Big Data and Discrimination

Talia B. Gillis† *& Jann L. Spiess*††

The ability to distinguish between people in setting the price of credit is often constrained by legal rules that aim to prevent discrimination. These legal requirements have developed focusing on human decision-making contexts, and so their effectiveness is challenged as pricing increasingly relies on intelligent algorithms that extract information from big data. In this Essay, we bring together existing legal requirements with the structure of machine-learning decision-making in order to identify tensions between old law and new methods and lay the ground for legal solutions. We argue that, while automated pricing rules provide increased transparency, their complexity also limits the application of existing law. Using a simulation exercise based on real-world mortgage data to illustrate our arguments, we note that restricting the characteristics that the algorithm is allowed to use can have a limited effect on disparity and can in fact increase pricing gaps. Furthermore, we argue that there are limits to interpreting the pricing rules set by machine learning that hinders the application of existing discrimination laws. We end by discussing a framework for testing discrimination that evaluates algorithmic pricing rules in a controlled environment. Unlike the human decision-making context, this framework allows for ex ante testing of price rules, facilitating comparisons between lenders.

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[1] (ECOA) forbids race, religion, age, and other

† Doctoral Student, Harvard Business School; John M. Olin Fellow in Empirical Law and Economics, Harvard Law School.

†† Post-doctoral Researcher, Microsoft Research New England.

For helpful feedback, we thank Oren Bar-Gill, Netta Barak-Corren, Yochai Benkler, John Beshears, Alexandra Chouldechova, Ellora Derenoncourt, Noah Feldman, Deborah Hellman, Howell Jackson, Cass Sunstein, Justin Wolfers, the editors and the participants of The University of Chicago Law Review Symposium on Personalized Law, and participants of the Law and Economics Colloquium at the University of Toronto and the Business Law Workshop at Oxford University. Talia Gillis acknowledges support provided by the John M. Olin Center for Law, Economics, and Business at Harvard Law School.

1 Pub L No 94-239, 90 Stat 251 (1976), codified as amended at 15 USC § 1691 et seq.

factors from being considered in setting credit terms,[2] and the Fair Housing Act[3] (FHA) prohibits discrimination in financing of real estate based on race, color, national origin, religion, sex, familial status, or disability.[4] Many of these rules were developed to challenge human discretion in setting prices and provide little guidance in a world in which 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 questions of when and how existing law can be applied to this novel setting, and when it must be adapted to remain effective.

In this Essay, 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 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

---

2 15 USC § 1691(a)(1)–(3).

3 Pub L No 90-284, 82 Stat 81 (1968), codified as amended at 42 USC § 3601 et seq.

4 42 USC § 3605(a). 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, Pub L No 95-128, 91 Stat 1111 (1977), codified at 12 USC § 2901 et seq, there are many state and local laws with discrimination provisions, such as fair housing laws.

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 facially neutral conduct has a discriminatory effect.[5] 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.[6] A comprehensive discussion of these two doctrines and their application to credit pricing is beyond the scope of this Essay, particularly because there are several aspects of these doctrines on which there is widespread disagreement.[7] 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.[8] The first approach is to focus on the "inputs" of the decision, stemming from the view that discrimination law is primarily concerned

---

5 For an overview of these two legal doctrines and their relation to theories of discrimination, see John J. Donohue, *Antidiscrimination Law*, in A. Mitchell Polinsky and Steven Shavell, 2 *Handbook of Law and Economics* 1387, 1392–95 (Elsevier 2007).

6 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, 2518 (2015), confirming the position of eleven appellate courts and various federal agencies, including the Department of Housing and Urban Development (HUD), which is primarily responsible for enforcing the FHA. See also generally 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 the ECOA, the Consumer Financial Protection Bureau and courts have found that the statute allows for a claim of disparate impact. See, for example, *Ramirez v GreenPoint Mortgage Funding, Inc*, 633 F Supp 2d 922, 926–27 (ND Cal 2008).

7 For further discussion of the discrimination doctrines under ECOA and FHA, see Michael Aleo and 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 L J 1, 22–38 (2008); Alex Gano, Comment, *Disparate Impact and Mortgage Lending: A Beginner's Guide*, 88 U Colo L Rev 1109, 1128–33 (2017).

8 We find it necessary to divide approaches to discrimination by their goal and focus because 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 for disparate impact or disparate treatment, such as in the three-part burden-shifting framework for establishing an FHA disparate impact case as formulated by HUD. See Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed Reg 11459, 11460–63 (2013), amending 24 CFR § 100.500.

with formal or intentional discrimination.[9] The second approach scrutinizes the decision-making process, policy, or conduct that then led to disparity. The third approach focuses on the disparity of the "outcome."[10] 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 that challenged mortgage pricing practices. In these cases, plaintiffs argued that black and Hispanic borrowers ended up paying higher interest rates and fees after controlling for the "par rate" set by the mortgage originator. The claim was that 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.[11] These types of assertions were made in the context of individual claims,[12] class actions,[13] and regulatory action.[14] What is most striking is that these cases do not directly scrutinize the broker decisions, treating them as a "black box," but focus instead on the mortgage originator's discretion policy.[15] Had the courts been able to analyze

---

9 This basic articulation is also used in Richard Primus, *The Future of Disparate Impact*, 108 Mich L Rev 1341, 1342 (2010). For a discussion on the different notions of intention, see Aziz Z. Huq, *What Is Discriminatory Intent?*, 103 Cornell L Rev 1212, 1240–65 (2018) (arguing that judicial theory of "intention" is inconsistent).

