February 18, 2020

# FALSE DREAMS OF ALGORITHMIC FAIRNESS: THE CASE OF CREDIT PRICING

*Talia Gillis* [*]

[*] Empirical Law and Finance Fellow and SJD Candidate, Harvard Law School. PhD Candidate in Business Economics, Department of Economics and Harvard Business School. For comments and suggestions on preliminary drafts, I thank Oren Bar-Gill, John Beshears, Doni Bloomfield, Sarith Felber, Noah Feldman, Jesse Fried, Becca Goldstein, Jon Gould, Deborah Hellman, Howell Jackson, David Laibson, Ela Leshem, Natasha Sarin, Joshua Simons, Holger Spamann, Jann Spiess, Rafi Stern, Cass Sunstein, and Crystal Yang.

## ABSTRACT

Credit pricing is changing. Traditionally, lenders priced consumer credit by using a small set of borrower and loan characteristics, sometimes with the assistance of loan officers. Today, lenders increasingly use big data and advanced prediction technologies, such as machine-learning, to set the terms of credit. These modern underwriting practices could increase prices for protected groups, potentially giving rise to violations of anti-discrimination laws.

What is not new is the concern that personalized credit pricing relies on characteristics or inputs that reflect preexisting discrimination or disparities. Fair lending law has traditionally addressed this concern through input scrutiny, either by limiting the consideration of protected characteristics or by attempting to isolate inputs that cause disparities.

But input scrutiny is no longer effective. Using data on past mortgages, I simulate algorithmic credit pricing and demonstrate that input scrutiny fails to address discrimination concerns. The ubiquity of correlations in big data combined with the flexibility and complexity of machine-learning means that one cannot rule out the consideration of a protected characteristic even when formally excluded. Similarly, in the machine-learning context, it may be impossible to determine which inputs drive disparate outcomes.

Despite these fundamental changes, prominent approaches to applying discrimination law in the algorithmic age continue to embrace the input-centered approach of traditional law. These approaches suggest that we exclude protected characteristics and their proxies, limit algorithms to pre-approved inputs, and use statistical methods to neutralize the effect of protected characteristics. Using my simulation exercise, I demonstrate that these approaches fail on their own terms, are likely unfeasible, and overlook the benefits of accurate prediction.

I argue that the shortcomings of current approaches mean that fair lending law must make the necessary, though uncomfortable, shift to outcome-focused analysis. When it is no longer possible to scrutinize inputs, outcome analysis provides a way to evaluate whether a pricing method leads to impermissible disparities. This is true not only under the legal doctrine of disparate impact, which has always cared about outcomes, but also, under the doctrine of disparate treatment, which historically has avoided examining disparate outcomes. Now, disparate treatment too can no longer rely on input scrutiny and must be considered through the lens of outcomes. I propose a new framework that regulatory agencies, such as the Consumer Financial Protection Bureau, can adopt to measure the disparities created when moving to an algorithmic world, enabling an explicit analysis of the tradeoff between prediction accuracy and other policy goals.

**INTRODUCTION** .......... **4**

**I. PRICING CREDIT BASED ON BIASED INPUTS** .......... **14**

1.1 THE CREDIT PRICING DECISION .......... 14
1.2 THE PROBLEM OF BIASED INPUTS .......... 17
*1.2.1 Biased world* .......... *17*
*1.2.2 Biased measurement* .......... *19*
1.3 TRADITIONAL FAIR LENDING LAW .......... 21

**II. THE CHANGING WORLD OF CREDIT LENDING** .......... **28**

2.1 WHAT IS CHANGING? .......... 29
*2.1.1 Nontraditional data* .......... *30*
*2.1.2 Advance prediction technologies* .......... *32*
*2.1.3 Automation* .......... *33*
2.2 SIMULATION EXERCISE – HYPOTHETICAL “NEW WORLD” CREDIT LENDER .......... 34
2.3 WHAT ARE THE CHALLENGES IN THE ALGORITHMIC CONTEXT? .......... 37
*2.3.1 Biased world inputs in algorithmic pricing* .......... *37*
*2.3.2. Biased measurement inputs in algorithmic pricing* .......... *40*

**III. APPROACHES TO ALGORITHMIC DISCRIMINATION** .......... **41**

3.1 EXCLUDING PROTECTED CHARACTERISTICS .......... 44
*3.1.1 Ineffective exclusion* .......... *45*
*3.1.2 Disparities may increase with exclusion* .......... *49*
3.2 EXCLUDING PROXIES FOR PROTECTED CHARACTERISTICS .......... 52
*3.2.1 What is a “proxy”?* .......... *53*
*3.2.2 Identifying proxies* .......... *55*
3.3 RESTRICTING ALGORITHM TO PREDETERMINED SET OF VARIABLES .......... 58
*3.3.1 Entrenching disadvantage* .......... *59*
*3.3.2 High cost to prediction accuracy* .......... *60*
3.4 ORTHOGONALIZING INPUTS .......... 63
*3.4.1 Instability of machine learning selection* .......... *64*
*3.4.2 Why machine learning is different* .......... *68*
3.5 REQUIRED SHIFT FROM CAUSATION TO CORRELATION .......... 68

**IV. TOWARD A SOLUTION** .......... **71**

4.1 TESTING OUTCOMES IN THREE STEPS .......... 72
*4.1.1 Comparing borrowers who are similarly situated* .......... *75*
*4.1.2 Considering incremental change* .......... *77*
4.2 REGTECH RESPONSE TO FINTECH .......... 79

**CONCLUSION** .......... **80**

**APPENDICES** .......... **82**

## INTRODUCTION

Many important decisions made primarily by humans in the past are now being automated using advance prediction technologies and big data. Algorithms are being used in a wide range of domains, from screening resumes[1] to determining criminal justice outcomes.[2] In consumer credit, there is a move towards reliance on algorithms to predict creditworthiness and to price credit accordingly. Credit pricing increasingly uses nontraditional data[3] and machine learning algorithms,[4] while decreasing its reliance on human decision-makers and a small set of creditworthiness indicators, such as FICO scores.[5]

Despite the efficiency and accuracy gained via these technologies, there has been increasing concern that machine learning algorithms may be biased or lead to unfair outcomes.[6] Bias, a general term loosely used to describe conduct or an outcome that is unfair to a vulnerable population or legally

[1] *See* Josh Bersin, *Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age*, FORBES (Feb. 17, 2013, 8:00 PM), http://www.forbes.com/sites/joshbersin /2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age/; Matt Richtel, *How Big Data Is Playing Recruiter for Specialized Workers*, N.Y. Times (Apr. 27, 2013), https://www.nytimes.com /2013/04/28 /technology/how-big-data-is-playing-recruiter-for-specialized-workers.html.

[2] *See* Ed Young, *A Popular Algorithm Is No Better at Predicting Crimes Than Random People*, ATLANTIC (Jan. 17, 2018), https://www.theatlantic.com/technology/archive/2018 /01/equivant-compas-algorithm/550646.

[3] *See infra* Section 2.1.1. *See also* Bureau of Consumer Financial Protection, Fair Lending Report of the Bureau of Consumer Financial Protection, June 2019, 84 C.F.R. 32420 (July 8, 2019) (describing the recent CFPB Report from June 28, 2019 on a symposium the Bureau held in which they "discussed the role of alternative data and modeling techniques can play in expanding access to traditional credit").

[4] *See infra* Section 2.1.3.

[5] FICO, previously known as Fair, Isaac & Co., "has developed a sophisticated algorithms for generating credit scores that characterized consumer financial creditworthiness" to predict whether consumers would default on their debt. FICO's credit score is based on information in the consumer credit report received from credit bureaus. *See* Fair Isaac Corp. v. Experian Information Solutions, Inc., 650 F.3d 1139 (8th Cir. 2011).

[6] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Matthew Adam Bruckner, *The Promise and Perils of Algorithmic Lenders' Use of Big Data*, 93 CHI.-KENT L. REV. 3, 25–29 (2018); Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 168 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2016–2017); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L. J. 2251, 2233 (2019); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 402 (2018). *See also* Megan Smith et al., *Big Risks, Big Opportunities: The Intersection of Big Data and Civil Rights*, WHITEHOUSE.GOV (2019), https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights.

protected group,[7] can occur in algorithms for several reasons. An algorithm could be trained with nonrepresentative data,[8] it could be set up to predict a human decision that is biased,[9] or it could have imperfect measures of the outcome of interest.[10] One type of concern, particularly important in the credit context and therefore the focus of this paper, is the use of characteristics or "inputs" that are biased because they reflect some preexisting disadvantage or because they are a noisy or biased measurement of borrower characteristics.[11]

On August 19, 2019, the Department of Housing and Urban Development (HUD) published its proposal to replace its rule on the implementation of the Fair Housing Act from 2013.[12] HUD's Proposed Rule on the Implementation

[7] *See* Mayson, *supra* note 7, at 2231 (discussing the ambiguity of the term "bias"). Often the language used to define "bias" is quite circular. *See, e.g.*, Kim, *supra* note 7, at 887 ("Similarly, data mining models built using biased, error- ridden, or unrepresentative data may be statistically biased").

[8] *See* Hurley & Adebayo, *supra* note 7, at 178 ("If credit scorers rely on non-neutral data collection tools that fail to capture a representative sample of all groups, some groups could ultimately be treated less favorably or ignored by the scorers final model"). It could also be that the dataset is simply flawed. For example, the Federal Trade Commission found that 21% of its sample of consumers had a confirmed error on at least one of three credit bureau reports. *See* FED. TRADE COMM'N, REPORT ON BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? (Jan. 2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf. This is of particular concern if certain groups, such as racial minorities, are more likely to have errors in their files. This is likely what happened when Amazon used AI to recruit workers, given that past hiring was predominantly male. *See* Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women* (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[9] *See, e.g.*, Bruckner, *supra* note 7, at 26 (discussing an example in which an algorithm was set up to predict admissions decision using a training set that was created by biased admission officers).

[10] This type of concern could arise when the outcome, or "label," is a noisy measurement of the true outcome of interest. *See, e.g.*, Mayson, *supra* note 7, at 2227 (arguing that past crime data is distorted relative to actual crime rates). Another concern arises when outcomes are only observed for a sub-group depending on an earlier decision that might itself be biased. This is often referred to as the "selective labels problem," and it is of particular concern in the credit context in which borrower default is only observed if they received a loan. *See* Himabindu Lakkaraju et al., *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*, PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 275 (KDD '17, ACM New York, NY, USA 2017) (developing a method to overcome the problem of selective labels).

[11] *See infra* Section 1.2.

[12] Department of Housing and Urban Development, Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460 (Feb. 15, 2013) (promulgating at 24 C.F.R. § 100.500) [hereinafter HUD Disparate Impact Rule 2013]. *See* Aaron Glantz

of the Fair Housing Act's Disparate Impact Standard[13] discusses for the first time how to determine whether an algorithm violates fair lending law. In fact, the Proposed Rule is one of the first attempts in the United States and worldwide to articulate discrimination law as applied to an algorithm. A crucial focus of the rule is how to scrutinize and justify the "inputs" into a lender's algorithm. Despite the Proposed Rule's attempt to facilitate "practical business choices . . . that sustain a vibrant and dynamic free-enterprise system,"[14] the proposed rule is confused and contradictory and reflects a lack of basic understanding of the technology at play.[15] These shortcomings suggest that fair lending law is likely to become a central battleground on which practitioners and scholars will argue over the application of discrimination law to algorithmic decision-making.

In this Article, I make two arguments. First, I argue that the leading approaches to algorithmic discrimination are misguided, even on their own terms. These solutions hold onto the input-centric view of traditional fair lending even as machine learning pricing makes this view obsolete and irrelevant. Second, I argue that when we recover from the overhang on input analysis we should explore ways to expand and emphasize regulatory output analysis through empirical testing of algorithmic outcomes.

Throughout the Article I use a simulation exercise in which a hypothetical lender analyzes past loans to make predictions about future borrowers. For this exercise, I combine the Boston Fed Home Mortgage Disclosure Act (HMDA) dataset, which contains information on mortgage applications, with simulated default rates disciplined by information on the loans.[16] My hypothetical lender uses a machine learning algorithm to predict default probability, which is then used to price credit for future borrowers. In this

and Emmanuel Martinez, *Can Algorithms Be Racist? Trump's Housing Department Says No*, REVEAL (Aug. 5, 2019), https://www.revealnews.org/article/can-algorithms-be-racist-trumps-housing-department-says-no/ (containing a link to the circulated draft: https://www.documentcloud.org/documents/6239364-Algorithm.html); Hannah Lang, *HUD Plan Would Raise Bar for Claims of Fair-lending Abuse*, NAT'L MORTGAGE NEWS (July 31, 2019) https://www.nationalmortgagenews.com/news/hud-plan-would-raise-bar-for-claims-of-fair-lending-abuse.

