* note 181.

251. *Id.* Other companies have not held back from considering this information for the very same purposes. *See* Joseph Walker, *Meet the New Boss: Big Data*, WALL ST. J. (Sept. 20, 2012), http://www.wsj.com/news/articles/SB10000872396390443890304578006252019616768 [https://perma.cc/6DHY-M429].

252. Calders & Žliobaitė, *supra* note 64, at 56.

B. External Difficulties

Assuming the internal difficulties can be resolved, there are further political and constitutional restraints on addressing Title VII's inadequacies with respect to data mining. Data mining discrimination will force a confrontation between the two divergent principles underlying antidiscrimination law: anticlassification and antisubordination.[253] Which of these two principles motivates discrimination law is a contentious debate, and making remedies available under antidiscrimination law will require a commitment to antisubordination principles that have thus far not been forthcoming from legislatures. This is not merely a political concern, as substantive remediation is becoming ever more suspect constitutionally as well.[254] While such remedies may be politically and legally impossible, the nature of data mining itself makes them practically necessary. Accordingly, these external difficulties may prevent antidiscrimination law from fully addressing data mining discrimination.

Two competing principles have always undergirded antidiscrimination law: anticlassification and antisubordination. Anticlassification is the narrower of the two, holding that the responsibility of the law is to eliminate the unfairness individuals in certain protected classes experience due to decision makers' choices.[255] Antisubordination theory, in contrast, holds that the goal of antidiscrimination law is, or at least should be, to eliminate status-based inequality due to membership in those classes, not as a matter of procedure, but of substance.[256]

Different mitigation policies effectuate different rationales. Disparate treatment doctrine arose first, clearly aligning with the anticlassification principle by proscribing intentional discrimination, in the form of either explicit singling out of protected classes for harm or masked intentional discrimination. Since disparate impact developed, however, there has never been clarity as to which of the principles it is designed to effectuate.[257] On the one hand, disparate impact doctrine serves anticlassification by being an "evidentiary dragnet" used to "smoke out" well-hidden disparate treatment.[258] On the other hand, as an effects-based doctrine, there is good reason to believe it was intended to address substantive inequality.[259] In this sense, the "business

253. Helen Norton, *The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality*, 52 WM. & MARY L. REV. 197, 206–15 (2010); *see also* Bagenstos, *supra* note 122, at 40–42, 40–41 nn.214–15 (collecting sources); Owen M. Fiss, *Groups and the Equal Protection Clause*, 5 PHIL. & PUB. AFF. 107, 157 (1976).

254. *See* Norton, *supra* note 253.

255. *Id.* at 209.

256. *Id.* at 206.

257. Primus, *supra* note 98, at 520–23.

258. *Id.*; Perry, *supra* note 138, at 526.

259. *See* Griggs v. Duke Power Co., 401 U.S. 424, 429–30 (1971) ("The objective of Congress in the enactment of Title VII is plain from the language of the statute. It was to achieve equality of

necessity" defense is a necessary backstop that prevents members of traditionally disadvantaged groups from simply forcing their way in without the necessary skills or abilities.[260]

Thus, the mapping from anticlassification and antisubordination to disparate treatment and disparate impact was never clean. Early critics of civil rights laws actually complained that proscribing consideration of protected class was a subsidy to black people.[261] This argument quickly gave way in the face of the rising importance of the anticlassification norm.[262] Over the years, the anticlassification principle has come to dominate the landscape so thoroughly that a portion of the populace thinks (as do a few Justices on the Supreme Court) that it is the only valid rationale for antidiscrimination law.[263]

The move away from antisubordination began only five years after disparate impact was established in *Griggs*. In *Washington v. Davis*, the Court held that disparate impact could not apply to constitutional claims because equal protection only prohibited intentional discrimination.[264] Since then, the various affirmative action cases have overwritten the distinction between benign and harmful categorizations of race in favor of a formalistic anticlassification principle, removed from its origins as a tool to help members of historically disadvantaged groups.[265] White men can now bring disparate treatment claims.[266] If antidiscrimination law is no longer thought to serve the purpose of improving the relative conditions of traditionally disadvantaged groups, antisubordination is not part of the equation.

