ISSN 1936-5349 (print)
ISSN 1936-5357 (online)

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS
FELLOWS' DISCUSSION PAPER SERIES

BIG DATA AND DISCRIMINATION

Talia B. Gillis
Jann L. Spiess

Discussion Paper No. 84

09/2018

Harvard Law School
Cambridge, MA 02138

July 19, 2018
Talia B. Gillis[§†] and Jann L. Spiess[†]

Big Data and Discrimination

I. Introduction

For many financial products, such as loans and insurance policies, companies distinguish between people based on their different risks and returns. However, the ability to distinguish between people by trying to predict future behavior or profitability of a contract is often restrained by legal rules that aim to prevent certain types of discrimination. For example, the Equal Credit Opportunity Act (ECOA) forbids race, religion, age and other factors from being considered in setting credit terms, and the Fair Housing Act (FHA) prohibits discrimination in financing of real estate based on race, color, national origin, religion, sex, familial status, or disability.[1] Many of these rules were developed to challenge human discretion in setting prices and provide little guidance in a world where firms set credit terms based on sophisticated statistical methods and a large number of factors. This rise of artificial intelligence and "big data" raises the question where and how existing law can be applied to this novel setting, and where it must be adapted to remain effective.

In this article, we bridge the gap between old law and new methods by proposing a framework that brings together existing legal requirements with the structure of algorithmic decision-making in order to identify tensions and lay the ground for legal solutions. Focusing on the example of credit pricing, we confront steps in the genesis of an automated pricing rule with their regulatory opportunities and challenges.

Based on our framework, we argue that legal doctrine is ill-prepared to face the challenges posed by algorithmic decision-making in a big-data world. While automated pricing rules promise increased transparency, this opportunity is often confounded. Unlike human decision-making, the exclusion of data from consideration can be guaranteed in the algorithmic context. However, forbidding inputs alone does not assure equal pricing and can even increase pricing disparities between protected groups. Moreover, the complexity of machine learning pricing limits the

[§] Harvard Business School and Harvard Law School.
[†] Microsoft Research New England.
For helpful feedback we thank Oren Bar-Gill, John Beshears, Ellora Derenoncourt, Noah Feldman, Howell Jackson, Cass Sunstein, and the participants of the symposium on Personalized Law at Chicago Law School. Talia Gillis acknowledges support provided by the John M. Olin Center for Law, Economics, and Business at Harvard Law School.
[1] These laws do not exhaust the legal framework governing discrimination in credit pricing. Beyond other federal laws that also relate to credit discrimination, such as the Community Reinvestment Act, there are many state and local laws with discrimination provisions, such as fair housing laws.

ability to scrutinize the process that led to a pricing rule, frustrating legal efforts to examine the "conduct" that led to disparity. On the other hand, the reproducibility of automated prices creates new possibilities for more meaningful analysis of pricing outcomes. Building on this opportunity, we provide a framework for regulators to test decision rules ex ante in a way that provides meaningful comparisons between lenders.

To consider the challenges to applying discrimination law to a context in which credit pricing decisions are fully automated, we consider both the legal doctrine of "disparate treatment", dealing with cases in which a forbidden characteristic is considered directly in a pricing decision, and "disparate impact", when a facially neutral conduct has a discriminatory effect.[2] While in general the availability of a disparate-impact claim depends on the legal basis of the discrimination claim , in the context of credit pricing, the law permits the use of disparate impact as a basis of a discrimination claim both under the FHA and the ECOA.[3] A comprehensive discussion of these two doctrines and their application to credit pricing is beyond the scope of this paper, particularly because there are several aspects of these doctrines on which there is widespread disagreement.[4] We therefore abstract away from some of the details of the doctrines and focus on the building blocks that create a discrimination claim. Developing doctrine that is appropriate for this context ultimately requires a return to the fundamental justifications and motivations behind discrimination law.

Specifically, we consider three approaches to discrimination.[5] The first approach is to focus on the "inputs" of the decision, stemming from the view that discrimination law is primarily concerned with formal or intentional discrimination.[6] The second approach scrutinizes the decision-making process, policy or conduct that then led to disparity. The third approach focuses

[2] For an overview of the legal doctrine and their relation to theories of discrimination, see: JOHN J DONOHUE, ANTIDISCRIMINATION LAW § 2 (2007).

[3] The Supreme Court recently affirmed that disparate impact claims could be made under the FHA in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc*, 135 S. Ct. 2507 (2015), confirming the position of eleven appellate courts and various federal agencies including the Department of Housing and Urban Development (HUD) primarily responsible for enforcing the FHA. See Robert G Schwemm, *Fair Housing Litigation After Inclusive Communities: What's New and What's Not*, 115 COLUM. L. REV. SIDEBAR 106 (2015). Although there is not an equivalent Supreme Court case with respect to ECOA, the Consumer Financial Protection Bureau and courts have found that the statute allows for a claim of disparate impact.

[4] For further discussion of the discrimination doctrines under ECOA and FHA, see: Michael Aleo & Pablo Svirsky, *Foreclosure fallout: The banking industry's attack on disparate impact race discrimination claims under the fair housing act and the equal credit opportunity act*, 18 BU PUB. INT. LJ (2008); Alex Gano, *Disparate Impact and Mortgage Lending: A Beginner's Guide*, 88 U. COLO. L. REV. (2017).

[5] We find it necessary to divide approaches to discrimination by their goal and focus since the doctrines of disparate treatment and disparate impact can be consistent with more than one approach depending on the exact interpretation and implementation of the doctrine. Moreover, legal doctrines often require more than one approach to demonstrate a case disparate impact or disparate treatment, such as in the three-part burden-shifting framework for establishing a FHA disparate impact case as formulated by the HUD, *Implementation of the Fair Housing Act's Discriminatory Effects Standard*, 24 CFR Part 100 (2013).

[6] This basic articulation is also used by Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341 (2009). For a discussion on the different notions of intention see: Aziz Z Huq, *Judging Discriminatory Intent*, (2017) (unpublished) (arguing that judicial theory of "intention" is inconsistent).

on the disparity of the "outcome".[7] We consider the options facing a social planner to achieve different policy ends and discuss how algorithmic decision-making challenges each of these options, without adopting a particular notion of discrimination.

Existing legal doctrine provides little guidance on algorithmic decision-making because the typical discrimination case focuses on the human component of the decision, which often remains opaque. Consider a series of cases from around 2008, which challenged mortgage pricing practices. These cases argued that Black and Hispanic lenders ended up paying higher interest rates and fees after controlling for the "par rate" set by the mortgage originator. The claim was the discretion given to the mortgage originator's employees and brokers in setting the final terms of the loans above the "par rate", and the incentives to do so, caused the discriminatory pricing.[8] These types of assertions are made in several cases of individual claims,[9] class actions,[10] and regulatory action.[11] What is most striking is that these cases do not directly scrutinize the broker decisions, treating them as a "black box", and focus instead on the mortgage originator's discretion policy.[12] Had the court been able to analyze the discriminatory decisions directly, we would have had a greater understanding of the precise conduct that is problematic and as a result of the scope and range of the legal doctrine which are important for the automated pricing context, but discrimination cases that involve opaque human decisions do not allow us to develop the exact perimeters of the doctrine.[13]

When algorithms make decisions, opaque human behavior is replaced by a set of rules constructed from data. Specifically, we consider prices that are set based on prediction of mortgage default. An algorithm takes as an input a training data set with past defaults and outputs a function that relates consumer characteristics, such as their income and credit score,

[7] We do not argue directly for any of these three approaches, rather we point to the opportunities and challenges that machine learning credit pricing creates for each approach.