[13] Department of Housing and Urban Development, Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42854 (Aug. 19, 2019) [hereinafter HUD Proposed Rule 2019].

[14] HUD Proposed Rule 2019, quoting Texas Dep't of Hous. and Cmty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507 (U.S. 2015).

[15] *See infra* Section 3.5.

[16] As explained in Section 2.2 and Appendix A, I fit a model that predicts whether an application is denied or rejected and then calibrate the rejection rates to publicly available statistics on default. Therefore, to the extent that there is some relation between a lending decision and borrower default, these simulated default rates may capture some of the relation between real-world default and borrower characteristics.

simulation exercise, the loan and borrower characteristics serve as the "inputs" to the credit decisions, while the predicted default probability is the "output."

Beginning with the question of why we might be concerned with algorithmic pricing and how it differs from traditional credit pricing, I provide structure and clarity to the discussion on algorithmic bias by distinguishing among different types of input biases. The current focus on algorithms and biases overlooks the fact that even traditional credit pricing relied on borrower characteristics that reflected preexisting disadvantage ("biased world" inputs)[17] or were inaccurately measured ("biased measurement" inputs).[18] In the algorithmic context, the use of biased inputs could increase disparities in some instances while decreasing them in others. On the one hand, differences in inputs could translate into greater variance in outcomes when using machine learning to predict default.[19] This means that the use of machine learning could further entrench pre-existing disparities to a larger extent than was likely under more basic statistical methods. On the other hand, the use of machine learning and big data could also mitigate some of the harm of "biased measurement" inputs if more information is available on individual borrowers.[20] Either type of input creates a concern that information about a protected characteristic is embedded in other information

[17] *See infra* Section 1.2.1. Credit pricing has always considered borrower characteristics that are likely to partially reflect pre-existing disadvantage or discrimination. For example, if women suffer discrimination in the labor market their income and debt-to-income ratios are "biased-world" inputs.

[18] *See infra* Section 1.2.1 If, for example, credit scores only consider certain types of creditworthiness indicators, such as timely loan payments but do not consider timely rent payments, and those indicators are less likely to be available for racial minority borrowers, then credit scores are a "biased measurement" input of creditworthiness. *See* Board of Governors of the Fed. Reserve Sys., *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit* (2007) (finding that recent immigrants have lower credit scores than implied by loan performance and recommending that the type of information supplied to credit-reporting agencies to include routine payments such as rent be expanded)

[19] *See infra* Section 2.3. One way to consider the change in the machine-learning setting is by comparing a linear regression to a machine-learning algorithm. As will be discussed further, the flexibility of machine-learning algorithms allows for a greater ability to differentiate among people based on their characteristics than on less flexible methods, such as an OLS regression. I demonstrate this using the Boston Fed HMDA dataset. *See also* Andreas Fuster et al., *Predictably Unequal? The Effects of Machine Learning on Credit Markets* 10 (2018), *available at* https://papers.ssrn.com/abstract=3072038 (presenting a useful demonstration of how a non-linear prediction allows for greater flexibility in differentiating between people based on their characteristics).

[20] *See infra* Section 2.3.2. One of the examples discussed in that Section is when a lender has information about a borrower's timely rent payments, it may not matter whether the borrower's FICO score does not take this information into account.

about the individuals.

Fair lending law is the primary lens to determine whether disparities created by biased inputs amount to discrimination in traditional credit pricing. Fair lending covers both the doctrine of disparate treatment, dealing with intentional discrimination, and the doctrine of disparate impact, dealing with a facially neutral rule that creates impermissible disparities.[21] Despite the ongoing disagreements over the boundaries and philosophical foundations of fair lending law, discussed but not resolved in this Article, the focus has been on scrutinizing decision inputs to determine whether lender pricing amounts to discrimination.[22] This has also been true of disparate impact, which, despite its name, has primarily focused on identifying and justifying inputs and policies that drive disparities.[23] Given that fair lending developed in the traditional credit pricing setting, in which pricing was based on few inputs and involved human discretion, there is a need to adapt the law to the algorithmic context.[24]

The second focus of this Article is to challenge some of the leading approaches in applying discrimination law to the algorithmic context.[25] These approaches are inadequate primarily on their own terms, because they are unable to satisfy their own loose definition of fairness. The approaches treat formal exclusion or inclusion of inputs as meaningful, when in fact in high dimensional data where correlations are ubiquitous, such exclusions have little or no effect. They are also undesirable for several other reasons. These approaches often cannot be practically implemented and are unsuitable for the machine learning setting. Furthermore, at times these approaches could restrict access to credit for vulnerable populations and further entrench disadvantage.

[21] *See infra* Section 1.4.

[22] *See infra* Section 1.4 and Section 3.5. For disparate treatment, the legal doctrine concerned with intentional discrimination, the central question is whether a borrower's protected characteristic played a role in setting the price and thereby served as an "input" in the decision. The legal doctrine of disparate impact, which is concerned with facially neutral policies that have an impermissible effect, also focuses on analyzing decision inputs after an initial demonstration of the outcome disparities. Disparate impact requires isolating the input that caused the disparity, and such an input is nevertheless permissible if it is related to a legitimate business justification. Therefore, as discussed in further detail below, although the prima facie case of disparate impact requires a showing of disparities, the analysis revolves around the cause of the disparities.

[23] *See infra* Section 1.4.

[24] *See* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019); Hurley & Adebayo, *supra* note 7, at 183.

This is also true of other areas of discrimination law. *See* Kim, *supra* note 7. *See generally* Barocas & Selbst, *supra* note 7, at 694.

[25] Many of these proposals are not only intended to apply to fair lending, but also have a direct bearing on how discrimination law would apply in algorithmic credit pricing.

I discuss and criticize four leading approaches. The first approach is the exclusion of protected characteristics, primarily as a method for negating a claim of intentional discrimination. This type of defense was used by Goldman Sachs in November 2019. In response to a claim that a man received a credit line twenty times higher than his wife,[26] Goldman Sachs argued that it was not possible for them to discriminate, as they do not make decisions "based on factors like gender" and that they "do not know your gender."[27]

The challenge, however, is that information about a person's protected characteristic is embedded in other information about the individual, so that a protected characteristic can be "known" to an algorithm even when it is formally excluded. I demonstrate this by predicting "age" and "marital status," two protected characteristics under fair lending law,[28] from the other variables within the HMDA dataset.[29]

There are several reasons we should be concerned about the ability to predict protected characteristics from other data. Consider an algorithmic lender who is required to comply with the Equal Credit Opportunity Act (ECOA) and cannot discriminate against borrowers based on their age.[30]

[26] Neil Vigdor, *Apple Card Investigated After Gender Discrimination Complaints*, NY TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.

[27] Shahien Nasiripour, Jennifer Surance, and Sridhar Natarajan, *Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks*, BLOOMBERG BUSINESSWEEK (Nov. 11, 2019, 3:01 PM), https://www.bloomberg.com/news/articles/2019-11-11/apple-card-s-ai-stumble-looks-familiar-to-old-school-banks.

[28] See 15 U.S.C. §1691(a) ("It shall be unlawful for any creditor to discriminate against any applicant, with respect to any aspect of a credit transaction – on the basis of … marital status, or age (provided the applicant has the capacity to contract).").

[29] The ability to predict "marital status" and "age" using the Boston Fed HMDA dataset is likely to be the lower bound on the ability to predict protected characteristics in the algorithmic context. This is because HMDA primarily contains traditional credit pricing variables, unlike "nontraditional" data discussed in Section 2.1.1.

[30] The requirement to not consider "age" under ECOA is more complex than would seem based on the text of ECOA alone. Regulation B contains specific provisions related to age. *See* The Consumer Financial Protection Bureau, Regulation B, 12 C.F.R. §1002.6(b)(2)). Whether and how a creditor can use age in a credit decision depends on the system used. According to §1002.6(b)(2)(ii), when using "an empirically derived, demonstrably and statistically sound, credit scoring system, a creditor may use an applicant's age as a predictive variable, provided that the age of an elderly applicant is not assigned a negative factor or value." Assuming algorithmic credit pricing meets the criteria of a "demonstrably and statistically sound" scoring system as defined in §1002.2(p), it is unclear how a lender using an algorithm will ever be able to show that they have met the requirement that "applicants age 62 years or older must be treated at least as favorably as applicants who are under age 62." (see Supplement I to Part 1002- Official Interpretation). This is because with algorithmic pricing, unlike expert based scoring, the weights are not pre-assigned to different characteristics. Similarly, as discussed in more detail in Section 3.4, one must be wary of interpreting the weight on "age" as the true and stable contribution of that variable to a

ECOA requires that lenders not directly or intentionally discriminate against an older borrower, or use a neutral rule that has a disproportionate effect on older borrowers. The lender is aware, however, that older borrowers are different from other borrowers. They often have less documented credit history and tend to use cash more frequently.[31] There is also a mechanical effect of age on default. An older person is less likely to live long enough to repay their loan before dying. The lender does not formally consider age but instead applies a machine learning algorithm to predict borrower default risk from Amazon purchase history. Given the close relationship between age and default risk, an algorithm may be particularly motivated to recover a borrower's age even when it is excluded from the algorithm. Even a basic machine learning algorithm would be able to ultimately consider a borrower's age, rendering this exclusion meaningless.

We should also be wary of excluding protected characteristics if we care about outcome disparities.[32] As I demonstrate through a simulated example, price disparities could decrease when algorithms are "race aware." This is because a characteristic may need to be interpreted differently for various racial group. However, when we exclude the race variable, we are imposing a similar interpretation of a characteristic for both white and non-white applicants, which may increase disparities.

The second approach I discuss expands the exclusion of inputs to proxies for protected characteristics. This approach recognizes that other inputs may act as "proxies" for protected characteristics and therefore should be excluded too. The approach, however, is not feasible when there is no agreed-upon definition of a proxy, and when complex interactions between variables are unidentifiable to the human eye. Even inputs that have traditionally been thought of as proxies for race, such as zip codes, may be less concerning than other ways in which a borrower's race can be recovered. Using the HMDA data, I demonstrate that there is a greater ability to predict "race" from the

prediction. *See* Kathryn P. Taylor, *Equal Credit for All – An Analysis of the 1976 Amendments to the Equal Credit Opportunity Act*, 22 ST. LOUIS U. L.J. 326, 338 (1978–1979) ("The Amendments set limits on the use of age in credit scoring systems, and prohibit the assignment of a negative value to the age of an elderly applicant"). I therefore conclude that it is unlikely that algorithmic credit pricing can consider age under current regulations.

[31] See Mary Jane Large, *The Credit Decision and Its Aftermath*, BANKING L.J. 4, 20–22 (1980) (discussing the background to the enactment of ECOA and the prohibition of discrimination based on age).

[32] As discussed further in Part III, the exclusion of protected characteristics may be considered a fair procedure, regardless on its impact on disparities. This question closely relates to the more general debate on procedural versus substantive justice. *See generally* Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181 (2004–2005). What is particularly striking about this context is the extent to which the formal exclusion of the characteristic is unlikely to mean the characteristic was not considered, regardless of the raw disparities among groups.

traditional credit pricing inputs in HMDA than from zip codes. Similarly, although it may be possible for example to require lenders to exclude clear proxies for age from datasets, borrower age can still be revealed by the combination of many consumer behaviors.

Despite the shortcomings of this approach, the Trump administration is promoting this very interpretation of disparate impact. In HUD's circulated draft of its Proposed Rule from August 2019, a lender can defend an algorithm by demonstrating that it does "not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act."[33] The proposed rule does not provide any guidance on how to identify a "proxy" or "substitute." This type of vague standard renders this approach impractical and unlikely to address the concern that an algorithm creates impermissible distinctions.