While the Court has clearly established that antisubordination is not part of constitutional equal protection doctrine, that it does not mean that antisubordination cannot animate statutory antidiscrimination law. Antisubordination and anticlassification came into sharp conflict in *Ricci v. Destefano*, a 2009 case in which the City of New Haven refused to certify a promotion exam given to its firefighters on the grounds that it would have produced a disparate impact based on its results.[267] The Supreme Court held that the refusal to certify the test, a facially race-neutral attempt to correct for perceived disparate impact, was in fact a race-conscious remedy that constituted disparate treatment of the majority-white firefighters who would

---

employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. Under the Act, practices, procedures, or tests neutral on their face, and even neutral in terms of intent, cannot be maintained if they operate to 'freeze' the status quo of prior discriminatory employment practices.").

260. *See* Tex. Dep't of Hous. & Cmty Aff. v. Inclusive Cmtys. Project, Inc., No. 13-1371, slip op. at 8 (Sup. Ct. 2015) (quoting *Griggs*, 401 U.S. at 431).

261. Primus, *supra* note 98, at 525–26.

262. *Id.*

263. *See* Bagenstos, *supra* note 122, at 41.

264. 426 U.S. 229, 246–48 (1976).

265. Rubenfeld, *supra* note 99, at 428, 433–36.

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

267. *Id.*

have been promoted based on the exam's results.[268] The Court held that disparate treatment cannot be a remedy for disparate impact without a "strong basis in evidence" that the results would lead to actual disparate treatment liability.[269]

Ricci was the first indication at the Supreme Court that disparate impact doctrine could be in conflict with disparate treatment.[270] The Court had previously ruled in essence that the antisubordination principle could not motivate a constitutional decision,[271] but it had not suggested that law effectuating that principle could itself be discriminatory against the dominant groups. That has now changed.[272]

The decision has two main consequences for data mining. First, where the internal difficulties in resolving discrimination in data mining described above can be overcome, legislation that requires or enables such resolution may run afoul of *Ricci*. Suppose, for example, Congress amended Title VII to require that employers make their training data and models auditable. In order to correct for detected biases in the training data that result in a model with a disparate impact, the employer would first have to consider membership in the protected class. The remedy is inherently race-conscious. The *Ricci* Court did hold that an employer may tweak a test during the "test-design stage," however.[273] So, as a matter of timing, data mining might not formally run into

268. *Id.*

269. *Id.* at 563.

270. Primus, *supra* note 92, at 1344; Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2162–63 (2013); Norton, *supra* note 253, at 229.

271. *See* Washington v. Davis, 426 U.S. 229, 239 (holding that discriminatory purpose is necessary to finding a violation of equal protection).

272. Primus, *supra* note 92, at 1343. While the decision was formally about Title VII only, and thus amenable to statutory resolution, the reasoning applied equally well to a future equal protection claim, endangering the future of disparate impact. *Id.* at 1385–87; Bradley A. Areheart, *The Anticlassification Turn in Employment Discrimination Law*, 63 ALA. L. REV. 955, 994 (2012); Norton, *supra* note 253, at 229–30. Justice Scalia stated as much in his concurrence. *Ricci*, 557 U.S. at 594 (Scalia, J., concurring) ("[The Court's] resolution of this dispute merely postpones the evil day on which the Court will have to confront the question: Whether, or to what extent, are the disparate-impact provisions of Title VII of the Civil Rights Act of 1964 consistent with the Constitution's guarantee of equal protection?"). But the Supreme Court seemed to pull back from the brink last term, approving of the use of disparate impact in a new setting—the Fair Housing Act—and engaging deeply with the constitutional issues that *Ricci* raised, settling them for now. Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After Inclusive Communities*, 101 CORNELL L. REV., at *11–12 (forthcoming 2016), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2642631 [https://perma.cc/WD43-XW2G]; Richard Primus, *Of Visible Race-Consciousness and Institutional Role: Equal Protection and Disparate Impact After* Ricci *and* Inclusive Communities, *in* TITLE VII OF THE CIVIL RIGHTS ACT AFTER 50 YEARS: PROCEEDINGS OF THE NEW YORK UNIVERSITY 67TH ANNUAL CONFERENCE ON LABOR 295 (2015).