[8] Most of these cases are disparate impact cases, although some of them are more ambiguous as to the exact grounds of the discrimination case and may be read as a disparate treatment case.

[9] See for example: Martinez v. Freedom Mortgage Team, Inc., 527 F. Supp. 2d 827 (N.D. 111. 2007).

[10] See for example: Ramirez v. Greenpoint Mortgage Funding, Inc. 633 F. Supp. 2d 922 (Dist. Court, ND California 2008); Miller v. Countrywide Bank, N.A., 571 F. Supp. 2d 251 (D. Mass. 2008).

[11] For a discussion of a series of complaints by the Justice Department against mortgage brokers that were settled see: IAN AYRES, GARY KLEIN & JEFFREY WEST, THE RISE AND (POTENTIAL) FALL OF DISPARATE IMPACT LENDING LITIGATION (Lee Anne Fennell and Benjamin J. Keys ed. 2017) (discussing a case against Countrywide (2011), Wells Fargo (2012) and Sage Bank (2015), all involving a claim that discretion to brokers resulted in discrimination).

[12] The use of a discretion policy as the conduct that caused the discriminatory effect has been applied by the CFPB to ECOA cases (see for example: The CFPB's Consent Order in the matter of American Honda Finance Corporation from 2015) as well as other areas, such as employment discrimination cases. The seminal case Watson v. Fort Worth Bank & Trust dealt with a disparate impact claim arising from discretionary and subjective promotion policies. The future of these types of class action cases is uncertain given Wal-Mart Stores, Inc. v. Dukes, 564 US 228 (2011). However, see the HUD's framing of the decision in the 2013 Regulation, page 11468 [Federal Register]

[13] It is important to note that this opaqueness is not only evidentiary, meaning the difficulty in proving someone's motivation and intentions in court. It is also a result of human decision-making often being opaque to the decision-maker themselves. There are decades of research showing that people have difficulty in recovering the basis for their decisions, particularly when they involve race. See for example the research reviews of the Kirwan Institute for the Study of Race and Ethnicity at the Ohio State University.

to the probability of default. Advances in statistics and computer science have produced powerful algorithms that excel at this prediction task especially when individual characteristics are rich and data sets large. These machine-learning algorithms search through large classes of complex rules to find a rule that works well at predicting the default of new consumers using past data. Since we consider the translation of the default prediction into a price as a simple transformation of the algorithm's prediction, we refer to the prediction and its translation into a pricing rule jointly as the "decision rule".[14]

We connect machine learning decision rules and current law by considering the three stages of a pricing decision, which we demonstrate in a simulation exercise. The data we use is based on real data on mortgage applicants using the Boston HMDA dataset[15] and we impute default probabilities from a combination of loan approvals and calibrate them to overall default rates.[16] The simulated data allows us to demonstrate several of our conceptual arguments and the methodological issues we discuss. However, given that crucial parts of the data are simulated, the graphs and figures in this paper should not be interpreted as reflecting real world observations, but rather methodological challenges and opportunities that arise in the context of algorithmic decision making.[17]

The remainder of this paper discuss each of the three steps of a pricing decision by underlining both the challenges and the opportunities presented by machine learning pricing to current legal rules. First, we consider the data input stage of the pricing decision and argue that exclusion of forbidden characteristics has limited effect, and would only satisfy a narrow understanding of anti-discrimination law. One fundamental aspect of anti-discrimination laws is the prohibition on the conditioning of the decision on the protected characteristics, which can formally be achieved in automated decision-making. However, the exclusion of the forbidden input alone may be insufficient when there are other characteristics that are correlated with the forbidden input, an issue that is exacerbated in the context of big data. On the other hand, we highlight the ways in which restricting a broader range of data inputs may have unintended consequences, such as increasing price disparity.

Second, we connect the process of constructing a pricing rule to the legal analysis of conduct and highlight which legal requirements can be tested from the algorithm. This stage of the firm's pricing decision is often considered the firm's "conduct" which can be scrutinized for identifying

[14] We assume that prices are directly obtained from predictions, and our focus of predicted default probabilities is therefore without loss of generality. Other authors, like Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, *Algorithmic decision making and the cost of fairness*, ARXIV PREPRINT ARXIV:1701.08230 (2017) instead consider a separate step that links predictions to decisions. In order to apply our framework to such a setup, we would directly consider the resulting pricing rule.

[15] Mortgage originators are required to disclose mortgage application information under the Home Mortgage Disclosure Act, including applicant race. The Boston HMDA dataset combines data from mortgage applications made in 1990 in the Boston area with a follow-up survey collected by the Federal Reserve Bank of Boston. Further information on the dataset can be found in the online appendix.

[16] The HMDA dataset only includes applicant status and so we need to simulate default rates to engage in a default prediction exercise.

[17] The online appendix contains more details on how this data was constructed.

the particular policy that led to the disparity. Unlike the context of human decision-making where conduct is not fully observed, in algorithmic decision-making we are able to observe the decision rule. We argue, however, that this transparency is limited to the types of issues that are interpretable in the algorithmic context. In particular, many machine-learning methods do not allow a reliable determination of which variables are important for the decision rule.

Third, we consider the statistical analysis of the resulting prices and argue that the observability of the decision rules expands the opportunities for controlled and preemptive testing of pricing practices. The analysis of the outcome becomes more central in the context of algorithmic decision-making given the inherent limitations of an analysis of the input and decision process stage. Moreover, outcome analysis in this new context is not limited to actual prices paid by consumers as we are able to observe the decision rule for future prices, allowing for forward-looking analysis of decision rules. This type of analysis is especially useful for regulators that enforce antidiscrimination law.

Our framework contributes to bridging the gap between the literature on algorithmic fairness and anti-discrimination law. Recent theoretical, computational and empirical advances in computer science and statistics provide different notions of when an algorithm produces fair outcomes and how these different notions relate to one another. However, many of these contributions focus solely on the statistical analysis of outcomes, but neither explicitly consider other aspects of the algorithmic decision process nor relate the notions of fairness to legal definitions of discrimination.[18] By providing a framework that relates the analysis of algorithmic decision-making to legal doctrine, we highlight how results from this literature can inform future law through the tools it has developed for the statistical analysis of outcomes.

II. Setup for Illustration and Simulation

Throughout this paper, we consider the legal and methodological challenges in analyzing algorithmic decision rules in a stylized setting that we illustrate with simulated data. A firm sets loan terms for new consumers based on observed defaults of past clients. Specifically, the company learns a prediction of loan default as a function of individual characteristics of the loan applicant from a training sample. It then applies this prediction function to new clients in a

[18] There are some exceptions. See for example: Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian, *Certifying and removing disparate impact*, ARXIV PREPRINT ARXIV:1412.3756 (2014). Although the paper attempts to provide a legal framework on algorithmic fairness, its focus on the Equal Employment Opportunity Commission 80 percent rule, and fairness as the ability to predict the protected class, does not capture the most significant aspects of discrimination law. Prior literature has suggested that big data may pose challenges to discrimination law particularly for Title VII employment discrimination. See for example Solon Barocas & Andrew D Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (focusing on several channels, primarily through biased human discretion in the data generating process, in which the data mining will reinforce bias). In contrast, our argument even applies when there is no human bias in past decisions. For a paper focused on the issues that big data cases for discrimination law in the context of credit scores, see Mikella Hurley & Julius Adebayo, *Credit scoring in the era of big data*, 18 YALE JL & TECH. 148 (2016) (focusing on the issues of transparency created by big data which will limit people's ability to challenge their credit score).

held-out dataset. This setup would be consistent with the behavior of a firm that aims to price loans at their expected cost.