10 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.

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

12 See, for example, *Martinez v Freedom Mortgage Team, Inc*, 527 F Supp 2d 827, 833–35 (ND Ill 2007).

13 See, for example, *Ramirez*, 633 F Supp 2d at 924–25; *Miller v Countrywide Bank, National Association*, 571 F Supp 2d 251, 253–55 (D Mass 2008).

14 For a discussion of a series of complaints by the Justice Department against mortgage brokers that were settled, see Ian Ayres, Gary Klein, and Jeffrey West, *The Rise and (Potential) Fall of Disparate Impact Lending Litigation*, in Lee Anne Fennell and Benjamin J. Keys, eds, *Evidence and Innovation in Housing Law and Policy* 231, 240–46 (Cambridge 2017) (discussing cases against Countrywide (2011), Wells Fargo (2012), and Sage Bank (2015), all involving a claim that discretion to brokers resulted in discrimination).

15 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, Consent Order, *In the Matter of American Honda Finance Corporation*, No 2015-CFPB-0014, *5–9 (July 14, 2015) (available on Westlaw at 2015 WL 5209146). This practice has also been applied in

the discriminatory decisions directly, we would have had a greater understanding of the precise conduct that was problematic. As a result of the scope and range of the legal doctrine, which are important for the automated pricing context, discrimination cases that involve opaque human decisions do not allow us to develop the exact perimeters of the doctrine.[16]

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 then outputs a function that relates consumer characteristics, such as their income and credit score, 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 are 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. Because 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."[17]

We connect machine learning, decision rules, and current law by considering the three stages of a pricing decision, which we

---

other areas, such as employment discrimination cases. For example, the seminal employment discrimination case *Watson v Fort Worth Bank & Trust*, 487 US 977, 982–85 (1988), 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 338, 352–57 (2011) (holding that, in a suit alleging discrimination in Wal-Mart's employment promotion policies, class certification was improper because an employer's discretionary decision-making is a "presumptively reasonable way of doing business" and "merely showing that [a company's] policy of discretion has produced an overall sex-based disparity does not suffice" to establish commonality across the class).

[16] 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 decisionmakers themselves. There are decades of research showing that people have difficulty recovering the basis for their decisions, particularly when they involve race. See, for example, Cheryl Staats, et al, *State of the Science: Implicit Bias Review 2015* *4–6 (Kirwan Institute for the Study of Race and Ethnicity, 2015), archived at http://perma.cc/4AJ6-4P4C.

[17] We assume that prices are directly obtained from predictions, and our focus on predicted default probabilities is therefore without loss of generality. Other authors instead consider a separate step that links predictions to decisions. See, for example, Sam Corbett-Davies, et al, *Algorithmic Decision Making and the Cost of Fairness* *2–3 (arXiv.org, Jun 10, 2017), archived at http://perma.cc/9G5S-JDT8. In order to apply our framework to such a setup, we would directly consider the resulting pricing rule.

demonstrate in a simulation exercise. The data we use is based on real data on mortgage applicants from the Boston HMDA data set,[18] and we impute default probabilities from a combination of loan approvals and calibrate them to overall default rates.[19] 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 Essay should not be interpreted as reflecting real-world observations but rather methodological challenges and opportunities that arise in the context of algorithmic decision-making.[20]

The remainder of this Essay discusses each of the three steps of a pricing decision by underlining both the challenges and the opportunities presented by applying machine-learning pricing to current legal rules. First, we consider the data input stage of the pricing decision and argue that excluding forbidden characteristics has limited effect and satisfies only a narrow understanding of antidiscrimination law.[21] One fundamental aspect of antidiscrimination laws is the prohibition on conditioning a 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.[22] In addition, we highlight the ways in which restricting a broader range of data inputs may have unintended consequences, such as increasing price disparity.[23]

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.[24] This stage of the firm's pricing decision is often considered the firm's "conduct," which can

---

[18] Mortgage originators are required to disclose mortgage application information, including applicant race, under the Home Mortgage Disclosure Act (HMDA), Pub L No 94-200, 89 Stat 1124 (1975), codified at 12 USC § 2801 et seq. The Boston HMDA data set 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. See Alicia H. Munnell, et al, *Mortgage Lending in Boston: Interpreting HDMA Data*, 86 Am Econ Rev 25, 30–32 (1996). Further information on the data set can be found in an online appendix.

[19] The HMDA data set includes only applicant status, so we need to simulate default rates to engage in a default prediction exercise. The calibration of overall default rates is based on Andreas Fuster, et al, *Predictably Unequal? The Effects of Machine Learning on Credit Markets* (2018), archived at http://perma.cc/LYY5-SAG2.

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

[21] See Part II.

[22] See Part II.A.

[23] See Part II.B.

[24] See Part III.

be scrutinized for identifying the particular policy that led to the disparity. Unlike in the context of human decision-making, in which 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 general-purpose determination about which variables are important for the decision rule absent further clarification regarding what "importance" would mean in this legal context.