273. *Ricci*, 557 U.S. at 585 (majority opinion) ("Title VII does not prohibit an employer from considering, before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race. And when, during the test-design stage, an employer invites comments to ensure the test is fair, that process can provide a common ground for open discussions toward that end.").

Ricci if the bias resulting in a disparate impact is corrected before applied to individual candidates. After an employer begins to use the model to make hiring decisions, only a "strong basis in evidence" that the employer will be successfully sued for disparate impact will permit corrective action.[274] Of course, unless every single model used by an employer is subject to a prescreening audit (an idea that seems so resource intensive that it is effectively impossible), the disparate impact will be discovered only when the employer faces complaints. Additionally, while *Ricci*'s holding was limited in scope, the "strong basis in evidence" standard did not seem to be dictated by the logic of the opinion, which illustrated a more general conflict between disparate treatment and disparate impact.[275]

Second, where the internal difficulties *cannot* be overcome, there is likely no way to correct for the discriminatory outcomes aside from results-focused balancing, and requiring this will pose constitutional problems. For those who adhere to the anticlassification principle alone, such an impasse may be perfectly acceptable. They might say that as long as employers are not intentionally discriminating based on explicitly proscribed criteria, the chips should fall where they may. To those who believe some measure of substantive equality is important over and above procedural equality, this result will be deeply unsatisfying.

An answer to the impasse created by situations that would require results-focused rebalancing is to reexamine the purpose of antidiscrimination law. The major justification for reliance on formal disparate treatment is that prejudice is simply irrational and thus unfair. But if an employer knows that his model has a disparate impact, but it is also his most predictive, the argument that the discrimination is irrational loses any force. Thus, data mining may require us to reevaluate why and whether we care about not discriminating.

Consider another example involving tenure predictions, one in which an employer ranks potential employees with the goal of hiring only those applicants that the company expects to retain for longer periods of time. In optimizing its selection of applicants in this manner, the employer may unknowingly discriminate against women if the historical data demonstrates that they leave their positions after fewer years than their male counterparts. If gender accounts for a sufficiently significant difference in employee tenure, data mining will generate a model that simply discriminates on the basis of gender or those criteria that happen to be proxies for gender. Although selecting applicants with an eye to retention might seem both rational and reasonable, granting significance to predicted tenure would subject women to systematic disadvantage if gender accounts for a good deal of the difference in tenure. If that is the case, any data mining exercise that attempts to predict

274. *Id.* at 585.
275. *See generally id.*

tenure will invariably rediscover this relationship. One solution could be for Congress to amend Title VII to reinvigorate strict business necessity.[276] This would allow a court to accept that relying on tenure is rational but not strictly "necessary" and that perhaps other factors could make up for the lack of predicted tenure.

But this solution and all others must rely on the antisubordination principle. Consider this question: should the law permit a company to hire no women at all—or none that it correctly predicts will depart following the birth of a child—because it is the most rational choice according to their model?[277] The answer seems obviously to be no. But why not? What forms the basis for law's objection to rational decisions, based on seemingly legitimate criteria, that place members of protected classes at systematic disadvantage? The Supreme Court has observed that, "Title VII requires employers to treat their employees as *individuals*, not 'as simply components of a racial, religious, sexual, or national class.'"[278] On the strength of that statement, the Court held that employers could not force women to pay more into an annuity because they, as women, were likely to live longer.[279] But it is not clear that this reasoning translates directly to data mining. Here, the model takes a great deal of data about an individual, and while it does make a determination based on statistics, it will make a different one if analyzing two different women. So if the model said to hire *no* women, it would be illegal, but, according to the doctrine, perhaps only because every woman ends up with the same result.