In order to analyze algorithmic credit pricing under different constraints, we simulate such training and hold-out samples from a model that we have constructed from the Boston HMDA dataset. While this simulated data includes race identifiers, our model assumes that race has no direct effect on default. We calibrate overall default probabilities to actual default probabilities form the literature, but since all defaults in this specific model are simulated and not based on actual defaults, any figures and numerical examples in this paper should not be seen as reflecting real-world observations. Rather, our simulation illustrates methodological challenges in applying legal doctrine to algorithmic decision making.

In the remaining paper, we highlight methodological challenges in analyzing algorithmic decision-making by considering two popular machine-learning algorithms, namely the random forest and the lasso. Both algorithms are well-suited to obtain predictions of default from a high-dimensional dataset. Specifically, we train both algorithms on a training sample with 2000 clients, with over 40 variables each (many of which are categorical). We then analyze their prediction performance on a hold-out dataset with 2000 new clients, drawn from the same model. While these algorithms are specific, we focus on general properties of algorithmic decision-making in big data.

III. Data Inputs and Input-Focused Discrimination

One aspect of many anti-discrimination regimes is to restrict inputs that can be used to price credit. Typically, this would mean that the protected characteristics, such as a race and gender, cannot be used in setting prices. Indeed, many discrimination regimes include rules on the exclusion of data inputs as a form of discrimination prevention. For example, ECOA regulation provides that: "Except as provided in the Act and this regulation, a creditor shall not take a prohibited basis into account in any system of evaluating the creditworthiness of applicants."[19] Moreover, the direct inclusion of a forbidden characteristic in the decision process could trigger the "disparate treatment" doctrine since the forbidden attribute could directly affect the decision.

Despite the centrality of input restriction to discrimination law, the enforcement of these rules is difficult when the forbidden attribute is observable to the decision maker.[20] When a decision maker, such as a job interviewer or mortgage broker, observes a person's race, for example, it is

[19] HUD 2013 Regulation, section 1002.6.

[20] There is a further issue that we do not discuss regarding the inherent tension between excluding certain characteristics from consideration on the one hand, and the requirement that a rule not have disparate impact, that requires considering these characteristics. For further discussion of this tension see: Richard A Primus, *Equal protection and disparate impact: Round three*, 117 HARV. L. REV. 494 (2003).

impossible to rule out that this characteristic played a role in the decision, whether consciously or sub-consciously. The vast majority of credit disparate impact cases deal with situations in which there is a human decision-maker,[21] meaning that it is impossible to prove that belonging to a protected group was not considered. As discussed in the introduction, in the series of mortgage lending cases in which mortgage brokers had discretion in setting the exact interest and fees of the loan, it is implicit that customers' race was known to the brokers who met face-to-face with the customers. Therefore, we cannot rule out that race was an input in the pricing outcome.

The perceived opportunity for algorithmic decision-making is that it allows for formal exclusion of protected characteristics, but we argue that it comes with important limitations. When defining and delineating the data that will be used to form a prediction, we can guarantee that certain variables or characteristics are excluded from the algorithmic decision. Despite this increased transparency that is afforded by the automation of pricing, we show that there are two main reasons that discrimination regimes should not focus on input restriction. First, we argue that if price disparity matters, input restriction is insufficient. Second, the inclusion of the forbidden characteristic may in fact decrease disparity, particularly when there is some measurement bias in the data.

A. Exclusion is limited

The formal exclusion of forbidden characteristics, such as race, would exclude any direct effect of race on the decision. This means that we would exclude any influence that race has on the outcome that is not due to its correlation with other factors. We would therefore hope that excluding race already reduces a possible disparity in risk predictions between race groups in algorithmic decision-making. However, when we exclude race in our simulation exercise, we show below that there is little change in how risk predictions differ between protected groups.

To demonstrate that disparity can indeed persist despite the exclusion of input variables, consider the three graphs below (Figure 1) that represent the probability density function of the predicted default rates of the customers in a new sample not used to train the algorithm, by race/ethnicity and using a random forest as a prediction algorithm. On the left the distribution of predicted default rates was created using the decision rule that included the group identity as an input. We can see that the predicted default distributions are different for Whites, Blacks and Hispanics. The median prediction for each group is represented by the vertical lines. The middle graph shows the distribution of predicted default rates when race is *excluded* as an input from the algorithm that produced the decision rule. Despite the exclusion of race, much of the difference between groups persists.

[21] See discussion in: Aleo & Svirsky, BU PUB. INT. LJ, (2008), page 33.

Indeed, if there are other variables that are correlated with race, then predictions may strongly vary by race even when race is excluded, and disparities persist. For example, if applicants of one group on average have lower education, and education is used in pricing, then using education in setting prices can imply different prices across groups. If many such variables come together, disparities may persist. In very high-dimensional data, and when complex, highly non-linear prediction functions are used, this problem that one input variable can be reconstructed jointly from the other input variables becomes ubiquitous.

One way to respond to the indirect effect of protected characteristics is to expand the criteria for input restriction. For example, if an applicant's neighborhood is highly correlated with an applicant's race, we may want to restrict the use of one's neighborhood in pricing a loan. A major challenge of this approach is the required articulation of the conditions under which exclusion of data inputs is necessary. One possibility would be to require the exclusion of variables that do not logically relate to default, an approach that relies on intuitive decisions since we do not know what causes default. Importantly, it is hard to reconcile these intuitive decisions with the data-driven approach of machine learning in which variables will be selected for carrying predictive power.

Furthermore, the effectiveness of these types of restrictions are called into question because even excluding other variables that are correlated with race has limited effect in big data. In the third graph, we depict the predicted default rates using a decision rule that was created while excluding race and ten variables that correlate the most with race. Despite significantly reducing the number of variables that correlate strongly with race, the disparity still persists for the three racial groups, even though it is now smaller. The purpose of these three graphs is to demonstrate the impact that correlated data has on the decision rule, even when we exclude the forbidden characteristics or variables that may be deemed closer to the forbidden characteristics. In big data, even excluding those variables that individually relate most to the "forbidden input" does not necessarily affect much how much pricing outputs vary with, say, race.[22]

[22] An alternative approach taken in the algorithmic fairness literature is to transform variables that correlate with the forbidden characteristic as a way of "cleaning" the training data. See for example: Feldman, et al., ARXIV PREPRINT ARXIV:1412.3756, (2014). For a general discussion of these approaches see: James E Johndrow & Kristian Lum, *An algorithm for removing sensitive information: application to race-independent recidivism prediction*, ARXIV PREPRINT ARXIV:1703.04957 (2017), page 3.