The third approach I discuss takes the reverse perspective of restricting algorithm inputs to pre-approved inputs, unlike the first two approaches, which allow all inputs other than certain forbidden inputs. Although this approach may allow for greater control over what algorithms use to price credit, the approach could ultimately restrict access to credit. This is because limiting an algorithm to traditional credit pricing inputs further perpetuates the exclusion of consumers lacking formal credit histories and could thus entrench disadvantage without the ability of big data to undo or mitigate the harm from "biased measurement" inputs. Moreover, using my simulation exercise, I demonstrate that the prediction of default based on fewer inputs decreases prediction accuracy. When lenders have less ability to differentiate among borrowers based on their risk, lenders are limited in their ability to price the lending risk.

The last approach I discuss, the orthogonalization approach, is based on a statistical method meant to prevent inputs that correlate with protected characteristics from serving as proxies. This is achieved by a statistical method, described in greater detail below,[34] in which a protected characteristic is used to *estimate* a prediction function but not to *apply* the prediction.[35] I argue that this approach goes wrong in the machine learning setting because it essentially involves lying to the algorithm. I use a technical demonstration to show how this approach is inappropriate for the machine learning context in which the sole purpose of the algorithm is to optimize prediction, and not estimate a model.

[33] See Implementation of the Fair Housing Act′s Disparate Impact Standard, 84 Fed. Reg. 42854 § 100 (U.S. Aug. 19, 2019), at page 33.

[34] *See infra* Section 3.4.

[35] *See* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. L. ANALYSIS (2018), for a discussion of the difference between the training of an algorithm and the application of the prediction (or the "screener").

These four approaches are unsuitable because they continue to scrutinize decision inputs, similar to traditional fair lending, when this strategy is no longer feasible or effective in the algorithmic context. They remain centered around the two causal questions on which traditional fair lending has focused: First, whether a protected characteristic had a causal effect on the credit decision (disparate treatment), and second, whether the inputs into credit decisions caused impermissible disparities (disparate impact).[36] However, machine learning is a world of correlation and not causation. When using a machine learning algorithm to predict an outcome, the focus is on the accuracy of the prediction.

I argue that the shortcomings of current approaches mean that fair lending law must make the necessary, yet uncomfortable, shift to outcome-focused analysis.[37] When it is no longer possible to scrutinize inputs, outcome analysis provides a way to evaluate whether a pricing method leads to impermissible disparities. This is true not only for the legal doctrine of disparate impact, which has always cared about outcomes even when it did so by scrutinizing inputs. Surprisingly, even disparate treatment, a doctrine that historically has been quite detached from disparate outcomes, can no longer rely on input scrutiny and must be considered through the lens of outcomes.

I end the Article by discussing possible paths forward. I argue that regulators should develop a framework for an ex ante consideration of the effects of an algorithmic pricing rule. This can be achieved by applying a credit pricing rule, before it is used by a lender, to a dataset of hypothetical lenders. The regulator can then examine the outcomes of the pricing rule to determine whether the pricing rule discriminates. This type of outcome-focused testing brings to the forefront the demonstration of disparities, which is formally part of the first stage of a disparate impact complaint in traditional fair lending law. My proposed testing framework develops this type of analysis and adapts it to the machine learning context.

The criteria to determine when disparities amount to discrimination depend heavily on the boundaries and interpretation of the doctrine of anti-discrimination, which continues to be disputed. Therefore, I do not provide an exact test, but argue that algorithmic outcomes can be used to answer

[36] *See e.g.*, Sheila R. Foster, *Causation in Antidiscrimination Law: Beyond Intent versus Impact*, HOUS. L. REV. 1469, 1472 (2004–2005) ("By definition, all discrimination claims require plaintiffs to demonstrate a causal connection between the challenged decision or outcome and a protected status characteristic"). This is further developed in Section 3.5.

[37] Some previous writing on discrimination and artificial intelligence has suggested that greater focus should be placed on outcomes. *See e.g.*, Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 24 ( "The focus on outcomes rather than how an algorithm operates seems especially useful as algorithms become increasingly complicated, even able to modify themselves.").

meaningful questions. The first question that my outcome analysis can answer is whether the pricing rule treats borrowers who are "similarly situated" equally. This can also be thought of as a second best inquiry when input analysis is unable to provide an answer to whether a protected characteristic was used in pricing. The second question is whether the pricing rule increases or decreases disparities relative to some baseline, such as the non-algorithmic credit pricing method. In the final Section, I describe how outcome analysis can be used to answer these two questions, and I highlight challenges to the implementation of outcome-based testing for future work.

My outcome-based approach to discrimination testing reflects the need to adopt an empirical and experimental approach to discrimination. In the algorithmic world we can no longer determine a priori how inputs relate to outcomes. We do not know whether an algorithm is using a protected characteristic from observing the algorithm's inputs. Similarly, we cannot determine whether an algorithmic method will increase or decrease disparities based only on whether it uses nontraditional inputs. An outcome-based approach seeks to test the actual effects of a credit pricing method providing an appropriate Regtech response to the Fintech industry.

The significance of how to implement discrimination law to the algorithmic context transcends fair lending. My analysis speaks directly to other legal settings in which there are ongoing debates on how to define and implement discrimination to algorithms, from employment to criminal justice.[38] In these domains, like fair lending, there is a misplaced focus on input scrutiny. The Article also contributes to discussions in the computer science and statistical literature on algorithmic fairness, by demonstrating how evaluation of algorithmic decisions should be informed by legal doctrine, and regulatory and institutional realities.

The Article proceeds in four parts. Part I focuses on the traditional world of credit lending and presents the distinction between "biased world" inputs and "biased measurement" inputs. Part II turns to the new world of algorithmic credit pricing, describing the primary changes and their meaning for the problem of biased inputs. Part III discusses the main approaches to discrimination law in the algorithmic context and shows that they are inadequate on their own terms and also otherwise undesirable. Part IV argues that the move to algorithmic pricing requires more fundamental shifts in fair lending law from input scrutiny to outcome analysis and develops an

[38] See other papers that discuss the application of discrimination in other areas of law. *See* Kim, *supra* note 7. (employment discrimination); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, CLR ONLINE 164 (2017); Allan G. King & Marko J. Mrkonich, *Big Data and the Risk of Employment Discrimination*, OKLA. L. REV. 555 (2015–2016). *See generally* Chander, *supra* note 38.

appropriate framework.

## I. PRICING CREDIT BASED ON BIASED INPUTS

When pricing credit, lenders often offer people different loan terms based on their individual predicted default probability using borrower characteristics and the loan specifics. In this Part, I discuss how there is commercial and social value in accurately predicting default risk, which underlies differential pricing. However, when characteristics vary by group because they reflect bias, their use to price credit differentially may entrench bias. As I elaborate on in this Part, traditional discrimination law addresses this tension by either prohibiting the direct use of a protected characteristic or by limiting pricing policies that could further bias.

In this Part, I focus on traditional credit lending, before discussing algorithmic credit pricing, to highlight what is likely to change. This is important because current concerns over the fairness of credit pricing algorithms overlook the fact that even traditional credit pricing relied on borrower characteristics that reflected pre-existing disadvantage ("biased world" inputs) or were inaccurately measured ("biased measurement" inputs).

I begin this Part by providing an overview of a credit pricing decision that presents the terminology I will use throughout the Article. I then discuss the distinction between "biased world" and "biased measurement" inputs and how they effect a pricing decision. I end the Part by discussing how traditional fair lending law has dealt with the tension between personalized pricing that relies on biased inputs and the benefits of accurate default prediction.[39]

### *1.1 The credit pricing decision*

Credit contracts are often personalized,[40] meaning that lenders will

[39] There are other concerns that can arise in the context of credit pricing that I do not fully address. For example, a lender could intentionally deny credit to a member of a protected group, motivated by animus, what economists typically refer to as "taste-based discrimination." *See* GARY S. BECKER, THE ECONOMICS OF DISCRIMINATION (2010). I focus on the problem of biased inputs, because of the prevalence of biased inputs in lending decisions, and second, because the use of biased inputs creates an opportunity and challenge for algorithmic pricing, as will be discussed in Section 2.3.

[40] Not all credit is personalized and not all credit is personalized to the same extent. The personalization of credit contracts can be costly so that the degree of personalization may depend on the magnitude of the credit contract. For mortgages, which are typically large loan contracts, there is likely to be a degree of personalization. But this can also be true of smaller loans and other types of debt, such as auto loans.

determine the specific terms of the contract based on the characteristics of the borrower and the specific loan. We can therefore articulate the pricing decision as one in which inputs, $x$, are used to determine the outcome, $y$. The inputs, $x$, are the variables or characteristics that the lender uses to determine the outcome. The outcome, $y$, could be the interest rate of the loan or the fees associated with the loan, or whether to approve the loan altogether.[41]

Pricing inputs could include borrower characteristics, such as the borrower's income or years of education, as well as the characteristics of the loan application, such as the loan amount. In traditional mortgage lending, a borrower's creditworthiness is assessed based on past credit behavior, often with the assistance of a credit bureau, such as Experian or Equifax, or relies on a borrower's FICO score. The borrower's income and future income is assessed to determine borrower liquidity. Lenders also use the specific characteristics of the loan, and the securitized property, to determine the terms of the loan. The exact terms of the loan could vary greatly across borrowers, and so there is a degree of personalization of the prices paid by borrowers.

Credit terms are also personalized because they are partially determined by lender employees or brokers (jointly "loan officers") who have discretion. In traditional mortgage lending the originator sets the lowest price at which they are willing to extend a loan. Borrowers then meet with loan officers who help set the exact terms of the loan. Loan officers are often incentivized to provide a more expensive loans.[42] These interactions are likely to increase the personalization of loans.[43]

[41] In this Article I focus on interest rates, but this is only one element of the cost of a mortgage. The overall cost of a mortgage is determined by other costs such as "discount points" and the compensation to the loan officer and broker. *See generally* Neil Bhutta et al., *Paying Too Much? Price Dispersion in the US Mortgage Market* (2019), *available at* https://papers.ssrn.com/abstract=3422904.

[42] The difference between the "par rate" and the final rate was known as the "yield spread premium" and was used to compensate loan officers. In the wake of the financial crisis, new regulations from 2010 prohibited loan officer compensation from directly being related to the interest rate. *See* Board of Governors of the Federal Reserve System, Truth in Lending, Regulation Z, 75 Fed. Reg. 185,58508 (Sep. 24, 2010). Even though loan officers are limited in their ability to be directly compensated for higher interest rates, more expensive loans are clearly more profitable for lenders and could ultimately affect loan officer compensation, albeit less directly. See Howell E. Jackson & Laurie Burlingame, *Kickbacks or Compensation: The Case of Yield Spread Premiums*, 12 STAN. J.L. BUS. & FIN. 289 (2006–2007), for a discussion on why yield spread premiums were problematic for consumers and for the argument that yield spread premiums lead to higher mortgage prices for consumers, which may fall disproportionately on the least sophisticated borrowers.

[43] This is because mortgage lenders often created borrower "bins" based on a limited set of characteristics in determining par rates. These bins were likely not based on sophisticated risk predictions but rather reflected more coarse divisions between lenders. As I have discussed elsewhere, it is not clear how exactly loan officers decided the final terms of the

Throughout this Article I focus on credit pricing that results from the prediction of default probability of the borrower. The lender predicts the default probability and then uses this default probability to directly set the price of the loan, such as the interest rate of the loan. I therefore refer interchangeably to the outcome $y$ as the predicted default probability and the loan price.[44]

There are several reasons that accurate default prediction might be beneficial for both lenders and borrowers, and provide reasons we would want to personalize credit pricing. The first reason is that default causes a loss for lenders. Therefore, when a lender can accurately predict default they can determine a cutoff for extending a loan or price risk accordingly.[45] Flat pricing could create significant harm because of the moral hazard caused by the selection of borrowers,[46] which could lead to the drying up of credit markets altogether. Inaccurate default prediction could also hurt consumers. The accurate prediction of default may mean that certain customers do not receive loans, however, this may be beneficial to consumers if we consider

loan. For example, we do not know whether loan officers were concerned with assessing credit worthiness or tried to learn borrower willingness-to-pay. *See* Gillis & Spiess, *supra* note 25.

[44] One can in theory separate the "prediction problem" from the "decision." *See e.g.*, Sam Corbett-Davies et al., *Algorithmic Decision Making and the Cost of Fairness*, PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 797 (KDD '17, ACM New York, NY, USA 2017).