The only escape from this situation may be one in which the relevance of gender in the model is purposefully ignored and all factors correlated with gender are suppressed. The outcome would be a necessarily less accurate model. The justification for placing restrictions on employers, and limiting the effectiveness of their data mining, would have to depend on an entirely different set of arguments than those advanced to explain the wrongfulness of biased data collection, poorly labeled examples, or an impoverished set of features. Here, shielding members of protected classes from less favorable treatment is not justified by combatting prejudice or stereotyping. In other words, any prohibition in this case could not rest on a procedural commitment to ensuring ever more accurate determinations. Instead, the prohibition would have to rest on a substantive commitment to equal representation of women in the workplace. That is, it would have to rest on a principle of antisubordination.

---

276. Remember that if there is disparate impact, but no liability, it is because the goal was deemed job-related or satisfied business necessity.

277. As a matter of case law, this question has essentially been answered. The Supreme Court has ruled that in the case of women being required to pay more into an annuity because they would likely live longer, pure market rationality is not a good enough answer. Ariz. Governing Comm. v. Norris, 463 U.S. 1073, 1083 (1983) (quoting City of Los Angeles Dep't of Water & Power v. Manhart, 435 U.S. 702, 708 (1978)).

278. *Id.*

279. *Id.*

The dilemma is clear: the farther the doctrine gets from substantive remediation, the less utility it has in remedying these kinds of discriminatory effects.[280] But the more disparate impact is thought to embody the antisubordination principle—as opposed to the "evidentiary dragnet" in service of the anticlassification norm—the more it will invite future constitutional challenges.[281]

This also raises a point about disparate *treatment* and data mining. Within data mining, the effectiveness of prohibiting the use of certain information exists on a spectrum. On one end, the prohibition has little to no effect because either the information is redundantly encoded or the results do not vary along lines of protected class. On the other end, the prohibition reduces the accuracy of the models. That is, if protected class data were not prohibited, that information would alter the results, presumably by making members of protected classes worse (or, in some cases, better) off. Thus, as a natural consequence of data mining, a command to ignore certain data has either no effect[282] or the effect of altering the fortunes of those protected classes in substantive ways. Therefore, with respect to data mining, due to the zero-sum nature of a ranking system, even *disparate treatment* doctrine is a reallocative remedy similar to affirmative action.[283] Once again, this erodes the legitimate rationale for on the one hand supporting an anticlassification principle but on the other, holding fast against antisubordination in this context. The two principles tend to accomplish the same thing, but one is less effective at achieving substantive equality.

This reveals that the pressing challenge does not lie with ensuring procedural fairness through a more thorough stamping out of prejudice and bias but rather with developing ways of reasoning to adjudicate when and what amount of disparate impact is tolerable. Abandoning a belief in the efficacy of procedural solutions leaves policy makers in an awkward position because there is no definite or consensus answer to questions about the fairness of specific outcomes. These need to be worked out on the basis of different normative principles. At some point, society will be forced to acknowledge that this is really a discussion about what constitutes a tolerable level of disparate impact in employment. Under the current constitutional order and in the political climate, it is tough to even imagine having such a conversation. But, until that happens, data mining will be permitted to exacerbate existing inequalities in difficult-to-counter ways.

---

280. *Id.* at 537.

281. Primus, *supra* note 98, at 536–37.

282. *See supra* text accompanying note 101.

283. For an argument that this is true more generally, see Bagenstos, *supra* note 90, and Owen M. Fiss, *A Theory of Fair Employment Laws*, 38 U. CHI. L. REV. 235, 313 (1971) (arguing that a key to understanding antidiscrimination prohibitions in the employment realm is that the prohibitions "confer[] benefits on a racial class—blacks").

CONCLUSION

This Essay has identified two types of discriminatory outcomes from data mining: a family of outcomes where data mining goes "wrong" and outcomes where it goes too "right." Data mining can go wrong in any number of ways. It can choose a target variable that correlates to protected class more than others would, reproduce the prejudice exhibited in the training examples, draw adverse lessons about protected classes from an unrepresentative sample, choose too small a feature set, or not dive deep enough into each feature. Each of these potential errors is marked by two facts: the errors may generate a manifest disparate impact, and they may be the result of entirely innocent choices made by data miners.