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.[25] The analysis of the outcome becomes attractive in the context of algorithmic decision-making given the 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 antidiscrimination 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.[26] 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.[27] By providing a framework that relates the analysis of

---

[25] See Part IV.

[26] See, for example, Jon Kleinberg, et al, *Algorithmic Fairness*, 108 AEA Papers and Proceedings 22, 22–23 (2018) (arguing that "across a wide range of estimation approaches, objective functions, and definitions of fairness, the strategy of blinding the algorithm to race inadvertently detracts from fairness").

[27] There are some exceptions. See, for example, Michael Feldman, et al, *Certifying and Removing Disparate Impact* *2–3 (arXiv.org, Jul 16, 2015), archived at http://perma.cc/ZAL7-6V75. Although the paper attempts to provide a legal framework for algorithmic fairness, its focus is on the Equal Employment Opportunity Commission's 80 percent rule and fairness as the ability to predict the protected class. The paper therefore does not capture the most significant aspects of antidiscrimination law. Prior literature has suggested that big data may pose challenges to antidiscrimination law, particularly for Title VII employment discrimination. See, for example, Solon Barocas and Andrew D.

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.

I. SETUP FOR ILLUSTRATION AND SIMULATION

Throughout this Essay, we consider the legal and methodological challenges in analyzing algorithmic decision rules in a stylized setting that we illustrate with simulated data. In our example, 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 held-out data set. 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 holdout samples from a model that we have constructed from the Boston HMDA data set. While this simulated data includes race identifiers, our model assumes that race has no direct effect on default.[28] We calibrate overall default probabilities to actual default probabilities from the literature, but because all defaults in this specific model are simulated and not based on actual defaults, any figures and numerical examples in this Essay 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 remainder of this Essay, 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 data set.

Selbst, *Big Data's Disparate Impact*, 104 Cal L Rev 671, 694–714 (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 applies even when there is no human bias in past decisions. For a paper focused on the issues that big data cases pose for antidiscrimination law in the context of credit scores, see generally Mikella Hurley and Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 Yale J L & Tech 148 (2016) (focusing on the transparency issues created by big data that will limit people's ability to challenge their credit score).

[28] Default rates may still differ between groups because individuals differ in other attributes across groups, but we assume in our model that the race identifier does not contribute variation beyond these other characteristics.

Specifically, we train both algorithms on a training sample with two thousand clients with approximately fifty variables each, many of which are categorical. We then analyze their prediction performance on a holdout data set with two thousand new clients drawn from the same model. While these algorithms are specific, we discuss general properties of algorithmic decision-making in big data.

II. DATA INPUTS AND INPUT-FOCUSED DISCRIMINATION

One aspect of many antidiscrimination regimes is a restriction on inputs that can be used to price credit. Typically, this means that protected characteristics, such as race and gender, cannot be used in setting prices. Indeed, many antidiscrimination regimes include rules on the exclusion of data inputs as a form of discrimination prevention. For example, a regulation implementing the ECOA provides: "Except as provided in the Act and this part, a creditor shall not take a prohibited basis into account in any system of evaluating the creditworthiness of applicants."[29] Moreover, the direct inclusion of a forbidden characteristic in the decision process could trigger the disparate treatment doctrine because 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 decisionmaker.[30] When a decisionmaker, such as a job interviewer or mortgage broker, observes a person's race, for example, it is impossible to rule out that this characteristic played a role in the decision, whether consciously or subconsciously. The most common type of credit disparate impact case deals with situations in which there is a human decisionmaker,[31] meaning that it is impossible to prove that belonging to a protected group was not considered. As we discuss 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' races were known to the brokers who met face-to-face with the customers.

---

[29] 12 CFR § 1002.6(b)(1).

[30] There is a further issue that we do not discuss, which is the inherent tension between excluding certain characteristics from consideration on the one hand and the requirement that a rule not have disparate impact, which requires considering those characteristics. For further discussion of this tension, see generally Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 Harv L Rev 493 (2003).

[31] See Aleo and Svirsky, 18 BU Pub Int L J at 33–35 (cited in note 7).

Therefore, we cannot rule out that race was an input in the pricing outcome.

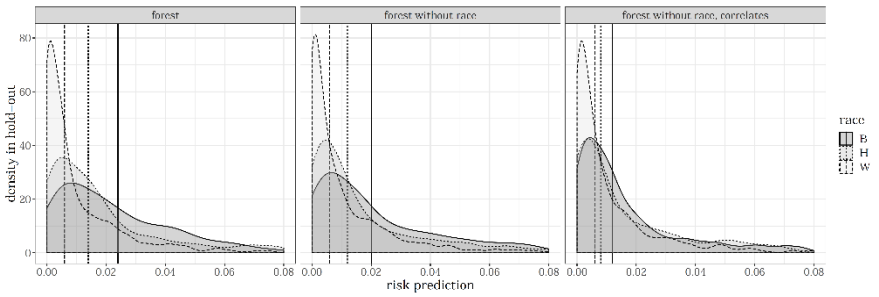
The perceived opportunity for algorithmic decision-making is that it allows for formal exclusion of protected characteristics, but we argue that it comes with important limitations. When defining and delineating the data that will be used to form a prediction, we can guarantee that certain variables 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 from fitting the algorithm 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 among groups persists.