Despite my focus on default probability, in reality default prediction is rarely the only metric used to personalize credit contracts. Personalization could reflect whether the loan is securitized or the purpose of the loan, as well the costs of administering the loan to the particular borrower. The personalized terms could also reflect the lender's assessment of the borrower's willingness to pay for the loan. A recent study suggests that there is a high degree of dispersion in the prices of mortgages. This suggests that many borrowers overpay for mortgages because they do not shop around or negotiate for a better rate. *See* Bhutta et al., *supra* note 42. I focus on default prediction personalization since this is arguably the least controversial basis for personalization. *See e.g.*, Robert Bartlett et al., *Consumer Lending Discrimination in the FinTech Era* 50, https://www.nber.org/papers/w25943. A continuation paper will discuss the different types of personalized pricing. This is particularly relevant to the "business justification" of disparate impact and whether the personalization of prices based on willingness-to-pay can qualify for this defense.

[45] This is often referred to as Risk Based Pricing. *See* Robert Phillips, *Optimizing Prices for Consumer Credit*, 12 J. REVENUE PRICING MGMT 360 (2013), ("a riskier customer should pay a higher price in order to compensate for the higher probability of default and the associated cost to the lender"). *See also* Michael Staten, *Risk-Based Pricing in Consumer Lending*, J.L. ECON. & POL'Y 33 (2015).

[46] *See* Dean Karlan & Jonathan Zinman, *Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment*, 77 ECONOMETRICA 1993 (2009) (using an experiment to document the existence of moral hazard in consumer credit markets).

that default and foreclosure are very costly for consumers.[47]

The accurate pricing of credit could also mean the ultimate expansion of access to credit. When lenders cannot distinguish between the risk of different borrowers, they may avoid lending to larger groups of applicants. The more accurate a lender's prediction, the more they are able to distinguish borrowers with different levels of risk. This may mean that some borrowers are less risky than previously believed which will expand access to credit, or than even riskier borrowers can receive a loan at a certain cost.[48] This is particularly likely to be the case in the tails of default prediction, meaning people with higher probability of default.

### *1.2 The problem of biased inputs*

Most inputs into a credit pricing decision in the traditional context reflect bias; however, the origin of that bias can vary greatly for different inputs. In this Section, I distinguish between a biased input that results from some historic or existing discrimination external to the lender itself ("biased world") and an input that is biased because of the way it defines and estimates a characteristic ("biased measurement"). Although analytically distinguishable, the difference between the two is often empirically indistinguishable.

A primary concern with personalized prices for credit is that it creates or further increases disparities among groups. Here I focus on bias that affects "protected groups," meaning the categories of people that discrimination law seeks to protect.[49] We therefore might be concerned that the way in which we predict default, and price credit accordingly, creates disparities among legally protected groups. As will be discussed further in Section 1.3 fair lending prohibits discrimination on the basis of race, religion, sex, marital status and age, among other grounds.

#### 1.2.1 Biased world

Lenders seeking to personalize credit terms to borrower's confront the

[47] *See* John Gathergood et al., *How Do Payday Loans Affect Borrowers? Evidence from the U.K. Market*, 32 REV. FIN. STUD. 496 (2019) (showing that at the credit score discontinuity for payday loans, consumers who received a loan were more likely to default and exceed bank overdraft limits).

[48] *See* Liran Einav at al.,*The impact of credit scoring on consumer lending,* 44 RAND J. ECON. 249 (2013) (showing that the adoption of an automated credit scoring at a large auto finance company led to higher-risk applicant lending).

[49] The two Act's that determine the protected groups for fair lending are the Equal Credit Opportunity Act (ECOA), (15 U.S.C. § 1691 (a)(1)-(2)), and the Fair Housing Act, (42 U.S.C. 3601).

problem that many of the factors used to determine individual risk are a product of pre-existing disadvantage or discrimination.[50] Although this is not the lender's fault, using these inputs exacerbates the effects existing discrimination in a new domain. There is no consensus on whether the use of biased world inputs gives rise to discrimination claims.[51]

There are several examples of "biased world" inputs. A central factor for determining repayment risk is a borrower's income. Past research has shown a significant racial and gender pay gap in the US.[52] These gaps may be a result of "pre-market factors,"[53] such as reduced access to higher education, or a result of labor market discrimination.[54] Similarly, higher rates of

[50] I use the term "discrimination" here to describe a reality in which a group was unfairly treated without considering whether those circumstances give formal rise to a claim of legal discrimination. This is sometimes referred to as "structural disadvantage." *See* Barocas & Selbst, *supra* note 7, at 691.

[51] *See infra* Section 1.3. According to some theories of discrimination, the use of "biased world" inputs does not give rise to a claim of discrimination, according to other theories, a situation of "compounding injustice" could trigger discrimination law. Deborah Hellman coined this term to describe a decision that "exacerbates the harm caused by the prior injustice because it entrenches the harm or carries it into another domain." *See* Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW (Hugh Collins & Tarunabh Khaitan eds., 2017).

[52] *See* Kayla Fontenot et al. *Income and Poverty in the United States: 2017* United States Census Bureau 6, 8-9 (2018), https://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-263.pdf.

Importantly, the black-white wage gap has increased as wage inequality has risen from 1979 to 2015. *See* Elise Gould, *State of Working America Wages 2018* (Feb. 20, 2019), https://www.epi.org/publication/state-of-american-wages-2018/.

[53] Pre-market factors are typically understood as factors that are used to "explain" wage gaps. The challenge is that these factors might themselves be a product of discrimination. For example, lenders often consider whether a borrower is self employed, which may be used to determine that their future income is less stable. *See* Alicia H. Munnell et al., *Mortgage Lending in Boston: Interpreting HMDA Data*, 86 AM. ECON. REV. 25 (1996) (finding that the probability that a loan request made by someone who is self-employed will be denied is roughly one third greater than the average denial rate, page 29); Todd J. Zywicki & Joseph D. Adamson, *The Law and Economics of Subprime Lending*, U. COLO. L. REV. 1, 9 (2009).

[54] Meaning that black workers with the same ability and education earn less than comparable white workers or face less employment opportunities.

While the racial wage gap in the labor market is well document, interpreting this gap and the extent to which it reflects either taste-based or statistical discrimination has proven difficult. *See* Dan Black et al., *Why Do Minority Men Earn Less? A Study of Wage Differentials among the Highly Educated*, 88 REV. ECON. & STAT. 300 (2006) (finding substantial wage gaps. The authors emphasize that they cannot rule out the possibility that this is a consequence of cultural or class prejudice.) *See also* Eric Grodsky & Devah Pager, *The Structure of Disadvantage: Individual and Occupational Determinants of Black-White Wage Gap*, 66 AM. SOC. REV. 542, 563 (2001) (find that although black men have gradually gained entry to highly compensated occupational positions, they have simultaneously become subject to more extreme racial disadvantages in respect to earning power). *See also* incarceration of racial minorities could also have a negative impact of credit scores.[55]

Levels of debt might also may reflect pre-existing disadvantage. For example, high interest lenders, such as payday lenders, often target minorities, leading to the accumulation of higher levels of debt.[56] There is also evidence that credit card lenders may also screen for minority consumers.[57]

When a lender who uses variables that reflect pre-existing disadvantage, or a biased world, that disadvantage is compounded and affects and new domain of lending. The bias input is then used to price credit that is more expensive for the disadvantaged group or even to deny credit altogether. Because credit is a way of potentially creating wealth this essentially could reinforce wealth gaps in the US.

### 1.2.2 Biased measurement

Many inputs into a pricing decision partially reflect measurement bias,

Roland G. Jr. Fryer, Devah Pager & Jorg L. Spenkuch, *Racial Disparities in Job Finding and Offered Wages*, 56 J.L. & ECON. 633, 690 (2013) (estimating that differential treatment accounts for at least one third of the black-white wage gap). Other studies have identified racial disparities in access to the labor market. *See e.g.*, Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 AM. ECON. REV. 991 (2004) (finding that white names triggered a callback rate that was 50% higher than that of equally qualified applicants with black names). *See also* John M Nunley et al., *Racial Discrimination in the Labor Market for Recent College Graduates: Evidence from a Field Experiment*, 15 B.E. J. OF ECON. ANALYSIS & POL'Y 1097 (2015).

[55] A recent paper documented the negative impact of incarceration on credit scores and income. *See* Abhay P. Aneja and Carlos F. Avenancio-Leon, *No Credit For Time Served? Incarceration and Credit-Driven Crime Cycles* (2018), *available at* https://abhayaneja.files.wordpress.com/2018/04/3bd94-incarcerationaccesstocredit-v022619.pdf. If black defendants are more likely to be incarcerated then the use of credit scores and income presents another way in which credit decisions rely on pre-existing disadvantage.

[56] *See* Oren Bar-Gill and Elizabeth Warren, *Making Credit Safer*, 157 U. PA. L. REV. 1, 66 (2008); Cassandra Jones Havard, *On the Take: The Black Box of Credit Scoring and Mortgage Discrimination*, B.U. PUB. INT. L.J. 241 (2010–2011) (arguing that subprime lending was incontrovertibly steered toward minority communities); Creola Johnson, *The Magic of Groups Identity: How Predatory Lenders Use Minorities to Target Communities of Color*, GEO. J. ON POVERTY L. & POL'Y 165, 169 (2010) (describing various marketing practices used by lenders to target minorities for predatory loans).

[57] *See* Andrea Freeman, *Payback: A Structual Analysis of the Credit Card Problem Financial Reform During the Great Recession: Dodd-Frank, Executive Compensation, and the Card Act*, ARIZ. L. REV. 151, 181 (2013) ("Credit card companies confine low-income individuals to a subprime market and attempt to steer many middle-class African American and Latinos into subprime loans.")

meaning that the way in which an input is defined or estimated is biased rather than the underlying characteristic being biased. While lenders may have more control over estimation that causes "biased measurement" inputs than "biased world" inputs, practically, these two types of biases are often indistinguishable.

The general reference to "borrower characteristics" masks the fact that any characteristic requires some sort of definition, measurement and estimation. For example, if we want to use a borrower's income, we must define what income is and how to determine a borrower's income. For instance, we will need to determine whether certain transfers, such as gifts from relatives are considered income, or whether to consider public assistance income.[58] It might also require a determination of the documentation needed to consider a transfer "income." When a definition systematically disadvantages a protected group, however, then it could be a case of a "measurement bias."[59]

Another type of "biased measurement" could arise when a substitute or a proxy is used in lieu of the characteristic that is of true interest. Often the variable that is of true interest is unobserved and so a lender might instead rely on a close substitute.[60]

In Section 3.1.1, I provide an example in which a borrower's "education" is used as a substitute for the borrower's "ability," which is relevant in determining future income. As "ability" is not observed by the lender, they could use borrower education as a proxy. If racial minorities are less likely to go to college for any given level of ability, this proxy will cause measurement bias. In this example, the problem I have highlighted is not necessarily created by pre-existing discrimination but by the imperfect measurement of the underlying variable of interest.

One central characteristic used to price credit, a borrower's credit score, may suffer from measurement bias. The exact inputs and models used to determine a credit score, such as a FICO score, is proprietary information so

[58] In fact, ECOA directly addresses this issue by prohibiting discrimination on the basis of whether an applicant is a recipient of public assistance income. The motivation behind adding this protected group was the conduct of lenders who refused to consider such income for the purpose of extending a loan. *See* Taylor, *supra* note 31, at 339.

[59] The type of measurement bias I discuss here is "feature bias," which is bias in the predictors x. There is a second type of measurement bias called "label bias," which is bias in y. Label bias could arise, for example, when a recorded late payment is a noisy measurement of a true late payment. If whites are better at avoiding a late payment being formally recorded, then there is label bias. *See* Chander, *supra* note 38, at 17 (arguing that label bias is the more severe bias).

[60] In the context of employment, this issue often arises when characteristics such as job performance are measured using information such as supervisor's evaluations, which may be biased. *See* Kim, *supra* note 7, at 876.

that it is hard to know for certain how these scores may be biased. However, we do know that credit scores have traditionally considered a few measures of creditworthiness like lending from large financial institutions and mortgage payments. Other measures of creditworthiness, such as timely rental payments or borrowing from smaller and more local financial institutions may also be predictive of default.[61]

Although the theoretical distinction between "biased world" and "biased measurement" is clear, in many cases a variable might combine the two types of biases. For example, a borrower's income could reflect both pre-existing discrimination in labor markets as well as some kind of measurement bias. This is problematic for the view that the use of variables that reflect a "biased world" are permissible while variables that reflect "biased measurement" are impermissible, discussed in more detail in Section 1.3.[62]

Moreover, it is unclear whether as an empirical matter it is possible to distinguish between these two types of biases. We can learn whether a certain variable correlates with race, but we might not be able to determine the origin of the correlation. Above I presented intuitive explanations for the reasons a variable might correlate with race, but this is a far cry from establishing the source and explanation for the correlation, or whether it stems from pre-existing discrimination or measurement bias.