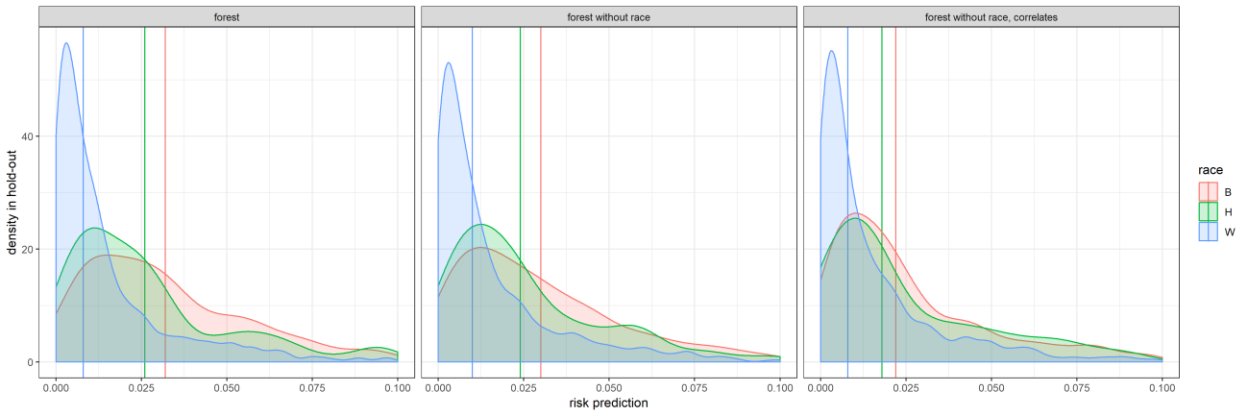

Figure 1: Distribution of risk predictions across groups for different inputs

If disparate impact is a proxy for disparate treatment or a means of enforcing disparate treatment,[23] we may find it sufficient that we can guarantee that there is no direct effect of race on the decision. Although it has long been recognized that a claim of disparate impact does not require a showing of intention to discriminate, which has traditionally been understood as the domain of disparate treatment, it is disputed whether the purpose of disparate impact is to deal with cases in which intention is hard to prove, or whether the very foundation of the disparate impact doctrine is to deal with cases where there is no intention to discriminate. There are several aspects of how disparate impact has been interpreted and applied that support the notion that it is a tool for enforcing disparate treatment rather than a theory of discrimination that is philosophically distinct.[24] According to the Supreme Court in *Texas Department of Housing & Community Affairs v. The Inclusive Communities Project*, "Recognition of disparate-impact liability under the FHA plays an important role in uncovering discriminatory intent: it permits plaintiffs to counteract unconscious prejudices and disguised animus that escape easy classification as disparate treatment."[25]

On the other hand, most formal articulations are clear that disparate impact can apply even when there is no discriminatory intent, and not only when discriminatory intent is not established.[26] This understanding of the disparate impact doctrine also seems more in line with perceptions of regulators and agencies that enforce discrimination law in the context of credit.[27] To the extent

[23] For a discussion of this view see for example: Richard Arneson, *Discrimination, Disparate Impact, and Theories of Justice*, *in* PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW (Deborah Hellman and Sophia Moreau ed. 2013). Also see: Primus, HARV. L. REV., (2003)

[24] A disparate impact claim can only be sustained if the plaintiff has not demonstrated a business necessity for the conduct. Conduct that lacks a business justification and led to a discriminatory outcome raises the suspicion that it is ill-intended. Moreover, many cases that deal with human decision-making seem to imply that there may have been intention to discrimination. For example, in Watson (see footnote X), the court emphasized that while the delegation of promotion decisions to supervisors may not be with discriminatory intent, it is still possible that the particular supervisors had discriminatory intent (page 990).

[25] Inclusive Communities, S. Ct.

[26] See HUD 2013 Regulation, page 11461 ("HUD… has long interpreted the Act to prohibit practices that have an unjustified discriminatory effect, regardless of intent")

[27] See HUD 2013 Regulation, page 11461 and ECOA Regulation B.

that disparate impact plays a social role beyond acting as a proxy for disparate treatment,[28] we may not find it sufficient to formally exclude race from the data considered.

B. Exclusion may be undesirable

Another criterion for the exclusion of inputs beyond the forbidden characteristics themselves are variables that may be biased. Variables could be biased because of some measurement error or because the variables reflect some historical bias, for example income that may correlate with race and gender as a result of labor market discrimination, or lending histories that may be a result of prior discrimination in credit markets.[29] The various ways variables can be biased has been discussed elsewhere.[30]

When data includes biased variables, it may not be desirable to exclude a protected characteristic since the inclusion of protected characteristics may allow the algorithm to correct for the biased variable.[31] For example, over the years there has been mounting criticism of credit scores, since they consider measures of creditworthiness that are more reflective of certain groups while overlooking indications of creditworthiness more prevalent for minority groups. One way this might happen is through credit rating agencies focusing on credit that come from mainstream lenders like depository banking institutions. However, if minority lenders are more likely to turn to finance companies that are not mainstream lenders, and this credit is treated less favorably by credit rating agencies,[32] the credit score may reflect the particular measurement method of the agency rather than the underlying creditworthiness in a way that is biased against minorities. If credit scores should receive less weight for minority borrowers, a machine-learning lender that uses a credit score as one of its data inputs would want to be able to use race as another data input in order to distinguish the use of credit scores for different groups.

Achieving less discriminatory outcomes by including forbidden characteristics in the prediction algorithm presents a tension between the input focused "disparate treatment" and the outcome focused "disparate impact" doctrines. This tension created by the requirement to ignore forbidden characteristics and yet assure that policies do not create disparate impact, thereby requiring a consideration of people's forbidden characteristics, has been debated in the past.[33] In the context of machine-learning credit pricing, including forbidden characteristics could

[28] This approach to disparate impact has been labeled as an "affirmative action" approach to disparate impact. See Arneson. 2013., page 105.
[29] See for example: Hurley and Adebayo, page 156 (discussing how past exclusion from the credit may affect future exclusion through credit scores).
[30] See for example the discussion in: Barocas & Selbst, CALIF. L. REV., (2016), page 677.
[31] Jens Ludwig, Jon Kleinberg, Sendhil Mullainathan, and Ashesh Rambachan. 5/2018. "Algorithmic Bias." AEA Papers and Proceedings, 108, Pp. 22-27.
[32] For a discussion see: Lisa Rice & Deidre Swesnik, *Discriminatory effects of credit scoring on communities of color*, 46 SUFFOLK UL REV. 935 (2013).
[33] See the discussion on Ricci v. DeStefano, 557 US 557 (2009) in Primus, HARV. L. REV., (2003).

potentially allow for the mitigation of harm of variables that suffer from biased measurement error.[34]

To summarize this section, despite the significant opportunity for increased transparency afforded by automated pricing, legal rules that focus on input regulation will have limited effect. On the one hand, unlike the human decision-making context we can guarantee that input has been excluded. However, if we care about outcome, we should move away from focusing on input restrictions as the emphasis of discrimination law. This is because input exclusion cannot eliminate and may even exacerbate pricing disparity.

IV. Algorithmic Construction and Process-Focused Discrimination

In the context of human credit pricing, most of the process of decision-making is opaque, leading to a limited ability to examine this process. Consider the mortgage lending cases described in the introduction, in which mortgage brokers determined the mark-up above the "par rate" set by the mortgage originator. In those cases, the broker decisions led to racial price disparity, however it is unclear why exactly the broker decisions led to these differences. The brokers could have considered customers' race directly and charged minorities higher prices, or perhaps brokers put disproportionate weight on variables that are correlated with race, such as borrower neighborhood. Although the exact nature of these decisions could lead to different conclusion as to the discrimination norm that was violated, these questions of the exact nature of the broker decisions remain speculative given that we have no record of the decision-making process of the broker.

To overcome the inherent difficulty in recovering the exact nature of the particular decision that may have been discriminatory, cases often abstract away by focusing on the facilitation of discriminatory decisions. The limited ability to scrutinize the decisions themselves leads courts and regulators to identify the discretion provided to brokers when setting the mortgage terms as the conduct that caused disparity.