FIGURE 1: DISTRIBUTION OF RISK PREDICTIONS ACROSS GROUPS FOR DIFFERENT INPUTS



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 may 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 nonlinear 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 because 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 is 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 by excluding race and the ten variables that most correlate 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.[32] In big data, even excluding those variables that individually relate most to the "forbidden input" does not necessarily significantly affect how much pricing outputs vary with, say, race.[33]

If disparate impact is a proxy for disparate treatment or a means of enforcing disparate treatment law,[34] 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 disparate impact claim 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 in which 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 law rather than a theory of discrimination that is philosophically distinct.[35] According to the Supreme Court in *Texas Department of Housing & Community*

---

[32] For a demonstration of the limited effect of excluding race from default prediction using data on real mortgage performance, see Fuster, et al, *Predictably Unequal* at *26–30 (cited in note 19).

[33] 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, *Certifying and Removing Disparate Impact* at *26 (cited in note 27). For a general discussion of these approaches, see James E. Johndrow and Kristian Lum, *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction* *3 (arXiv.org, Mar 15, 2017), archived at http://perma.cc/CR67-48PN.

[34] For a discussion of this view, see, for example, Richard Arneson, *Discrimination, Disparate Impact, and Theories of Justice*, in Deborah Hellman and Sophia Moreau, eds, *Philosophical Foundations of Discrimination Law* 87, 105 (Oxford 2013). See also generally Primus, 117 Harv L Rev at 493 (cited in note 30).

[35] A disparate impact claim can be sustained only if the plaintiff has not demonstrated a "business necessity" for the conduct. *Texas Department of Housing & Community Affairs v Inclusive Communities Project, Inc*, 135 S Ct 2507, 2517 (2015). 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 intent to discriminate. For example, in *Watson v Fort Worth Bank & Trust*, 487 US 977 (1988), 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. Id at 990.

Affairs v Inclusive Communities Project,[36] "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."[37]

On the other hand, most formal articulations are clear that disparate impact can apply even when there is no discriminatory intent, not only when discriminatory intent is not established.[38] This understanding of the disparate impact doctrine also seems more in line with perceptions of regulators and agencies that enforce antidiscrimination law in the context of credit.[39] To the extent that disparate impact plays a social role beyond acting as a proxy for disparate treatment,[40] 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 may correlate with race and gender as a result of labor market discrimination, and lending histories may be a result of prior discrimination in credit markets.[41] The various ways variables can be biased has been discussed elsewhere.[42]

When data includes biased variables, it may not be desirable to exclude a protected characteristic because the inclusion of protected characteristics may allow the algorithm to correct for the

---

36 135 S Ct 2507 (2015).

37 Id at 2511–12.

38 See 78 Fed Reg at 11461 (cited in note 8) ("HUD . . . has long interpreted the Act to prohibit practices with an unjustified discriminatory effect, regardless of whether there was an intent to discriminate.").

39 See id. See also 12 CFR § 1002.6(a) ("The legislative history of the [ECOA] indicates that the Congress intended an 'effects test' concept . . . to be applicable to a creditor's determination of creditworthiness.").

40 This approach to disparate impact has been labeled as an "affirmative action" approach to disparate impact. See Arneson, *Discrimination, Disparate Impact, and Theories of Justice* at 105–08 (cited in note 34). For further discussion of the different theories of disparate impact and their application to antidiscrimination policy in the algorithmic context, see generally Tal Z. Zarsky, *An Analytical Challenge: Discrimination Theory in the Age of Preditive Analytics*, 14 I/S: J L & Pol Info Society 11 (2017).

41 See, for example, Hurley and Adebayo, 18 Yale J L & Tech at 156 (cited in note 27) (discussing how past exclusion from the credit market may affect future exclusion through credit scores).

42 See, for example, Barocas and Selbst, 104 Cal L Rev at 677 (cited in note 27).

biased variable.[43] For example, over the years there has been mounting criticism of credit scores because they consider measures of creditworthiness that are more predictive for certain groups while overlooking indications of creditworthiness that are more prevalent for minority groups.[44]

One way this might happen is through credit rating agencies focusing on credit that comes from mainstream lenders like depository banking institutions. However, if minority borrowers are more likely to turn to finance companies that are not mainstream lenders, and if this credit is treated less favorably by credit rating agencies,[45] the credit score may reflect the particular measurement method of the agency rather than 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.[46]

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 requirements 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.[47] In the context of machine-learning credit pricing,

---

43 See generally Kleinberg, et al, 108 AEA Papers and Proceedings at 22 (cited in note 26). See also Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?* *9 (arXiv.org, Feb 28, 2018), archived at http://perma.cc/KH2Q-Z64F (discussing the shortcomings of algorithms that train on data with group membership but are group blind when used to make predictions and arguing that a transparent use of group membership can better achieve "impact parity").

44 See, for example, Lisa Rice and Deidre Swesnik, *Discriminatory Effects of Credit Scoring on Communities of Color*, 46 Suffolk U L Rev 935, 937 (2013) ("Credit-scoring systems in use today continue to rely upon the dual credit market that discriminates against people of color. For example, these systems penalize borrowers for using the type of credit disproportionately used by borrowers of color.").