### *1.3 Traditional fair lending law*

Fair lending law is the primary lens through which to consider the personalization of credit pricing. Therefore, this Section provides an overview of fair lending law, which covers both the doctrine of disparate treatment, dealing with intentional discrimination, as well as disparate impact, dealing with facially neutral rules that have an impermissible impact. Because there are ongoing disputes with respect to the foundations and scope of the disparate impact doctrine, I discuss how the different positions view the problem of biased inputs, but I do not adopt a particular interpretation. I end by highlighting how the disagreements over the boundaries of the

[61] The fact that only certain types of behaviors are measured by credit scores could mean that some borrowers are not scored at all. Many consumers have thin credit files because they are less likely to access the types of financial services that report to the traditional credit bureaus. *See* Persis Yu et al., *Big Data, a Big Disappointment for Scoring Consumer Creditworthiness*, NAT'L CONSUMER L. CENTER page 12 Mar. 6, 2013. According to the CFPB, African-Americans and Latinos are more likely to be credit invisible, at rates of around 15% in comparison to 9% for whites. *See* The Consumer Financial Protection Bureau Office of Research, *Data Point: Credit Invisibles*, at 24-25 (May 2015), https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.

[62] See Kleinberg et al., *supra* note 36, for the distinction between "group differences in the raw data" and biases for the "choice of predictors."

doctrine have come to the forefront with a new proposed rule by the Department of Housing and Urban Development (HUD).

The two laws that form the core of credit pricing discrimination are the Fair Housing Act (FHA) of 1968 and the Equal Credit Opportunity Act (ECOA) of 1974. The FHA, also known as Title VIII of the Civil Rights Act of 1968, protects renters and buyers from discrimination by sellers or landlords and covers a range of housing related conduct including the setting of credit terms.[63] The FHA prohibits discrimination in the terms of credit based on race, color, religion, sex, disability, familial status and national origins.[64] In 1974, Congress passed the Equal Credit Opportunity Act (ECOA), banning discrimination in all types of credit transactions. ECOA therefore complements FHA by expanding discrimination provisions to other credit contexts beyond housing related credit. Initially ECOA only covered sex and marital status discrimination but was then amended in 1976 to also cover race, color, religion, and other grounds of discrimination.[65]

ECOA and FHA cover both discrimination doctrines of "disparate treatment," dealing with the direct condition of a decision on a protected characteristic, often with intent to discriminate, and "disparate impact," which typically involves a facially neutral rule that has a disparate effect on protected groups. ECOA and FHA do not explicitly recognize the two discrimination doctrines in the language of the law itself. However, the disparate impact doctrine has been recognized in the case of credit pricing by courts and agencies in charge of enforcing the laws. The Supreme Court recently affirmed that disparate impact claims could be made under the FHA in *Inclusive Communities*, [66] confirming the position of eleven appellate

[63] In 1988 the Fair Housing Amendments Act was passed, strengthening the mortgage lending provisions of the FHA.

[64] The Equal Credit Opportunity Act (ECOA) (15 U.S.C. § 1691 (a)(1)-(2)) and the Fair Housing Act (42 U.S.C. 3604).

[65] There are other laws that have additional provisions relating to credit pricing discrimination that are not my focus in this Article. The Community Reinvestment Act (CRA) of 1977, encourages banks and other lenders to address the needs of low-income households within the areas they operate, which often overlaps with serving racial minority areas. The CRA does not give a right to private action but rather instructs the relevant supervisory agency on how to oversee that institutions are serving the lending needs of their community. Another federal law related to credit pricing discrimination is the Home Mortgage Disclosure Act (HMDA) (12 U.S.C. § 2801) (1989), which requires that certain financial institutions make regular disclosures to the public on mortgage applications and lending. Although HMDA does not contain any explicit discrimination provisions, one of its purposes is to allow the public and regulators to consider whether lenders are treating certain borrowers in certain areas differently. The empirical sections of this Article rely on HMDA data.

[66] Texas Dep't of Hous. and Cmty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507 (U.S. 2015).

courts and various federal agencies including HUD, the agency primarily responsible for enforcing the FHA.[67] 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.[68]

Disparate treatment involves the direct conditioning of the decision on a protected characteristic and therefore focuses on the *causal* connection between a protected characteristic and a credit decision.[69] The doctrine could be triggered by directly considering a protected characteristic, such as race, in a specific credit decision or when a protected characteristic is used in setting general lending policy, such as in the case of "redlining."[70] Disparate treatment identifies cases in which a protected characteristic directly influenced a credit decision, it therefore is concerned with the *causal* relationship between protected characteristics and decisions.

Disparate impact, the second discrimination doctrine under FHA and ECOA, covers cases in which a facially neutral rule has an impermissible

[67] *See* Robert G. Schwemm, *Fair Housing Litigation after Inclusive Communities: What's New and What's Not*, COLUM. L. REV. SIDEBAR 106, 106 (2015). ("The Court's 5-4 decision in the *ICP* case endorsed forty years of practice under the FHA…", during which the impact theory of liability had been adopted by all eleven federal appellate courts to consider the matter.")

[68] *See e.g.*, Ramirez v. GreenPoint Mortgage Funding, Inc, 633 F. Supp. 2d 922, 926–27 (N.D. Cal. 2008). The CFPB has recently proposed abandoning disparate impact liability under the ECOA. *Compare* Consumer Financial Protection Bureau, *Consumer Laws and Regulations: Equal Credit Opportunity Act* (June, 2013) ("The ECOA has two principal theories of liability: disparate impact and disparate theory.") *with* Mick Mulvaney, *Statement of the Bureau of Consumer Financial Protection on Enactment of SJ Res 57* (Consumer Financial Protection Bureau, May 21, 2018) (stating that the CFPB will reexamine its guidance on disparate impact liability under the ECOA in light of "recent Supreme Court decisions distinguishing between antidiscrimination statutes that refer to the consequences of action and those that refer only to the intent of the actor"). For a skeptical view of whether the statutory language of ECOA supports disparate impact *see* Peter N. Cubita & Michelle Hartmann, *The ECOA Discrimination Proscription and Disparate Impact - Interpreting the Meaning of the Words That Actually Are There Survey - Consumer Financial Services Law*, BUS. LAW 829 (2005–2006).

[69] In the employment discrimination context, see Sullivan, *supra* note 7, at 408 (suggesting that in the employment context, one way to read Title VII is that it "embraces ad causal view of what we call disparate treatment.")

[70] Redlining is the practice of denying credit to borrowers from predominantly minority neighborhoods and is typically considered a case of disparate treatment. Some early trial cases established the disparate treatment claim under the theory of "redlining". *See* Laufman v. Oakley Building & Loan Company, 408 F. Supp. 489 (S.D. Ohio 1976). The theory behind redlining is that the racial composition of an area was used to make a loan decision and therefore the decision depended directly on a protected characteristic. Moreover, for many years geographical lines were so strongly associated with racial divisions that it seemed natural for litigants to consider geographical criteria as being close to racial criteria.

disparate effect. A disparate impact case typically follows the burden-shifting framework that was developed primarily in the Title VII employment discrimination context.[71] The first step of the framework is the plaintiff's *prima facie* showing of a disparate outcome for a protected group.. This requires the plaintiff to identify the specific conduct or policy that led to the disparate outcomes. Once a plaintiff has established the disparate outcome and the cause of the outcome, the burden shifts to the defendant to demonstrate that there was a business justification for the conduct or policy that led to disparity.[72] The burden then shifts back to the plaintiff to demonstrate whether there was a less discriminatory way to achieve that same goal.

In spite of the formally coherent structure of a disparate impact claim, there is significant disagreement over the philosophical foundations of the doctrine and whether case law and regulatory action are consistent with those foundations. One of the most important disagreements is over the extent to which disparate impact is meant to address cases that are more about effect than intent.[73] According to one theory, which I call the "intent-based" theory, disparate impact treats unjustified discriminatory effects as a proxy for the true concern of interest which is the discriminatory intent.[74] This account

[71] Disparate impact first entered US law in the 1971 breakthrough case Griggs v. Duke Power Company, 401 U.S. 424 (1971), in which hiring requirements of a high school diploma and an aptitude test were challenged. A formal burden shifting framework was articulated in the subsequent employment decision Albermarle Paper Co. v. Moody, 422 U.S. 405 (1975), and this was articulated into the three-step burden-shifting approach that is applied today. This burden-shifting framework was formalized into the language of Title VII in §703(k), added by the Civil Rights Act of 1991. A similar language exists in the HUD Disparate Impact Rule 2013. *See* Regulation B, §202.6, footnote 2 (discussing the relevance of Title VII for interpreting fair lending disparate impact). See also Equal Credit Opportunity, 41 Fed. Reg. 29870, 29874 (July 20, 1976) ("Congress intended certain judicial decisions enunciating this "effects test" from the employment area to be applied in the credit area.")

[72] A central question in this context is what type of business justification can be considered legitimate. See Louis Kaplow, Balancing Versus Structured Decision Procedures: Antitrust, Title VII Disparate Impact, and Constitutional Law Strict Scrutiny (forthcoming), for a detailed discussion of this burden shifting framework in the context of employment discrimination.

[73] For an articulation of these disagreements see Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, HARV. L. REV. 494, 520 (2003–2004). There are other debates around disparate impact, or "indirect discrimination," a similar doctrine in the Europe and many other countries. For example, one important question is the extent to which disparate impact and indirect discrimination represents a moral wrong, or whether there is some other policy justification for the doctrine. A related question is whether disparate impact should in fact be considered discrimination at all. *See generally* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67 (2017–2018).

[74] *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake*, 53 UCLA L. REV. 701, 708 (2005–2006) (tracing the origins and implementation of disparate impact in the

emphasizes disparate impact's ability to unearth cases in which there is a discriminatory motive that is hard to prove.[75]

A second theory of the disparate impact doctrine is that disparate outcomes are a concern in of themselves and the doctrine should be understood as an attempt to "dismantle racial hierarchies regardless of whether anything like intentional discrimination is present."[76] This second theory has also characterized disparate impact as "disturbing in itself, in the sense that a practice that produces such an impact helps entrench something like a caste system."[77] This theory of disparate impact, which I call the "effect-based" theory of disparate impact, views intent as irrelevant not only for evidentiary reasons.[78]

context of Title VII to argue that it may have limited a more expansive theory of intent under disparate theory). *See also* Nicholas O. Stephanopoulos, *Disparate Impact, Unified Law*, 128 YALE L. J. 1566 (2019), for a discussion of this theory in context of voting discrimination.

[75] *See* Primus, *supra* note 74, at 518 (discussing the view that "disparate impact doctrine is an evidentiary dragnet designed to discover hidden instances of intentional discrimination", in the context of Title VII). Another distinction that is often made, primarily in the context of the Equal Protect Clause, is between legal scholars who argue that discrimination law is meant to target arbitrary misclassification of individuals ("anticlassification") and scholars who assert that discrimination law targets practices that disadvantage groups or perpetuate disadvantage ("antisubordination"). *See e.g.*, Jack M. Balkin & Reva B. Siegel, *American Civil Rights Tradition: Anticlassification or Antisubordination Fiss's Way: The Scholarship of Owen Fiss: I. Equality*, U. MIAMI L. REV. 9 (2003–2004). Balkin and Siegel's primary focus is focus on the Equal Protection Clause, however, they point out that an anticlassification reading of Title VII disparate impact would view the doctrine as primarily concerned with implicit disparate treatment. *Id.* at 22.

[76] *See* Primus, *supra* note 74, at 518. Primus provides a more detailed discussion of the different possible motives of Title VII disparate impact. *See also* Stephanopoulos, *supra* note 75, at 1604 (discussing the view that purpose the disparate impact doctrine is to improve the position of minorities by "preventing their existing disadvantages from spreading into new areas, and ultimately to undermine the entrenched racial hierarchies of American society.") *See also* Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law after Inclusive Communities*, CORNELL L. REV. 1115, 1132 (2015–2016). *See also* Richard Primus, *The Future of Disparate Impact*, MICH. L. REV. 1341, 1352 (2009–2010) ("Disparate impact doctrine was widely understood as a means of redressing unjust but persistent racial disadvantage in the workplace, and antidiscrimination law was broadly tolerant of deliberate measures intended to improve the position of disadvantaged minority groups.")