Algorithmic decision-making presents an opportunity for transparency. Unlike the human decision-making context in which many aspects of the decision remain highly opaque, sometimes even to the decision makers themselves, in the context of algorithmic decisions-making we can observe many aspects of the decision and therefore can scrutinize these decisions to a greater extent. The decision process that led to a certain outcome can theoretically be recovered in the context of algorithmic decision-making providing for potential transparency that is not possible with human decision-making.[35]

[34] For an example of an application see: Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan and Ashesh Rambachan, Algorithmic Fairness, AER Papers & Proceedings (2018).

[35] This may not true for aspects of the process that involve human discretion, such as the label and feature selection.

However, this transparency is constrained by the limits to interpretability of decision rules. Prior legal writing on algorithmic fairness often characterizes algorithms as opaque and uninterpretable,[36] however whether an algorithm is interpretable depends on the question being asked. Despite the opaqueness of the mortgage broker decisions, these decisions are not referred to as "uninterpretable". Instead analytical and legal tools have been developed to consider the questions that can be answered in that context. Similarly, in the context of machine learning we need to understand what types of questions can be answered and analyzed and then develop the legal framework to evaluate these questions. There are many ways in which algorithms can be interpreted thanks to the increased replicability of their judgements. Indeed, we highlight in the last section a crucial way in which algorithms can be interpreted for the purposes of *ex ante* regulation.

One potential way to interpret algorithmic decisions is to consider which variables are used by the algorithm, equivalent to interpreting coefficients in the context of regression analysis. Typically, in social science research, the purpose of regression analysis is the interpretation of the coefficients of the independent variables which often reflect a causal effect of the independent variable on the dependent variable. Analogously, in the case of machine learning, the decision rule is constructed by an algorithm, providing two related opportunities: first, the algorithm provides a decision rule (prediction function) that can be inspected and from which we can presumably determine which variables matter for the prediction, and second, we can inspect the construction of the decision rule itself and attempt to measure which variables were instrumental in forming the final rule. In the case of a prediction rule that creates differing predictions for different groups, we may want to look to the variables used to make a prediction to understand what is driving the disparate predictions.

However, in the context of machine-learning prediction algorithms, the contribution of individual variables is often hard to assess. We demonstrate the limited expressiveness of the variables an algorithm uses by running the prediction exercise in our simulation example repeatedly. Across ten draws of data from that same population, we fit a logistic lasso regression where in every draw we let the data choose which of the many characteristics to include in the model, expecting that each run should produce qualitatively similar prediction functions. Although these samples are not identical because of the random sampling, they are drawn from the same overall population and we therefore expect that the algorithmic decisions produce similar output.

The outcome of our simulation exercise documents the problems with assessing an algorithm by the variables it uses. The specific representation of the prediction functions, and which variables are used in the final decision rule, vary considerably in our example. A graphic representation of this instability can be found in Figure 2. This figure records which characteristics were included in the logistic lasso regressions we ran on ten draws from the population. Each column represents a draw, while the vertical axis enumerates the over 80 dummy-encoded variables in our dataset.

[36] See for example, Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, CHICAGO-KENT LAW REVIEW (2018)page 44 (discussing how consumers may find it difficult to protect themselves because "many learning algorithms are thought to be quite opaque")

The black lines in each column reflect the particular variables that were included in the logistic lasso regression for that sample draw. While some characteristics (rows) are included in the model persistently, there are few discernible patterns, and an analysis of these prediction functions based on which variables were included would yield different conclusions from draw to draw, despite originating from similar data.

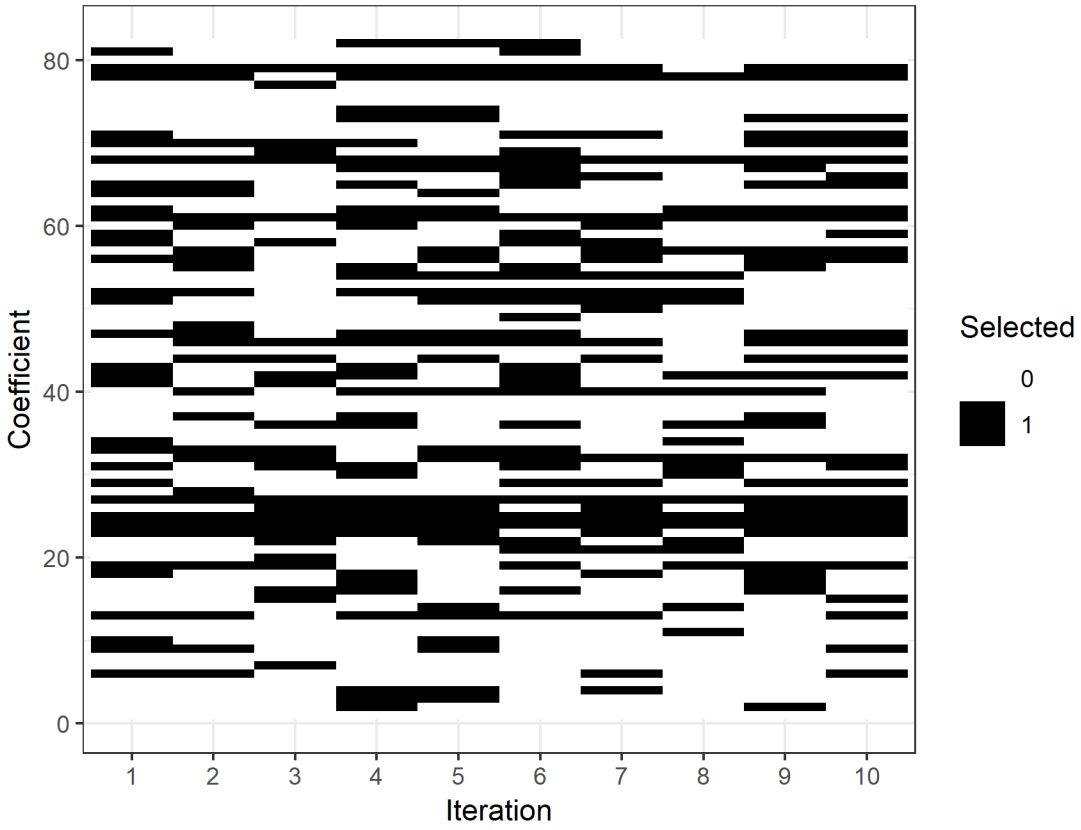

Figure 2: Included predictors in a lasso regression across ten samples from the same population

Importantly, despite these rules looking vastly different, their overall predictions indeed appear qualitatively similar. Figure 3 shows the distribution of default predictions by group for the first three draws of our ten random draws, documenting that they are qualitatively similar with respect to their pricing properties across groups. So while the prediction functions look very different, the underlying data, the way in which they were constructed, and the resulting price distributions are all similar.