45 See Rice and Swesnik, 46 Suffolk U L Rev at 949 (cited in note 44).

46 Prior economic literature on affirmative action has argued that group-blind policies may be second best in increasing opportunities for disadvantaged groups relative to group-aware policies. See, for example, Roland G. Fryer Jr and Glenn C. Loury, *Valuing Diversity*, 121 J Pol Econ 747, 773 (2013).

47 See, for example, the discussion of Primus, 117 Harv L Rev 494 (cited in note 30), in Justice Antonin Scalia's concurrence in *Ricci v DeStefano*, 557 US 557, 594 (2009) (Scalia concurring).

including forbidden characteristics could potentially allow for the mitigation of harm from variables that suffer from biased measurement error.[48]

* * *

To summarize this Part, despite the significant opportunity for increased transparency afforded by automated pricing, legal rules that focus on input regulation will have limited effect.[49] On the one hand, unlike in 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 antidiscrimination law. This is because input exclusion cannot eliminate and may even exacerbate pricing disparity.

III. ALGORITHMIC CONSTRUCTION AND PROCESS-FOCUSED DISCRIMINATION

In the context of human credit-pricing, most of the decision-making process is opaque, leading to a limited ability to examine this process. Consider the mortgage lending cases we describe in the Introduction, in which mortgage brokers determined the markup above the "par rate" set by the mortgage originator. In those cases, the broker decisions led to racial price disparity. However, it is unclear exactly why the broker decisions led to these differences. The brokers could have considered customers' race directly and charged minorities higher prices or perhaps the 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 conclusions as

---

[48] See generally, for example, Kleinberg, et al, 108 AEA Papers and Proceedings 22 (cited in note 26) (applying this logic to a hypothetical algorithm to be used in college admissions; arguing that facially neutral variables like SAT score can be correlated with race for a variety of reasons, such as the ability to take a prep course; and generating a theorem showing that excluding race from consideration while leaving in variables correlated with race leads to less equitable outcomes).

[49] In our simulation, we focus on excluding group identities when fitting the prediction function. In the context of linear regression, using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. See also Indrė Žliobaitė and Bart Custers, *Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models*, 24 Artificial Intelligence & L 183 (2016), (proposing a procedure that uses the sensitive attribute in the training data but then producing a decision rule without it); Devin G. Pope and Justin R. Sydnor, *Implementing Anti-discrimination Policies in Statistical Profiling Models*, 3 Am Econ J 206 (2011) (making a similar proposal).

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.

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 decisionmakers themselves—in the context of algorithmic decision-making, we can observe many aspects of the decision and therefore 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.[50]

However, this transparency is constrained by the limits on interpretability of decision rules. Prior legal writing on algorithmic fairness often characterized algorithms as opaque and uninterpretable.[51] 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 judgments. Indeed, we highlight in Part IV a crucial way in which algorithms can be interpreted for the purposes of ex ante regulation.

---

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

[51] See, for example, Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 Chi Kent L Rev 3, 44 (2018) (discussing how consumers may find it difficult to protect themselves because "many learning algorithms are thought to be quite opaque").

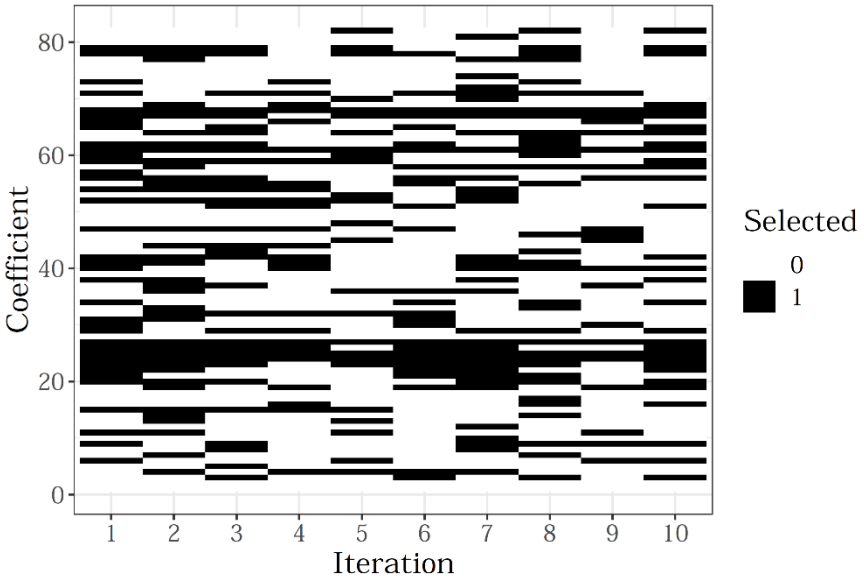
One potential way to interpret algorithmic decisions is to consider which variables are used by the algorithm, equivalent to interpreting coefficients in regression analysis. Typically, in social science research, the purpose of regression analysis is to interpret the coefficients of the independent 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. 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: 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 should produce similar outputs.