[77] Cass R. Sunstein, *Algorithms, Correcting Biases*, SOCIAL RESEARCH 6 (2018).

[78] Despite the large conceptual difference between intent-based and effect-based theories of disparate impact, many cases are somewhat consistent with both understandings of the doctrine. *See* Bagenstos, *supra* note 77, at 1132, arguing that *Griggs* is consistent with both understandings of disparate impact. In the context of fair lending, disparate impact cases are somewhat vague, such as challenging the practice of mortgage originators allowing loan officer's discretion in setting loan terms. These cases argue that the loan officer discretion leads to higher rates for minority borrowers. *See* Ian Ayres et al., *The Rise and (Potential) Fall of Disparate Impact Lending Litigation*, *in* EVIDENCE AND INNOVATION IN HOUSING

Under both theories, the need to establish causal connections between "policies" and "outcomes" is at the heart of disparate impact. Under either approach the plaintiff must establish the causal link between a policy and a disparate outcome to make a *prima facie* claim of disparate impact.[79] The stringency of this requirement will determine how broad or limited a disparate impact claim can be. Similarly, in the second stage of the burden shifting framework, a defendant can demonstrate the *causal* link between the policy that caused disparity and a business justification.

The emphasis on establishing these causal connections reflects the centrality of input scrutiny for both disparate treatment and disparate impact. Disparate treatment is concerned with the direct conditioning on a protected characteristic, thereby scrutinizing whether a protected characteristic was an input to the decision. Disparate impact, despite its name, is also concerned

LAW AND POLICY 231 (2017).) The central question is how the loan officers were exercising their discretion, and whether they were intentionally discriminating against racial minorities. Based on this ambiguity, Schwemm and Taren argue that these cases may be considered a hybrid impact/intention case. This is because the conduct being scrutinized is the discretion provided to brokers that resulted in a disparate impact on minorities. However, it is this discretion that may have allowed brokers to intentionally discriminate against minorities. *See* Robert G. Schwemm & Jeffrey L. Taren, *Discretionary Pricing, Mortgage Discrimination, and the Fair Housing Act*, HARV. C.R.-C.L. L. REV. 375, 406 (2010). Proponents of the intent-based theory would emphasize the fact that underlying these cases is the concern that mortgage brokers were directly considering race in setting terms condition. On the other hand, proponents of the effect-based theory can point to the fact that the cases do not claim any intent on the part of the mortgage originators who are the target of the discrimination claim. See Michael Selmi, *Indirect Discrimination and the Anti-Discrimination Mandate*, *in* PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW 257 (Deborah Hellman & Sophia Moreau eds., 2013) for a discussion of how disparate impact's limited effect in practice is linked to its difficulty to relate to employer "fault."

[79] According to the HUD Joint Policy statement, the plaintiff needs to "identify the specific practice that cause the alleged discriminatory effect". *See* Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18266 (Apr. 15, 1994). The Supreme Court in *Inclusive Communities*, emphasized the causality requirement: "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… A plaintiff who fails to allege facts at the pleading stage or produce statistical evidence demonstrating a causal connection cannot make out a prima facie case of disparate impact." Texas Dep't of Hous. and Cmty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507 (U.S. 2015). For a discussion of whether and how *Inclusive Communities* differs from the HUD joint policy, see: Schwemm, *supra* note 68. *See also* Implementation of the Fair Housing Act′s Disparate Impact Standard, 84 Fed. Reg. 42854 § 100 (U.S. Aug. 19, 2019). The test in *Inclusive Communities* was recently incorporated in HUD Proposed Rule 2019, at 42858. *Also see* OCC 97 Bulletin 97-24. For the OCC to find that credit score meets can be justified by a business necessity the variable causing the disparity must have "an understandable relationship to an individual applicant's creditworthiness" (page 11). In the employment context, legal scholars have argued that there needs to be some causal link between job performance and input, and not correlation alone is insufficient.

with the inputs into a decision. Although the *prima facie* case requires an analysis of effects or outcomes of a policy, the focus quickly shifts to what inputs created this disparate and whether they relate to a legitimate business justification.

To return to the two categories of biased inputs, does the use of "biased world" or "biased measurement" inputs trigger discrimination law? On one account the use of bias inputs should not trigger the doctrine of disparate treatment because there is no direct conditioning on a protected characteristic. Similarly, the use of biased inputs should not give rise to a claim of disparate impact because "biased world" inputs are not a result of any actions on the part of the mortgage originator and will continue to exist regardless of the actions of the mortgage originator.[80] However, the effect-based theory of disparate impact may be wary of the use of "biased world" inputs if they entrench and compound existing disadvantage. Under either approach we may be concerned when a biased input highly correlates with a protected characteristic that it becomes a "proxy" for the characteristic.[81]

The use of "biased measurement" inputs arguably gives rise to more liability on the part of the lender. This is because the lender may have a choice as to how to measure an underlying characteristic or may be able to exert effort to avoid biased measurement. For example, a lender could create a procedure for verifying income from multiple employers and sources, and measure income that is less consistent or formal.[82]

However, both legal scholars and the law overstate lenders' ability to choose between "biased measurement" inputs and "biased world" inputs. As discussed above, many inputs are a hybrid of both biased world and biased measurement. A further issue relates to what is reasonable to expect from a lender in avoiding measurement bias inputs. As mentioned above, credit scores are likely to be a biased measurement of credit worthiness because they focus on certain behaviors that signal creditworthiness and not others, such as timely rental payments. It seems unreasonable to expect a lender to collect all the information a credit bureau would collect along with other consumer payment behaviors in order to address the issue of biased

[80] This may depend on the interpretation of the business justification. If a lender used biased inputs to predict willingness-to-pay, and this type of prediction is not a legitimate business justification, then the conduct could trigger discrimination law. Typically, the prediction of default as the basis for pricing is the least controversial of the business justifications a lender can provide.

[81] I discuss this in detail *infra* Section III.

[82] The use of a "biased measurement" input may also reflect discriminatory intent. Once a lender faces a choice in the way they define and measure a variable, a lender's intention may come into play. This Article does not fully address the issue of a lender who disguises their discriminatory intent through algorithmic decision-making. For further discussion of this type of discrimination see Kleinberg et al., *supra* note 36, at 29.

measurement.

In Part III, I discuss current positions on how to apply discrimination law to an algorithmic context given the challenge of biased inputs. I analyze four positions that represent a range of views on how to understand the role and definition of discrimination law. I begin by discussing the approach of excluding protected characteristics. This approach has been argued as sufficient to negate a discrimination claim, both disparate treatment and disparate impact, according to the intent-based theory of disparate impact.[83] I then discuss approaches that further exclude inputs that correlate with protected characteristics, more in line with the effects-based theory of disparate impact.[84] I end my discussion with a statistical approach to orthogonalizing inputs to an algorithm.[85]

In conclusion, although there is often agreement that fair lending law covers both the disparate treatment and disparate impact doctrines, there is disagreement on the theoretical basis and the boundaries of disparate impact. These disagreements have implications for the legality of using biased inputs, an issue that will become more pronounced in the algorithmic context, as discussed in the next part.[86]

## II. THE CHANGING WORLD OF CREDIT LENDING

Credit pricing is moving away from a process that relies on few variables and involves human discretion in setting the final terms to a world in which big data and machine learning are used instead. This is likely to change the ways in which we determine whether a pricing method amounts to "disparate treatment" or whether it causes "disparate impact."

I begin this Part by describing the changes taking place in the context of credit pricing on which I focus in this Article. I then present the central methodology of my Article, which is a simulation exercise in which a hypothetical lender uses machine learning to price credit. Building on the simulation exercise, the Part ends with a discussion of the meaning of the changes for pricing using biased inputs and the application of fair lending law. My conclusion is that algorithmic pricing could in some cases exacerbate the problem of biased inputs but in other cases mitigate the harm.

[83] *See infra* Section 3.1.

[84] *See infra* Section 3.2 and 3.3.

[85] *See infra* Section 3.4.

[86] For further discussion of how case law and a statutory language does not fully support any one theory of disparate impact, see Primus, *supra* note 74, at 518–36 ("As one might expect from a doctrine with polyglot origins, no single theory makes sense of all of the data. The statutory text is sketchy, and the cases speak in more than one voice.")

### *2.1 What is changing?*

Changes in how people receive credit is related to the larger revolution brought on by the Fintech industry, a term used to describe the segment of financial services characterized by digital innovations and technology-enabled business model innovations.[87] In this Article, I focus on the technological change in the pricing of credit and the role of artificial intelligence (AI).[88] I discuss on three aspects that are reshaping the personalization of credit pricing, namely the use of non-traditional data and advanced prediction technologies, and the automation of lending decisions.[89] Many lenders have incorporated a version of all three trends while other lenders have only partially adopted some of these changes.

There are an increasing number of Fintech companies that act as alternative credit providers to traditional lenders. These alternative lenders operate in several domains including mortgages, auto loans,[90] credit card lending, and personal loans.[91] In addition, many traditional lenders are using the services of third parties that engage in alternative ways of predicting

[87] "Fintech" covers a large range of financial activity, including payment and trading systems. In mortgage markets, one of the biggest changes in recent years has been the increase in online platforms that offer mortgages, which allows consumers to conduct the full process of mortgage origination online. This creates automation not only with respect to determining the price of credit but also related to the bureaucratic aspects of a mortgage.

[88] There are many ways in which artificial intelligence can assist with the process of lending in ways that are separate from their prediction of credit worthiness. For example, AI can help with organizing and reading paperwork, which is especially onerous in the case of mortgage.

[89] In analyzing the changes in credit pricing and their implications, a central question that arises concerns the baseline for the comparison. One can consider a range of credit pricing, from human decision-making to machine learning. For some of my analysis, the focus is on the move from similar empirical methods, like linear regression pricing, to machine learning pricing. When discussing changes in human discretion in setting the terms, I primarily focus on the change from the loan officer pricing to machine-learning pricing.

[90] *See* Becky Yerak, *AI Helps Auto-loan Company Handle Industry's Trickiest Turn*, WALL STREET J (Jan. 3, 2019), https://www.wsj.com/articles/ai-helps-auto-loan-company-handle-industrys-trickiest-turn-11546516801 (using 2,700 characteristics instead of the few it was using before). Other companies that have embraced this type of lending. For example, Synchrony Financial and Ford Motor Credit Co.

[91] For example, Upstart (https://www.upstart.com), uses education and other academic variables to set the price of credit, based on the idea that these variables measure propensity to pay that may not be reflected in characteristics like FICO scores. Another company, Lendbuzz (https://lendbuzz.com), targets populations that may not have easy access to credit, such as foreign students who are less likely to have US credit histories. The alternative lender, Crest Financial, for example uses the software of DataRobot for underwriting decisions, https://builtin.com/artificial-intelligence/ai-finance-banking-applications-companies.

creditworthiness and pricing credit.[92]

The Fintech market share in borrowing services is significant and increasing. According to one estimate, 82% of lenders report using nontraditional and alternative data in lending decisions.[93] The segment of the lending sector that relies on machine learning and big data is also likely to increase over time. A recent survey by Fannie Mae found that 27% of mortgage originators currently use machine learning and artificial intelligence in their origination process whereas 58% of mortgage originators expect to adopt the technology within two years.[94]

### 2.1.1 Nontraditional data

The first change taking place in the world of credit is the expansion of credit decision "inputs" to nontraditional data. Data such as payment and consumer behavior, social media behavior, and digital footprints are being increasingly used to price credit.[95]This is unlike traditional lending, which

[92] For example, ZestFinance (https://www.zestfinance.com), uses machine-learning to predict creditworthiness by providing modeling services that utilize the data already held by lenders. In approving personal loans it helps lenders use information from the loan application process to identify individuals who are likely not to pay back the loan. *See* AnnaMaria Andriotis, *Shopping at Discount Stores Could Help Get You a Loan*, WALL STREET J. (Mar. 4, 2019), https://www.wsj.com/articles/use-a-landline-that-could-help-you-get-a-loan-from-discover-11551695400. *See also* https://www.underwrite.ai; Lenddo, SAS, Equifax, and Kreditech

[93] *See* Aite, *Alternative Data Across the Loan Life Cycle: How FinTech and Other Lenders Use It and Why* (2018). https://www.experian.com/assets/consumer-information/reports/Experian_Aite_AltDataReport_Final_120418.pdf?elqTrackId=7714eff9f5204e7ca8517e8966438157&elqaid=3910&elqat=2. *See also* The Financial Stability Board, *Stability Implications from FinTech – Supervisory and Regulatory Issues that Merit Authorities' Attention*, 35 (2017), *available at* https://www.fsb.org/wp-content/uploads/R270617.pdf. ("Innovations in financial services are applying rapidly evolving technologies in new ways and leveraging different business models. New technologies include big data, artificial intelligence, machine learning, cloud computing and biometrics.")