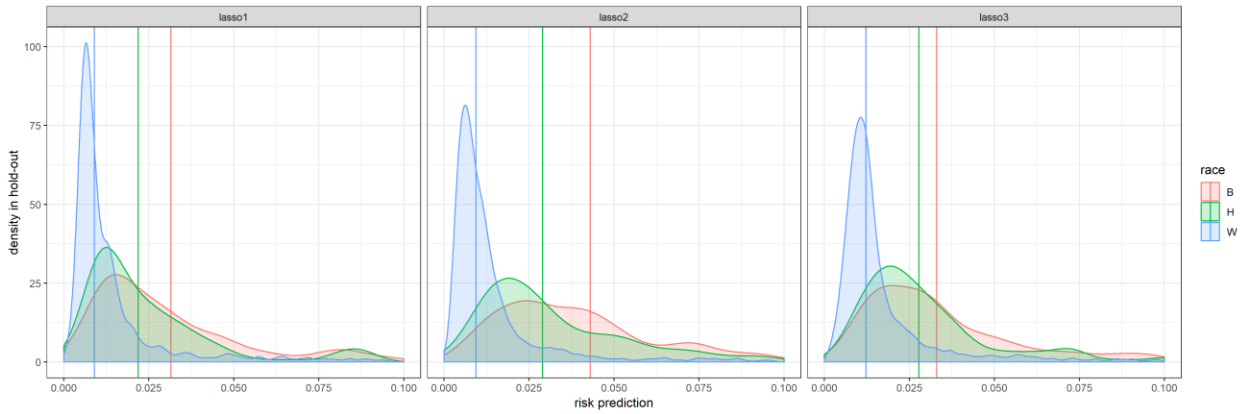

Figure 3: Distribution of default predictions for the three first lasso predictors

The instability of the variables chosen for the prediction suggests we should be skeptical about any type of inclusion of variables as an interpretation of how variables relate to the ultimate prediction. The primary object of a machine learning algorithm is the accuracy of the prediction and not a determination of the effect of specific variables in determining the outcome. When there are many possible characteristics that predictions can depend on, and algorithms choose from a large, expressive class of potential prediction functions, then many rules that look very different have qualitatively similar prediction properties. Which of these rules is chosen in a given draw of the data then may come down to a flip of a coin. When data is high-dimensional and complex machine-learning algorithms are used, a determination of conduct based on variable-importance measures is thus limited. The specific representation of prediction functions, and specifically which variables are used in the final decision rule, is therefore not generally an appropriate description of the actual predictions that the rule yields.[37]

The problem with interpretability illustrated by this instability is important for how law approaches the evaluation of algorithms. We demonstrate that the deconstruction of the prediction in hope of recovering the causes of disparity and maybe even to consider which variables should be omitted from the algorithm to reduce disparity is limited. Even without this issue of instability of the prediction rule, it may be hard to intelligently describe the rule when it is constructed from many variables, all receiving only marginal weight. Therefore, legal rules that seek to identify the cause or root of disparate decisions cannot be easily applied.

Thus, legal doctrines that put weight on identifying a particular conduct that caused disparity will not be able to rely on variable inclusion or importance analysis alone. Courts have interpreted the FHA , ECOA and their implementing regulations as requiring that the plaintiff demonstrate a causal connection between a discriminatory outcome and a specific practice in order to establish a *prima facie* case of discrimination.[38] The Supreme Court recently affirmed the requirement for the identification of a particular policy that caused the disparity in *Inclusive Communities*, where is said that: "a disparate-impact claim that relies on a statistical disparity must fail if the plaintiff

[37] See Sendhil Mullainathan & Jann Spiess, *Machine learning: an applied econometric approach*, 31 JOURNAL OF ECONOMIC PERSPECTIVES (2017) for a discussion, in particular Figure 2 for a discussion of instability.
[38] See for example *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989). Also see HUD 2013 Regulation.

cannot point to a defendant's policy or policies causing that disparity."[39] In the mortgage lending cases discussed in the introduction the conduct was the discretion given to mortgage lender employees and brokers. In other housing contexts, policies that have been found to cause a discriminatory effect include landlord residency preferences that favor people with local ties over outsiders and land-use restrictions that prevent housing proposals that are of particular value to minorities.[40] At first blush, it may seem appropriate to ask which of the variables that are included in the decision rule are those that led to the pricing being differential, as corresponding to the conduct identification requirement of the discrimination doctrine.[41] However, our example above demonstrates that such an analysis is questionable and unlikely to be appropriate as the algorithmic equivalent of identifying conduct.[42] Discrimination doctrine should therefore move away from this type of abstract decision rule analysis as a central component of discrimination law.

V. Implied Prices and Outcome-Focused Discrimination

In the previous section we argued that, despite its purported transparency, the analysis of machine-learning pricing is constrained by limits to the interpretability of abstract pricing rules. In this section, we argue that the replicability that comes with automation still has meaningful benefits for the analysis of discrimination when the pricing rule is applied to a particular population. The resulting price menu is an object that can be studied and analyzed and therefore should play a more central role in discrimination analysis. We consider a novel *ex ante* form of regulation which we call "discrimination stress-testing" that exploits the opportunity that automated decision rules can be evaluated before they are applied to actual consumers.

The final stage of a lending decision is the pricing "outcome", meaning the prices paid by consumers. In a world in which credit pricing involves mortgage brokers in setting the final lending terms, do not know their pricing outcome until the actual prices have materialized, and see prices only for the actual consumers. When pricing is automated, however, we also have information about pricing even before customers receive loans from inspecting the pricing rule.

---

[39] Inclusive Communities, S. Ct. at 2523.

[40] For further discussion of examples of challenged policies , see: Robert G Schwemm & Calvin Bradford, *Proving Disparate Impact in Fair Housing Cases after Inclusive Communities*, 19 NYUJ LEGIS. & PUB. POL'Y 685 (2016).

[41] Although automated credit systems have been challenged in court, court decisions rarely provide guidance on this question. For example, in the case of *Beaulialice v. Federal Home Loan Mortgage Corporation* 2007 WL 744646, *4 (M.D. Fla. Mar. 6, 2007, the plaintiff challenged the automated system used to determine their eligibility for a mortgage. Although the defendant's motion for summary judgement was granted, the basis for the decision was not because the plaintiff had not raised a plausible conduct for a claim of disparate impact. Alternatively, if the mere decision to use an algorithm is the "conduct" that caused discrimination this will essentially devoid the requirement of any meaningful content, strengthening the conclusion that the identification of a policy should be replaced with a greater emphasis on other elements of the analysis, such the outcome analysis discussed in the next section.

[42] There may be situations in which a particular aspect of the construction of the algorithm can be identified as leading to discrimination. As discussed in prior literature, biased outcomes may be a result of human decisions regarding the use and construction of the data. See for example: Barocas & Selbst, CALIF. L. REV., (2016); Hurley & Adebayo, YALE JL & TECH., (2016).

Furthermore, the pricing rule can be applied to any population, real or theoretical, to understand the pricing distribution that the pricing rule creates. Therefore, the set of potential outcomes that a legal regime can analyze is broader than the set of outcomes that can be analyzed in the case of human decision-making and the richness of information that is available at an earlier point in time means that the practices of the lender can be examined before waiting a period of time to observe actual prices.

Although pricing outcome analysis plays an important conceptual role in discrimination law, it is debatable how to practically conduct this analysis. Formally, outcome analysis that shows that prices provided to different groups diverge is part of the *prima facie* case of disparate impact. However, despite the centrality of outcome analysis, there is surprisingly little guidance on how exactly to conduct outcome analysis for the purposes of a finding of discrimination.[43] For example, we know little about the criteria to use when comparing two consumers to determine whether they were treated differently, or, in the language of the legal requirement, whether two "similarly situated" people obtained different prices. In addition, there is little guidance on the question of the relevant statistical test to use. As a result, output analysis in discrimination cases often focusses on simple comparisons and regression specifications,[44] moving quite swiftly to other elements of the case that are afforded a more prominent role, such as the discussion of the particular conduct or policy that lead to a disparate outcome.

In the case of machine learning, we argue that outcome analysis becomes central to the application of discrimination law. As is discussed in the previous two sections, both input regulation and decision process scrutiny are limited in the context of machine-learning pricing. Crucial aspects of current discrimination law, which focus on the procedure of creating the eventual prices, are undermined by the inability to properly interpret the decision rule that leads to the disparity and the challenges of closely regulating data inputs. Therefore, discrimination law will need to shift its focus to outcome analysis in the context of machine-learning credit pricing.