Where data mining goes "right," data miners could not have been any more accurate given the starting point of the process. This very accuracy, exposing an uneven distribution of attributes that predict the target variable, gives such a result its disparate impact. If the data accurately models inequality, attempts to devise an alternative way of making the same prediction will only narrow the disparate impact if these efforts reduce the accuracy of the decision procedure. By now, it should be clear that Title VII, and very likely other similarly process-oriented civil rights laws, cannot effectively address this situation.

This means something different for the two families, and it should be slightly more surprising for the former. At a high level of abstraction, where a decision process goes "wrong" and this wrongness creates a disparate impact, Title VII and similar civil rights laws should be up to the task of solving the problem; that is ostensibly their entire purpose. But aside from a few more obvious cases involving manifest biases in the dataset, it is quite difficult to determine ahead of time what "correct" data mining looks like. A decision maker can rarely discover that the choice of a particular target variable is more discriminatory than other choices until after the fact, at which point it may be difficult and costly to change course. While data miners might have some intuitions about the influence that prejudice or bias played in the prior decisions that will serve as training data, data miners may not have any systematic way of measuring and correcting for that influence. And even though ensuring reliable samples before training a model is a possibility, the data may never be perfect. It may be impossible to determine, ex ante, how much the bias contributes to the disparate impact, it may not be obvious how to collect additional data that makes the sample more representative, and it may be prohibitively expensive to do so. Companies will rarely be able to resolve these problems completely; their models will almost always suffer from some deficiency that results in a disparate impact. A standard that holds companies liable for any amount of theoretically avoidable disparate impact is likely to ensnare all companies. Thus, even at this level of abstraction, it becomes clear that holding the decision makers responsible for these disparate impacts is at

least partly troubling from a due process perspective. Such concerns may counsel against using data mining altogether. This would be a perverse outcome, given how much even imperfect data mining can do to help reduce the very high rates of discrimination in employment decisions.

If liability for getting things "wrong" is difficult to imagine, how does liability for getting things "right" make any more sense? That proxy discrimination largely rediscovers preexisting inequalities suggests that perhaps Title VII is not the appropriate remedial vehicle. If what is at stake are the results of decades of historical discrimination and wealth concentration that have created profound inequality in society, is that not too big a problem to remedy through individual lawsuits, assuming affirmative action and similar policies are off the table? Thus, perfect data mining forces the question: if employers can say with certainty that, given the status quo,[284] candidates from protected classes are on average less ready for certain jobs than more privileged candidates, should employers specifically be penalized for hiring fewer candidates from protected classes?

Doctrinally, the answer is yes, to some extent. Professor Christine Jolls has written that disparate impact doctrine is akin to accommodation in disability law—that is, both accommodations and disparate impact specifically require employers to depart from pure market rationality and incur costs associated with employing members of protected classes.[285] Similarly, the Title VII annuity cases[286] and Title VII's ban on following racist third-party preferences[287] each require a departure from market rationality. Thus, Title VII makes that decision to a degree. But to what degree? How much cost must an employer bear?

Title VII does not require an employer to use the least discriminatory means of running a business.[288] Likewise, Title VII does not aim to remedy historical discrimination and current inequality by imposing all the costs of restitution and redistribution on individual employers.[289] It is more appropriately understood as a standard of *defensible* disparate impact. One route, then, to addressing the problems is to make the inquiry more searching and put the burden on the employer to avoid at least the easy cases. In a system that is as unpredictable as data mining can be, perhaps the proper way of

284. We cannot stress enough the import of these caveats. Certainty is a strong and unlikely precondition, and the status quo should not be taken as a given, as we explain below.

285. *See generally* Jolls, *supra* note 90.

286. *See* Ariz. Governing Comm. v. Norris, 463 U.S. 1073, 1083 (1983); City of Los Angeles Dep't of Water & Power v. Manhart, 435 U.S. 702, 708 (1978).