[94] *See* Fannie Mae, Mortgage Lender Sentiment Survey: How Will Artificial Intelligence Shape Mortgage Lending (Oct. 4, 2018), http://www.fanniemae.com/resources/file/research/mlss/pdf/mlss-artificial-intelligence-100418.pdf. It is important to keep in mind that this is the utilization of AI in all aspects of the process, not only risk assessment. For example, use of AI to enhance consumer experience.

[95] See Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11183 (Feb. 21, 2017), for a definition of traditional data. *See* Hurley & Adebayo, *supra* note 7, at 162, for a useful overview of some of the non-traditional data sources. The use of non-traditional data is also taking place in other domains in which algorithms are used to make decisions. In the context of employment decisions see Kim, *supra* note 7, at 861.

relied on relatively few defined characteristics.

Lenders are increasingly using borrower characteristics that, though intuitively relevant to creditworthiness, were not traditionally used to price credit. For example, information on education, such as the school attended and degree attained,[96] and GPA and SAT scores,[97] intuitively relate to a borrower's future income and are therefore relevant to default risk. This type of information is particularly valuable for young borrowers who have yet to build up a credit history and therefore typically have difficulty obtaining certain types of loans.[98]

Credit scores have traditionally only used loan payments to large and established financial institutions to determine creditworthiness. Lenders are therefore increasingly using information on timely payment of utility bills and rent payments as indicators of creditworthiness for people without credit history.[99] Similarly, data on phone bills and short-term loans, which were often not included in credit files, are now used by Fintech lenders. Companies with rich information on consumer behavior, such as Alibaba, are using this information to create alternative credit scores.[100]

Consumer behaviors discernable at the time the loan is requested are also being used in pricing credit. For example, a recent paper looks at the use of "digital footprint" data, such as the device and operating system used by the consumer when using a furniture purchasing website, to determine creditworthiness.[101] These digital footprints predicted default slightly better than traditional credit bureau scores, suggesting that the digital footprints hold information that is not contained  in credit scores.[102] The use of these types of data may be particularly valuable for short-term lenders and consumer websites that offer "ship-first pay-later," creating a quasi short-

[96] *supra* note 94., page 10.

[97] *See* for example Upstart (https://www.upstart.com). This is information that was is typically not considered in credit scoring, such as FICO scores. *See* Hurley & Adebayo, *supra* note 7, at 9.

[98] In some cases, traditional credit rating agencies, recognizing the problem that many people do not have adequate credit histories, have begun to develop their own alternative credit files. FICO Expansion, for example, considers debit data and utility data among other types of data

[99] *supra* note 94.Aite, page 7. Credit bureaus are becoming increasingly aware of this problem and so solutions, such as Experian's RentBureau allow consumers to incorporate information about rent payment history into their credit file. This indicates that non-traditional data may over time be incorporated into traditional metrics.

[100] For example, Sesame Credit.

[101] *See* Tobias Berg et al., *On the Rise of FinTechs – Credit Scoring Using Digital Footprints* (2018).

[102] *See Id.* The combination of the digital footprints with traditional bureau scores provided the most accurate prediction. This suggests that digital footprints and traditional scores are complements rather than substitutes.

term loan.

Fintech lenders are also using social media to price credit and to verify borrower information. Although social media data might not intuitively seem related to credit worthiness, third parties are using this information to provide lenders with alternative or additional data on borrowers.[103] Social media data can also be used to verify borrower information.

The use of nontraditional data not only contains the potential for more accurate creditworthiness predictions but also may allow for the expansion of credit to populations that have traditionally been excluded from credit markets. In the United States, 11% of adults have no credit record at all whereas an additional 8.3% have thin credit records that deem them "unscorable,"[104] so any lending that requires such a score will automatically not be accessible to nearly one-fifth of the population. The use of nontraditional datasets would give more people access to credit.[105]

### 2.1.2 Advanced prediction technologies

Traditional credit pricing uses simple models for differentiating among people in terms of their default risk, as discussed in Part I. In recent years, credit pricing is increasingly using more complex prediction methods, such as machine learning, that allow for more accurate default prediction. These advanced prediction technologies can be differentiated from more traditional types of credit scoring in which the weight that various variables receive is determined at the outset.[106] In the case of machine learning, the algorithm

[103] See discussion in Chris Brummer & Yesha Yadav, *Fintech and the Innovation Trilemma*, 107 GEO. L.J. 235, 265 (2018–2019) ("Importantly, unlike in earlier decades, when information on underlying loans or mortgage-backed securities was sourced through central nodes of information—such as credit rating agencies or conventional news organizations—today the production of digital data is often decentralized. Specifically, data emerges from a diffuse proliferation of websites, social media, and various genres of news sources and databases"). *See also* Rose Eveleth, Credit Scores Could Soon Get Even Creepier and More Biased, VICE (Jun. 13, 2019), https://www.vice.com/en_us/article/zmpgp9/credit-scores-could-soon-get-even-creepier-and-more-biased.

[104] The Consumer Financial Protection Bureau Office of Research, *Data Point: Credit Invisibles*, at 24-25 (May 2015), https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf. ("As of 2010, 26 million consumers in the United States were credit invisible, representing about 11 percent of the adult population. An additional 19 million consumers, or 8.3 percent of the adult population, had credit records that were treated as unscorable by a commercially-available credit scoring model. These records were about evenly split between those that were unscored because of an insufficient credit history (9.9 million) and because of a lack of recent history (9.6 million)"). *See also* discussion on page Hurley & Adebayo, *supra* note 7, at 155.

[105] *See also* Smith et al., *supra* note 7, at 13.

[106] *See* Hurley & Adebayo, *supra* note 7, at 162. (In the case of FICO scores, the "model assigns a numeric value for each of these five variables, and then applies a pre-determined

itself determines which inputs to use and what weights to assign them in reaching an accurate prediction.

The increased use of nontraditional data and machine learning are closely related to one another. This is because the use of nontraditional data increases the number of characteristics used to predict creditworthiness, and neither traditional prediction techniques nor human decision-makers are well-suited for high-dimensional data, a term used to describe data that contain many characteristics. Moreover, when characteristics do not bear an immediate and intuitive relation to the outcome of interest, it is difficult to determine which model to use in relating inputs to outcomes. Machine learning is optimal for this setting because it is designed to overcome difficulties in high-dimensional data and uses nonintuitive correlations to form accurate predictions.

The increase in prediction accuracy comes at a price of lower interpretability. Because machine learning algorithms are set up to optimize prediction accuracy, and not to produce a meaningful model of how inputs relate to outcomes, the algorithm outputs are not always easy to interpret. This issue that has received considerable attention in both academic and policy circles and has been the motivation behind legislation that attempts to mitigate the harms that stem from uninterpretable algorithms.[107]

### 2.1.3 Automation

Another important trend in credit lending is the automation of credit pricing— meaning the reduction of human involvement and discretion is setting prices. In an automated context, once the characteristics of the borrower and loan are set, the price of credit is automatically determined by some function or algorithm. This is a significant departure from some categories of traditional lending, particularly larger loans such as mortgages, which typically involved a broker and employee who would meet face-to-face with borrowers to determine the exact terms of the loan. Although these loans included a formulaic or automated aspect,[108] the ultimate loan terms

weight (in percentage terms) to each of these input values and averages them to arrive at a final credit score.")

[107] See discussions on the right to an explanation: Lillian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 19, 65-67 (2017); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SECURITY & PRIVACY 46, 40 (2018).

[108] For example, credit scores typically use some sort of algorithm to determine creditworthiness. Credit scores can either be used as a dimension used to price credit or the only determinant of credit price. In addition, Fannie Mae and Freddie Mac have typically used some type of algorithm to determine the price at which they purchase mortgages.

could not be known unless a borrower completed the application process.

Automation can offer several benefits. First, it may allow for a more efficient process of pricing and approving loans and a greater ability to adjust to changes in lending markets.[109] In addition, it may avoid errors in human judgment with respect to evaluating credit worthiness.[110] Typically, the literature refers to algorithms as "black boxes" and opaque. However, it is harder to imagine a decision-making process that is more of a "black box" that human decision-making. Automation brings an added level of transparency which provides important regulatory opportunities, as discussion in Section IV.

### *2.2 Simulation exercise – hypothetical "new world" credit lender*

To consider the implications of the change on credit pricing, I use a hypothetical "new world" lender. This lender takes data on past loans and their performance to predict the default risk of new borrowers. The lender then uses the predicted default risk to price credit. For example, the lender may determine that people above a certain risk of default will pay a higher interest rate on the loan. This hypothetical lender is a "new world" lender because it uses past loan information to form predictions using machine learning.[111]

My hypothetical lender uses loan information reported by mortgage lenders under the Home Mortgage Disclosure Act (HMDA)[112] to predict credit worthiness. Specifically, I use the Boston Fed HMDA dataset to which I add simulated default rates. Details on the Boston Fed HMDA dataset and the model I use to simulate default rates can be found in Appendix A. Although these default rates are based on real-world data, because they are simulated, any figures and numerical examples in this Article that show default rates should not be seen as reflecting real-world observations.

The prediction of loan default as a function of individual characteristics of the loan applicant from the training sample is made either by using a

[109] *See* Andreas Fuster et al., *The Role of Technology in Mortgage Lending*, 32 REV. FIN. STUD. 1854 (2019).

[110] A recent paper demonstrates how loan officers that have discretion may make worse decisions when busy, for example. *See* Dennis Campbell et al., *Making Sense of Soft Information: Interpretation Bias and Loan Quality*, J. ACCT. & ECON. 101240 (2019).

[111] New world credit lenders are likely to rely on data from past loans or data collected by third parties when predicting credit worthiness. Some lenders have information on their past loans, while some financial institutions may have other information about consumers, which they can then use to form predictions. New lenders or lenders seeking to improve predictions might rely on third parties that collect information on consumer and payment behaviors.

[112] Home Mortgage Disclosure Act (HMDA) (12 U.S.C. § 2801))

"random forest," in which the machine learning algorithm makes the prediction using decision trees,[113] or a "lasso regression," another common machine learning algorithm in which the algorithm selects the variables it deems most important for the prediction.[114] The algorithm is trained on a sample with 2000 clients, with more than 40 variables each (many of which are categorical). This function can then be applied to new borrowers, which is a subset of borrowers from the HMDA dataset not used to train the algorithm. One thing to note is that unfortunately, due to data limitations, this lender does not include many of the types of the nontraditional data discussed in Section II.2.1.1. Nonetheless, the lender uses more types of variables than what is typically used by mortgage originator in setting the "par-rate" in traditional lending.[115]

The purpose of this exercise is to demonstrate how advanced algorithms change lending decision-making and whether current approaches to discrimination law in the new context are likely to be effective. This methodology, which Jann Spiess and I first developed in "Big Data and Discrimination,"[116] allows for a meaningful analysis of the legal and methodological challenges in analyzing algorithmic decision rules in a stylized setting.

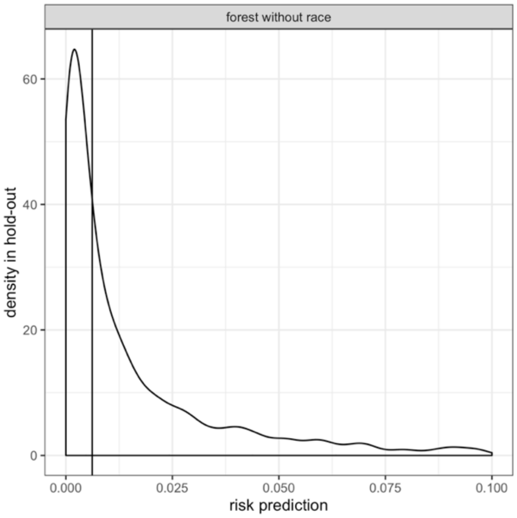


*Figure 1: Distribution of predicted risk. The graph shows the distribution risk for all borrowers in the holdout set of 2,000 borrowers. The graph is cutoff at 10%, meaning that only borrowers with a default*

[113] Leo Breiman, *Random Forests*, 45 MACHINE LEARNING 5 (2001).