The type of *ex-ante* analysis, which we call "discrimination stress-testing", is most similar to bank stress-testing, which also evaluates an outcome using hypothetical parameters. Introduced in February 2009 as part of the Obama Administration's Financial Stability Plan, and later formalized in Dodd Frank,[45] stress-tests require certain banks to report the bank's stability under hypothetical financial scenarios. These scenarios are determined by the Federal Reserve and specify the macroeconomic variables, such as the GDP growth and housing prices, that the bank needs to assume in its predicted portfolio risk and revenue. The results of these test help to determine whether the bank should increase its capital and provides and general assessment of the bank's resilience. This allows for a form of regulation that is forward-looking and provides a

[43] See Schwemm & Bradford, NYUJ LEGIS. & PUB. POL'Y, (2016) (arguing that neither *inclusive Communities* nor HUD provide any guidance on how to establish differential pricing for a *prima facie* case of discrimination under FHA and showing lower courts rarely followed the methodology established under Title VII.).
[44] See AYRES, et al. 2017
[45] Dodd- Frank Wall Street Reform and Consumer Protection Act (Pub. L. 111–203, 2010).

consistent estimate across banks.[46] In a discrimination stress-test, the regulator would apply the pricing rule of the lender to some hypothetical population and then evaluate whether the pricing meets some criteria of disparity before the lender implements the rule.[47]

Developing the precise discrimination stress-test requires articulating how the test will be implemented and the criteria used to judge pricing outcomes. A full analysis of these issues is beyond the scope of this paper. Instead we highlight two main concerns in the development the discrimination stress-test. First, we demonstrate the significance of selecting a particular population to which the pricing rule is applied. Second, we discuss the importance of the particular statistical test used to evaluate pricing disparity.

A. Population Selection

The first aspect of the discrimination stress-test that we wish to highlight is that disparity highly depends on the particular population to which the pricing rule is applied. The opportunity in the context of machine-learning pricing is that prices can be analyzed *ex ante*. However, this analysis can only be conducted when applying the rule to a particular population. Therefore, regulators and policy-makers need to determine what population to use when applying a forward-looking test.

The decision what population to use for testing is important because the disparity created by a pricing rule is highly sensitive to the particular population. If price disparity is created by groups having different characteristics beyond the protected characteristic, like race, the correlations of characteristics with race will determine price disparity.

We demonstrate the sensitivity of disparate outcomes to the particular borrower population by applying the same price rule to two different populations. We split our simulated sample into two geographical groups. One group covers lenders from Suffolk County, which covers some of the more urban areas of the Boston metropolitan area, and the other group covers more rural areas. Using the same prediction rule of default, we plot the distribution by race. While the rule – in this case a prediction based on a random forest – is exactly the same, the distribution of default predictions is qualitatively different between applicants in Suffolk county (right panel of Figure 4) and those in more rural areas of the Boston metropolitan area (left panel). Specifically, the same rule may induce ether a very similar (left) or quite different (right) distribution of predictions by group.

[46] MICHAEL S BARR, HOWELL E JACKSON & MARGARET E TAHYAR, FINANCIAL REGULATION: LAW AND POLICY (Foundation Press St. Paul. 2016), page 308.

[47] The power to announce future regulatory intention already exists within the CFPB's regulatory toolkit in the form of a No-Action Letter in which it declares that it does not intend to recommend the initiation of action against a regulated entity for a certain period of time. For example, in September 2017, the CFPB issued a No-Action Letter to Upstart, a lender that uses nontraditional variables to predict creditworthiness, in which it announced that it had no intention to initiate enforcement or supervisory action against Upstart on the basis of ECOA.

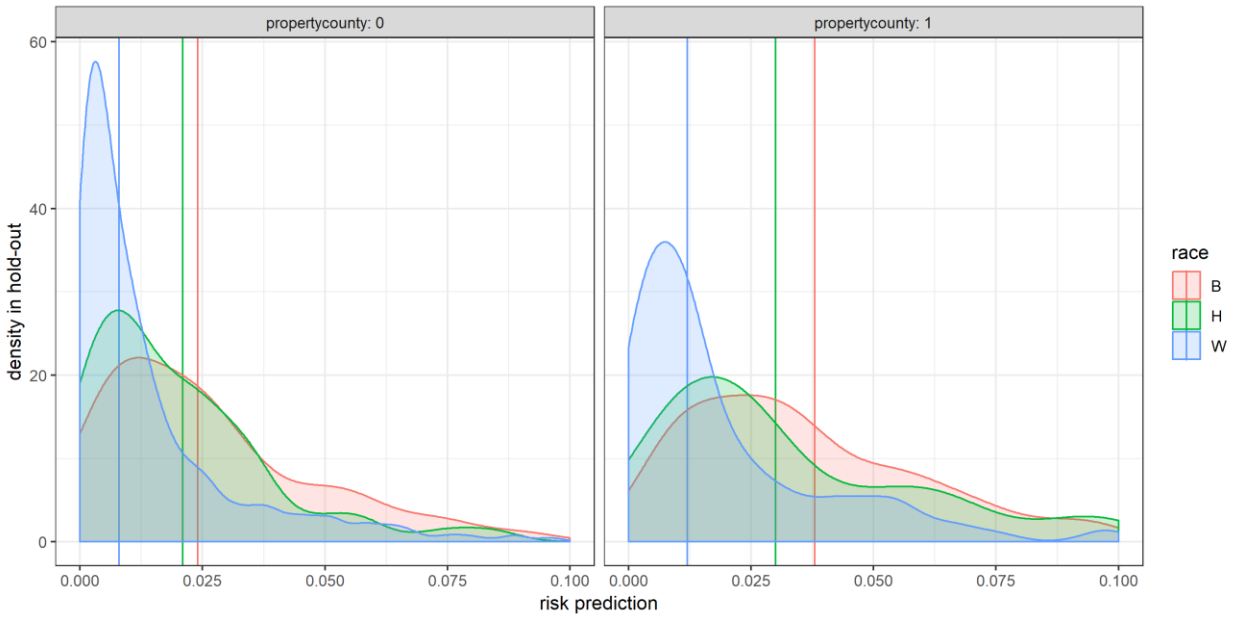

Figure 4: Risk predictions from the same prediction function across different neighborhoods

The sensitivity of outcomes to the population highlights two important considerations for policy-makers. First, it suggests that regulators should be deliberate in their selection of the population to use when testing. For example, they may want to select a population in which characteristics are highly correlated with race, or that represent more vulnerable lenders.[48] Second, if regulators wish to compare lender pricing rules, they should keep the population constant across lenders. This would provide for a comparable measure of disparity between lenders. Such meaningful comparisons are not possible with human decision-making when there is no pricing rule but only materialized prices. If regulators evaluate lending practices using *ex post* prices, differences between lenders may be driven by differences in decision rules or the composition of the particular population that received the loan.

The sensitivity of price disparities to the population also suggests that regulators should not disclose the exact sample they use to test discrimination. In this respect, the design of the discrimination stress-test could be informed by financial institution stress-testing. While the general terms of the supervisory model are made public, many of the details used to protect revenue and losses are kept confidential by regulators and are changed periodically limiting financial institution's ability to game the specifics of the stress test.[49] Similarly, for discrimination stress-testing, the exact dataset used could be kept confidential so that lenders are not able to create decision rules that minimize disparity for the specific dataset alone.