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 eighty dummy-encoded variables in our data set. 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 consistently included in the model, there are few discernible patterns, and an

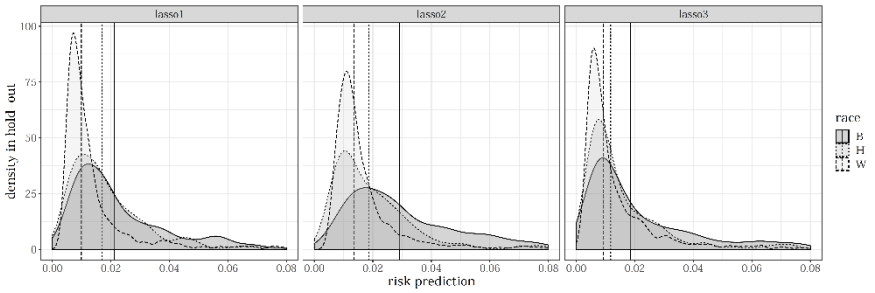
analysis of these prediction functions based on which variables were included would yield different conclusions from draw to draw, despite originating from similar data.

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 FIRST THREE LASSO PREDICTORS



The instability of the variables chosen for the prediction suggests that we should be skeptical about looking at the inclusion of certain variables to evaluate the process by which the decision rule was constructed and to determine the relationship between those variables and the ultimate decision. 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, many rules that look very different have qualitatively similar prediction properties. Which of these rules is chosen in a given draw of the data may come down to a flip of a coin. While these rules still differ in their predictions for some individuals, to the degree that we care only about the rules' overall prediction performance or overall pricing distributions, the specific representation of prediction functions may therefore not generally be a good description of relevant properties of the decisions.[52] In general, when data is high dimensional and complex machine-learning algorithms are used, a determination of conduct based on variable-importance measures is limited absent a specific notion of "importance" or interpretability that encapsulates the considerations relevant for the discrimination context.[53]

---

52 See Sendhil Mullainathan and Jann Spiess, *Machine Learning: An Applied Econometric Approach*, 31 J Econ Perspectives 87, 97 (2017) ("Similar predictions can be produced using very different variables. Which variables are actually chosen depends on the specific finite sample. . . . This problem is ubiquitous in machine learning.").

53 See Zachary C. Lipton, *The Mythos of Model Interpretability* *2 (arXiv.org, Mar 6, 2017), archived at http://perma.cc/3JQD-6CJ4 (highlighting that there is no common notion of "interpretability" for machine-learning models because the goals of interpretation differ); Leilani H. Gilpin, et al, *Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning* *3–5 (arXiv.org, Jun 4, 2018), archived at http://perma

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 the hope of recovering the causes of disparity and maybe even consideration of 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 of which receive only marginal weight. Therefore, legal rules that seek to identify the cause or root of disparate decisions cannot be easily applied.

Legal doctrines that put weight on identifying a particular conduct that caused disparity will not be able to rely on variable inclusion or general-purpose importance analysis alone. Courts have interpreted the FHA, the 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.[54] The Supreme Court recently affirmed the requirement for the identification of a particular policy that caused the disparity in *Inclusive Communities*, in which it said: "[A] disparate-impact claim that relies on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity."[55] In the mortgage lending cases we discuss 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.[56] 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 differential, in accordance with the conduct identification requirement of the discrimination

.cc/B789-LZBL (critically reviewing attempts to explain machine-learning models and noting the diversity of their goals).

[54] See, for example, *Wards Cove Packing Co v Atonio*, 490 US 642, 657–58 (1989) (establishing a "specific causation requirement" for disparate impact claims under Title VII of the Civil Rights Act of 1964). See also 12 CFR Part 100.

[55] *Inclusive Communities*, 135 S Ct at 2523.

[56] For additional examples of challenged policies, see Robert G. Schwemm and Calvin Bradford, *Proving Disparate Impact in Fair Housing Cases after* Inclusive Communities, 19 NYU J Legis & Pub Pol 685, 718–60 (2016).

doctrine.[57] However, our example above demonstrates that such an analysis is questionable and unlikely to be appropriate as the algorithmic equivalent of identifying conduct absent context-specific importance and transparency measures that apply to the legal context.[58] Antidiscrimination doctrine should therefore move away from this type of abstract decision rule analysis as a central component of antidiscrimination law.

IV. IMPLIED PRICES AND OUTCOME-FOCUSED DISCRIMINATION

In Part III we argue that, despite its purported transparency, the analysis of machine-learning pricing is constrained by limits to the interpretability of abstract pricing rules. In this Part, 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 an ex ante form of regulation that we call "discrimination stress testing," which exploits the opportunity that automated decision rules can be evaluated before they are applied to actual consumers. Our focus is on how to evaluate whether a pricing rule is fair, not on how to construct a fair pricing rule.

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 setting the final lending terms, pricing outcomes are not known until the actual prices have materialized for actual consumers. When pricing is automated, however, we also have information about pricing, even before customers receive loans, from inspecting the pricing rule. Furthermore,

---

[57] Although automated credit systems have been challenged in court, court decisions rarely provide guidance on this question. For example, in *Beaulialice v Federal Home Loan Mortgage Corp*, 2007 WL 744646, *4 (MD Fla), the plaintiff challenged the automated system used to determine her eligibility for a mortgage. Although the defendant's motion for summary judgment was granted, the basis for the decision was not that the plaintiff had not demonstrated conduct for a plausible claim of disparate impact. Alternatively, if the mere decision to use an algorithm is the "conduct" that caused discrimination, the requirement will be devoid 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. This analysis is the outcome analysis the next Section discusses.