287. *See* 29 C.F.R. § 1604.2(a)(1)(iii) (2015) (stating the EEOC's position that "the preferences of coworkers, the employer, clients or customers" cannot be used to justify disparate treatment).

288. *See, e.g.*, El v. Se. Pa. Transp. Auth., 479 F.3d 232, 242 (3d Cir. 2007).

289. *See* Steven L. Willborn, *The Disparate Impact Model of Discrimination: Theory and Limits*, 34 AM. U. L. REV. 799, 809–10 (1985).

thinking about the solution is a duty of care, a theory of negligent discrimination.[290]

But if Title VII alone cannot solve these problems, where should society look for answers? Well, the first answer is to question the status quo. Data mining takes the existing state of the world as a given and ranks candidates according to their predicted attributes in *that* world. Data mining, by its very nature, treats the target variable as the only item that employers are in a position to alter; everything else that happens to correlate with different values for the target variable is assumed stable. But there are many reasons to question these background conditions. Sorting and selecting individuals according to their apparent qualities hides the fact that the predicted effect of possessing these qualities with respect to a specific outcome is also a function of the conditions under which these decisions are made. Recall the tenure example from Part III.B. In approaching appropriate hiring practices as a matter of selecting the "right" candidates at the outset, an employer will fail to recognize potential changes that he could make to workplace conditions. A more family-friendly workplace, greater on-the-job training, or a workplace culture more welcoming to historically underrepresented groups could affect the course of employees' tenure and their long-term success in ways that undermine the seemingly prophetic nature of data mining's predictions.

These are all traditional goals for reducing discrimination within the workplace, and they continue to matter even in the face of the eventual widespread adoption of data mining. But data can play a role here, too. For example, comparing the performance of equally qualified candidates across different workplaces can help isolate the formal policies and institutional dynamics that are more or less likely to help workers flourish. Research of this sort could also reveal areas for potential reform.[291]

Education is also important. Employers may take some steps to rectify the problem on their own if they better understand the cause of the disparity. Right now, many of the problems described in Part I are relatively unknown. But the more employers and data miners understand these pitfalls, the more they can strive to create better models on their own. Many employers switch to data-driven practices for the express purpose of eradicating bias;[292] if employers discover that they are introducing new forms of bias, they can correct course.

Even employers seeking only to increase efficiency or profit may find that their incentives align with the goals of nondiscrimination. Faulty data and data

290. *See generally* David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899 (1993).

291. Solon Barocas, *Putting Data to Work*, DATA AND DISCRIMINATION: COLLECTED ESSAYS 58, 60 (Seeta Peña Gangadharan, Virginia Eubanks & Solon Barocas eds., 2014).

292. Claire Cain Miller, *Can an Algorithm Hire Better than a Human?*, N.Y. TIMES (June 25, 2015), http://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html [https://perma.cc/UR37-83D4].

mining will lead employers to overlook or otherwise discount people who are actually "good" employees. Where the cost of addressing these problems is at least compensated for by a business benefit of equal or greater value, employers may have natural incentives to do so.

Finally, employers could also make more effective use of the tools that computer scientists have begun to develop.[293] Advances in these areas will depend, crucially, on greater and more effective collaboration between employers, computer scientists, lawyers, advocates, regulators, and policy makers.[294]

This Essay is a call for caution in the use of data mining, not its abandonment. While far from a panacea, data mining can and should be part of a panoply of strategies for combatting discrimination in the workplace and for promoting fair treatment and equality. Ideally, institutions can find ways to use data mining to generate new knowledge and improve decision making that serves the interests of both decision makers and protected classes. But where data mining is adopted and applied without care, it poses serious risks of reproducing many of the same troubling dynamics that have allowed discrimination to persist in society, even in the absence of conscious prejudice.

293. *See* list *supra* note 217.

294. Joshua A. Kroll, et al., Accountable Algorithms, 165 U. PA. L. REV. __ (forthcoming 2017).