[48] Regulators are often interested in who the lender actually serviced, which may reveal whether the lender was engaging in redlining or reverse redlining. Clearly, this hypothetical is inappropriate for that analysis. For further discussion of redlining and reverse redlining see: Gano, U. COLO. L. REV., (2017).

[49] See BARR, et al. 2016 313 ("In essence, the Federal Reserve Board, by changing the assumptions and keeping its models cloaked, is determined that its stress tests cannot be gamed by the financial sector.").

Another benefit of population selection is that it allows for price disparity testing even when lenders do not collect data on race. Although mortgage lenders are required to collect and report race data under HMDA, other forms of lending do not have an equivalent requirement. This creates significant challenges for private and public enforcement of ECOA, for example. Discrimination stress-testing offers a solution to the problem of missing data on race. With discrimination stress-testing, the lender itself would not have to collect race data for an evaluation of whether the pricing rule causes disparity, as long as the population the regulator uses for the test includes protected characteristics. Since the regulator evaluates the pricing rule based on the prices provided to the hypothetical population, it can evaluate the effect on protected groups regardless of whether this data is collected by the lender.

B. Test for Disparity

Once the pricing rule is applied to a target population, the price distribution needs be evaluated. We focus on two aspects of this test, namely, the criterion by which two groups are compared and the statistical test used to conduct the comparison. A large literature originating in computer science discusses when algorithms should be considered fair.[50]

The first aspect of a test to disparity is the criterion used to compare groups. For example, we could ignore any characteristics that vary between individuals and simply consider whether the price distribution is different by group. Alternatively, the criterion for comparison could deem that certain characteristics, which may correlate with group membership, should be "controlled" for when comparing between groups. Meaning that in testing for disparity only individuals that share these characteristics are compared. Courts consider this issue by asking whether "similarly situated" people from the protected and non-protected group were treated differently.[51] Suppose that individuals with the same income, credit score and job tenure are considered "similarly situated". Then the distribution of prices is allowed to vary across groups provided that this variation only represents variation of the composition with respect to those characteristics that represent "similarly situated" individuals.[52] For example, prices may still differ between Hispanic and White applicants to the degree that those differences represent differences in income, credit score and job tenure.

---

[50] See for example: Emma Pierson Sam Corbett-Davies, Avi Feller, Sharad Goel & Aziz Huq, Algorithmic decision making and the cost of fairness (ACM 2017);Sendhil Mullainathan Jon Kleinberg, Sendhil & Manish Raghavan, *Inherent trade-offs in the fair determination of risk scores*, ARXIV PREPRINT ARXIV:1609.05807 (2016); Feldman, et al., ARXIV PREPRINT ARXIV:1412.3756, (2014).For a recent overview of the different notions of fairness see: Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMBERLAND L. REV. 102 (2018).

[51] This requirement has also been referred to as the requirement that demonstration of disparate impact focus on "appropriate comparison groups". See discussion in Jennifer L Peresie, *Toward a coherent test for disparate impact discrimination*, 84 IND. LJ 773 (2009), page 698; Also see Schwemm & Bradford, NYUJ LEGIS. & PUB. POL'Y, (2016) at page 698.

[52] Moritz Hardt Cythia Dwork, Toniann Pitassi, Omer Reingold, & Richard Zemel, *Fairness through awareness*, PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE (2012) provide concept that can be seen as an implementation of "similarly situated" people being treated the same, through connecting a metric of distance between people to how different their outcomes can be.

The longer the list of the characteristics that make people "similarly situated" the less likely it is that there will be a finding of disparity.[53] Despite the importance of this question, there is little guidance in cases and regulatory documents on which characteristics make people similarly situated.[54] An approach that considers what is predictive as making people similar would mean that by definition the algorithm is not treating similarly situated people differently. Especially in a big-data world with a large number of correlated variables, a test of statistical parity thus requires a clear implementation of similar situated to have any bite. Therefore, a determination of what makes people "similarly situated" is primarily a normative question that law-makers and regulators should address.

The determination of who is "similarly situated" is distinct from an approach of input restriction. Restricting inputs to "similarly situated" characteristics would guarantee that there is no disparity, however this is not necessary. Although a complete discussion of the conditions under which input variables that do not constitute "similarly situated" characteristics do not give rise to a claim of disparity is beyond this paper, we highlight two considerations. First, as argued throughout the paper, the particular correlations of the training set and holdout set will affect pricing disparity, and so little can be determined from the outset. If, for example, a characteristic does not correlate with race it may be that its inclusion in the algorithm will not lead to disparity. Second, the statistical test should include a degree of tolerance set by the regulator. When this tolerance is broader it is more likely that characteristics included in the algorithm may not give rise to a claim of disparity even when they are not "similarly situated" characteristics.

In addition to the criterion by which to compare groups, the regulator requires a test in order to determine whether indeed there is disparity.[55] Typically, for such a test the regulator needs to fix a tolerance level that expresses how much the distribution of risk predictions may deviate across groups between similarly situated individuals.

VI. Conclusion

[53] Ayres characterizes the problem as a determination of what variables to include as controls when regressing for the purpose of disparate impact. See: Ian Ayres, *Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of" included variable" bias*, 48 PERSPECTIVES IN BIOLOGY AND MEDICINE (2005)

[54] One exception is the Policy Statement on Discrimination in Lending, by HUD, the Department of Justice and other agencies from 1994 which suggested that the characteristics in HMDA do not constitute an exhaustive list of the variables that make people similarly situated. See section C.

[55] The algorithmic fairness literature includes many different test, some of which are summarized by MacCarthy, CUMBERLAND L. REV., (2018), page 88. One of the only examples of an articulated statistical test is the "four-fifths rule" adopted by the Equal Employment Opportunity Commission (EEOC) in 1979. We do not discuss the rule because its formulation does not seem natural in a context like credit pricing in which there is not a single criterion with pass rates. In addition, it is not clear the extent to which this is binding test given the tendency of courts to overlook the test. For further discussion see: Schwemm & Bradford, NYUJ LEGIS. & PUB. POL'Y, (2016). For an application of this test in the algorithmic fairness literature see: Feldman, et al., ARXIV PREPRINT ARXIV:1412.3756, (2014).

In this article, we present a framework that connects the steps in the genesis of an algorithmic pricing decision to legal requirements developed to protect against discrimination. We argue that there is a gap between old law and new methods that can only be bridged by resolving normative legal questions. These questions have thus far received little attention because they were of less practical importance a world where anti-discrimination law focused on opaque human decision-making.

While algorithmic decision-making allows for pricing to become traceable, the complexity and opacity of modern machine-learning algorithms limit the applicability of existing legal anti-discrimination doctrine. Simply restricting an algorithm from using specific information, for example, would at best satisfy a narrow reading of existing legal requirements, and typically have limited bite in a world of big data. On the other hand, scrutiny of the decision process is not always feasible in the algorithmic decision-making context, suggesting a greater role for outcome analysis.

Prices set by machines also bring opportunities for effective regulation provided that open normative questions are resolved. Our analysis highlights an important role for the statistical analysis of pricing outcomes. Since prices are set by fixed rules, discrimination stress tests are an opportunity to check pricing outcomes in a controlled environment. Such tests can draw upon criteria from the growing literature on algorithmic fairness which can also illuminate the inherent tradeoffs between different notions of discrimination and fairness.

This is a watershed moment for anti-discrimination doctrine, not only because the new reality requires an adaptation of an anachronistic set of rules, but because philosophical disagreements over the scope of discrimination law now have practical and pressing relevance.