[58] 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 and Selbst, 104 Cal L Rev at 677–93 (cited in note 27); Hurley and Adebayo, 18 Yale J L & Tech at 173 (cited in note 27).

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 based on algorithmic decision-making 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.[59] 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 relevant statistical test to use. As a result, output analysis in discrimination cases often focuses on simple comparisons and regression specifications[60] and then moves 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 led to a disparate outcome.

In the case of machine learning, we argue that outcome analysis becomes central to the application of antidiscrimination law. As Parts II and III discuss, both input regulation and decision process scrutiny are limited in the context of machine-learning pricing. Crucial aspects of current antidiscrimination law that focus on the procedure of creating the eventual prices are limited by the

---

[59] See Schwemm and Bradford, 19 NYU J Legis & Pub Pol at 690–92 (cited in note 56) (arguing that neither HUD Regulation 12 CFR Part 11 nor *Inclusive Communities*, both of which endorse discriminatory effects claims under the FHA, provide any guidance on how to establish differential pricing for a prima facie case of discrimination and showing that lower courts rarely followed the methodology established under Title VII).

[60] See Ayres, Klein, and West, *The Rise and (Potential) Fall of Disparate Impact Lending Litigation* at 236 (cited in note 14) (analyzing *In re Wells Fargo Mortgage Lending Discrimination Litigation*, 2011 WL 8960474 (ND Cal), in which plaintiffs used regression analysis to prove unjustified disparate impacts, as an example of how plaintiffs typically proceed).

difficulties of interpreting the decision rule that leads to the disparity and the challenges of closely regulating data inputs. Therefore, antidiscrimination law will need to increase its focus on outcome analysis in the context of machine-learning credit pricing.[61]

The type of ex ante analysis that 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,[62] stress tests require certain banks to report their stability under hypothetical financial scenarios.[63] 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 tests help to determine whether the bank should increase its capital and provide a general assessment of the bank's resilience. This allows for a form of regulation that is forward-looking and provides a consistent estimate across banks.[64] In a discrimination stress test, the regulator would apply the pricing rule of the lender to some hypothetical population before the lender implements the rule to evaluate whether the pricing meets some criteria of disparity.[65]

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 Essay. Instead, we highlight two main concerns in the development the discrimination stress test. First, we demonstrate the significance of selecting a particular population

---

61 Professor Pauline Kim argues that, in the algorithmic context, employers should be allowed to rely on the "bottom-line defense," thereby recognizing an increased role for outcome-based analysis in this context. See Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm & Mary L Rev 857, 923 (2016).

62 Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub L No 111-203, 124 Stat 1376 (2010), codified at 12 USC § 1503 et seq.

63 12 USC § 5365.

64 See Michael S. Barr, Howell E. Jackson, and Margaret E. Tahyar, *Financial Regulation: Law and Policy* 313 (Foundation 2016).

65 The power to announce future regulatory intent already exists within the CFPB's regulatory toolkit in the form of a No-Action Letter, through 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. See Consumer Financial Protection Bureau, *No-Action Letter Issued to Upstart Network* (Sept 14, 2017), archived at http://perma.cc/N4SU-2PRS.

to which the pricing rule is applied. Second, we discuss the importance of the particular statistical test used to evaluate pricing disparity.

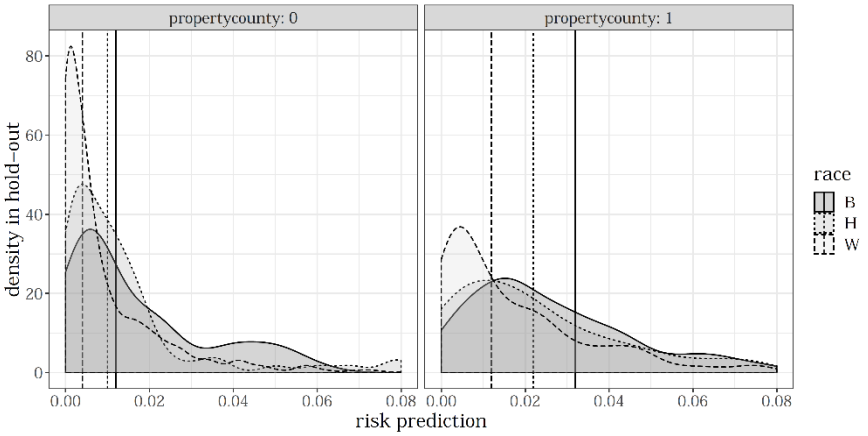
A. Population Selection

The first aspect of the discrimination stress test that we 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 be conducted only when applying the rule to a particular population. Therefore, regulators and policymakers need to determine what population to use when applying a forward-looking test.

The decision of which 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, such as 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 either a very similar (left) or quite different (right) distribution of predictions by group.

FIGURE 4: RISK PREDICTIONS FROM THE SAME PREDICTION FUNCTION ACROSS DIFFERENT NEIGHBORHOODS



The sensitivity of outcomes to the selected population highlights two important considerations for policymakers. 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 one that represents more vulnerable lenders.[66] Furthermore, regulators should select the sample population based on specific regulatory goals. If, for example, regulators seek to understand the impact of a pricing rule on the specific communities in which a lender operates, regulators may select a sample population of those communities rather than a nationally representative sample. 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 and 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

[66] 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, 88 U Colo L Rev at 1124–28 (cited in note 7).

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 institutions' ability to game the specifics of the stress test.[67] Similarly, for discrimination stress testing, the exact data set used could be kept confidential so that lenders are not able to create decision rules that minimize disparity for the specific data set alone.

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 the HMDA, other forms of lending do not have an equivalent requirement. This creates significant challenges for private and public enforcement of the 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. Because 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.[68]

---

[67] See Barr, Jackson, and Tahyar, *Financial Regulation: Law and Policy* at 313 (cited in note 63) ("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.").

[68] See, for example, Corbett-Davies, et al, *Algorithmic Decision Making* at *2 (cited in note 17). See also generally Feldman, et al, *Certifying and Removing Disparate Impact* at *2–3 (cited in note 27). 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 Cumb L Rev 67, 89–102 (2017).

The first aspect of a disparity test 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 that may correlate with group membership should be controlled for when comparing between groups. When controlling for these characteristics in disparity testing, only individuals that share these characteristics are compared.[69] Courts consider this issue by asking whether "similarly situated" people from the protected and nonprotected group were treated differently.[70] Suppose that individuals with the same income, credit score, and job tenure are considered "similarly situated." The distribution of prices is then allowed to vary across groups provided that this variation represents only variation with respect to those characteristics that define "similarly situated" individuals.[71] 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.[72]

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.[73] Despite the importance of this question, there is little

---

69 A formalization of this idea appears in Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao, *On Conditional Parity as a Notion of Non-discrimination in Machine Learning*, (arXiv.org, Jun 26, 2017), archived at http://perma.cc/A92T-RGZW. They argue that the main notions of nondiscrimination are a form of conditional parity.

70 See *BP Energy Co v Federal Energy Regulatory Commission*, 828 F3d 959, 967 (DC Cir 2016). This requirement has also been referred to as the requirement that demonstration of disparate impact focus on "appropriate comparison groups." See Schwemm and Bradford, 19 NYU J Legis & Pub Pol at 698 (cited in note 56). See also Jennifer L. Peresie, *Toward a Coherent Test for Disparate Impact Discrimination*, 84 Ind L J 773, 776–79 (2009).

71 See Cynthia Dwork, et al, *Fairness through Awareness*, Proceedings of the Third Innovations in Theoretical Computer Science Conference 214, 215 (2012) (providing a 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).

72 See generally Robert Bartlett, et al, *Consumer-Lending Discrimination in the Era of FinTech*, *1 (UC Berkeley Public Law Research Paper, Oct 2018), archived at http://perma.cc/6BM8-DHVR. In that paper, varying credit prices and rejection rates for ethnic groups are decomposed into effects driven by "life-cycle variables" and ethnic disparities that are not driven by these variables and therefore, according to the authors, discriminatory.

73 Professor Ian 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 & Med S68, S69–70 (2005).

guidance in cases and regulatory documents on which characteristics make people similarly situated.[74] An approach that considers what is predictive of being similarly situated 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 lawmakers 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 the scope of this Essay, we highlight two considerations. First, as we argue throughout the Essay, 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, its inclusion in the algorithm may 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 used to compare groups, the regulator requires a test in order to determine whether there is indeed disparity.[75] Typically, for such a test, the regulator needs to

---

[74] One exception is the 1994 Policy Statement on Discrimination in Lending by HUD, the Department of Justice, and other agencies, which suggested that the characteristics listed in the HMDA do not constitute an exhaustive list of the variables that make people similarly situated. Department of Housing and Urban Development, Interagency Policy Statement on Discrimination in Lending, 59 Fed Reg 18267 (1994).

[75] The algorithmic fairness literature includes many different tests, some of which are summarized by MacCarthy, 48 Cumb L Rev at 86–89 (cited in note 67). One of the only examples of an articulated statistical test is the "four-fifths rule" adopted by the Equal Employment Opportunity Commission in 1979. 29 CFR § 1607.4(D):

> A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.

fix a tolerance level that expresses how much the distribution of risk predictions may deviate across groups between similarly situated individuals.

CONCLUSION

In this Essay, 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 be bridged only by resolving normative legal questions. These questions have thus far received little attention because they were of less practical importance in a world in which antidiscrimination 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 antidiscrimination doctrine. Simply restricting an algorithm from using specific information, for example, would at best satisfy a narrow reading of existing legal requirements and would 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. Because prices are set by fixed rules, discrimination stress tests are opportunities to check pricing outcomes in a controlled environment. Such tests can draw on 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 antidiscrimination doctrine, not only because the new reality requires an adaptation of an anachronistic set of rules but because philosophical disagreements over the scope of antidiscrimination law now have practical and pressing relevance.

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, the extent to which this test is binding is not clear given the tendency of courts to overlook it. For further discussion, see Schwemm and Bradford, 19 NYU J Legis & Pub Pol at 706–07 (cited in note 56). For an application of this test in the algorithmic fairness literature, see generally Feldman, et al, *Certifying and Removing Disparate Impact* (cited in note 27).