[114] The objective of the lasso is to minimize the sum of squares between the true outcome and predicted outcome (like a linear regression), subject to regularization that restricts the magnitude of coefficients.

[115] For a full description of the variables in the Boston Fed HMDA dataset see Munnell et al., *supra* note 55.

[116] *See* Gillis & Spiess, *supra* note 25.

*risk of less than 10% are plotted. The vertical line is the median borrower (of the full sample, not just the borrowers with a risk below 10%).*

At the first stage I run a random forest algorithm on my training data, and then apply the resulting model to a new set of borrowers. In Figure 1, the model's prediction function is applied to a holdout set, meaning a subset of 2,000 borrowers that is drawn from the same distribution but was not used to train the prediction function. In the real world, this is likely to be a group of new applicants for which the lender is deciding whether to extend a loan and at what price. Borrowers who are to the left of the distribution have a lower probability of default. When credit pricing is based on default probability, these borrowers will pay a lower interest rate for a loan because they are less likely to default. Borrowers who fall on the right side of the distribution are more likely to default and therefore will pay a higher interest rate.[117]

The algorithm used to plot Figure 1 was race blind in the sense that it did not use the variable "race" to form its prediction.[118] However, the holdout dataset to which the prediction is applied does contain a "race" variable. We can therefore separately plot the default distribution for white and non-white and Hispanic applicants ("minority applicants"). Figure 2 shows the default distribution for white applicants (on the left) and minority applicants.

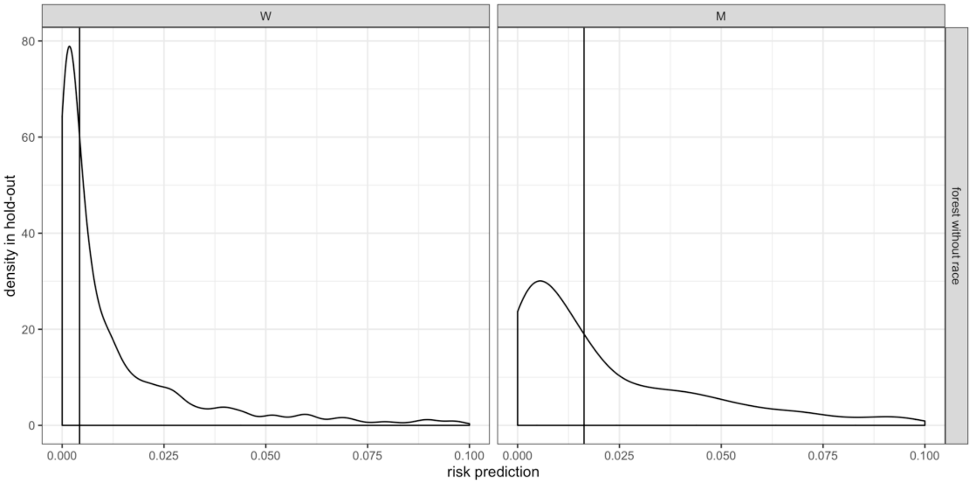


*Figure 2: Distribution of default risk for white (W) and minority (M) applicants. Both graphs are cut off at 10% default risk. The vertical line plots the median default risk for the full sample.*

[117] I emphasize the use of default risk as a way to set the price of the loan. But the default risk could also be used to decide who to approve for a loan altogether. A lender might have a cutoff for lending altogether so that applicants who are predicted to default above a threshold default probability will be denied a loan altogether.

[118] Throughout most of this Article I consider a lender who does not use the variable "race" in forming a prediction. This is simply because a lender who does not have a clear intention to discriminate is unlikely to use this variable. In Section 3.1 I discuss the exclusion of a protected characteristic in more detail.

Figure 2 shows that the default distribution is further to the left for white borrowers, reflecting a higher proportion of white borrowers who are low risk. This can also be seen by the vertical line, signifying the median applicant, which is further to the left for the white applicants than for the minority applicants. This simulation of an algorithmic lender will be used in the next part to demonstrate how this type of pricing changes how consumers are differentiated. In Section III, this simulation will be used to demonstrate the shortcomings of current approaches to algorithmic discrimination.

### *2.3 What are the challenges in the algorithmic context?*

It is important to understand how biased inputs affect credit pricing decisions. Although the problem of biased inputs is not new to the algorithmic context, its consequences may be different in the traditional and algorithmic setting.

On the one hand algorithmic pricing could exacerbate the problem of "biased world" because it increases the variance in predictions and may expand the number of "biased world" inputs due to its use of nontraditional data. Algorithmic pricing also allows for a greater ability to recover protected characteristics, as will be discussed in detail in Section III.[119] On the other hand, the algorithmic context could mitigate the harms of "biased measurement" by providing an increased amount of information on individuals.

#### 2.3.1 Biased world inputs in algorithmic pricing

The first way in which the move to the new world of credit pricing could increase the disparities between protected groups is by broadening input variables to include additional "biased world" inputs. This is the change that receives the most attention in the media and in legal writing. If algorithmic credit pricing differentiates between people along dimensions that correlate with race, then clearly the outcome disparities will increase.[120]

[119] This has drawn significant scholarly and policy attention. *See, e.g.*, Hurley & Adebayo, *supra* note 7.

[120] Another concern is that as the number of inputs increases, so will the number of inaccurate inputs. In general, the accuracy of the data used to price credit (and score consumers) is highly regulated. The Fair Credit Reporting Act stresses the importance of accurate data by obligating all "consumer reporting agencies" to ensure that credit histories are accurate (15 U.S.C. § 1681e(b) and see also the amendments to this act enacted as the Fair and Accurate Credit Transaction Act); *See* Yu et al., *supra* note 62, at 29 (expressing concerns about the use of big data in the lending process since "Expanding the number of data points also introduces the risk that inaccuracies will play a greater role in determining

Another way in which machine learning pricing could increase the disparities of credit prices is due to the greater ability of machine learning to personalize prices. The flexibility of the machine learning regression means that in forming predictions, the algorithm can better distinguish between individuals creating more granular predictions. This could mean that differences among individuals are more likely to translate into greater differences in predicted outcomes than would be true with other less flexible prediction technologies, such as a linear regression.[121] Even small differences between individuals could translate into greater gaps between the price for credit paid by white and non-white borrowers. One way to describe the increase in price personalization is through the variance of the distribution. A higher variance in the default probability means that people are more spread out in terms of the price they pay for credit, creating a greater range of predictions.

To consider how machine learning can increase the price variance, I compare a simple function using just a few variables with a machine learning algorithm that uses many variables. For the simple prediction function, I use an OLS (ordinary least squares) regression to predict default with a small subset of the variables available in HMDA.[122] For the machine learning prediction, I use a random forest with the full set of HMDA variables, other than "race." Therefore, the comparison between the simple function and the machine learning function differ along two dimensions; the number of variables and the prediction technology.

creditworthiness"); *See* also Robert B. Avery et al., *Credit Report Accuracy and Access to Credit*, FED. RES. BULL. 297 (2004) (a "follow up" research that examines the possible effects of data limitations in consumer credit report, including inaccuracies, on consumers).

[121] It is not clear that an OLS regression is the right comparison here since the typical "old world" pricing method relied on human discretion and perhaps human discretion is a more flexible prediction than some machine-learning regressions. However, the par-rate set by the mortgage originator is likely to rely on a function closer to an OLS regression if not more basic (such as default means within bins).

[122] The four variables used for this example are – "housingdti," "totaldti," "fixedadjustable," "loanterm"

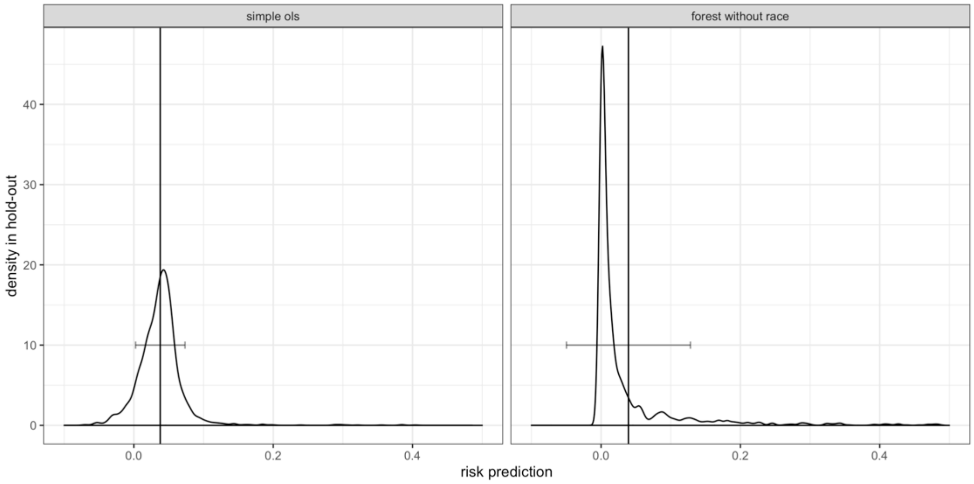


*Figure 3: Increase of spread with machine learning. For the graph on the left, an OLS regression was used to predict default with the indepent variables -"housingdti,""totaldti," “fixedadjustable,” “loanterm.” The prediction function was then applied to a “hold out” set. The graphs show the distribution of predicted default probabilities. For the graphs on the right, a random forest algorithm was used to predict default using all variables in the Boston Fed HMDA dataset, other than race. The prediction function was then applied to the same holdout set as the OLS prediction. The graph on the right shows the distribution of predicted default probabilities. The vertical lines are the mean default predictions, and the horizontal bars are the standard errors. Together, the mean and standard errors, demonstrate the “spread” of the prediction.*

Comparing the two distributions in Figure 3 demonstrates how the use of machine learning algorithm leads to borrowers being more spread out. This reflects the fact that the machine learning algorithm’s predictions have higher variance than the simple regression. That is, the price of credit is more personalized. The greater variance of the random forest prediction, represented by the wider horizontal bar, is the combined effect of the use of more inputs and a more flexible prediction technology.[123]

The increased variance of the random forest prediction has implications for racial disparities even though Figure 3 does not directly measure these disparities. When new credit pricing uses new data sources in which there are large differences between people who belong and do not belong to protected groups, the use of machine learning could translate the differences in inputs into larger outcome disparities. For example, if male and female borrowers are different with respect to inputs that predict default, a more flexible

[123] In reality, these two effects are also likely to be combined because, when confronted with big data, classic regression analysis leads to overfitting—constructing a model that corresponds so closely to the data at hand that it is unable to make meaningful predictions in other samples. Big data and machine learning therefore often go hand in hand. It is also important to keep in mind that both graphs do not use the type of nontraditional data that real-world algorithmic lenders are using, so that these graphs are understating the extent to which algorithmic pricing will increase variance.

prediction technology can increase the differences in predicted default for men and women.

The ultimate welfare implications of increased personalization are unclear in the real world.[124] As will be discussed in further detail below, the more accurate prediction may allow for certain groups previously denied credit altogether to now receive credit. Because of the ability to estimate their risk more accurately, a lender may agree to extend credit to this group at a higher interest rate than would be available to safer borrowers. This means that groups that were previously completely excluded from credit markets might be able to receive credit, albeit at a higher price than that made available to more creditworthy borrowers.

### 2.3.2. Biased measurement inputs in algorithmic pricing

Many of the concerns of the effects of big data and machine learning credit pricing discussed in the context of biased world also apply to variables that reflect biased measurement. The added variables and the increased flexibility that follows from the use of machine learning could increase the credit pricing disparities.

On the other hand, the use of big data and advance prediction technologies could lead to a decreased reliance on a biased proxy. For example, FICO scores may be biased because they reflect creditworthiness as measured by past mortgage payments but not timely rental payments, which are more prevalent for minorities. If big data provides lenders with the opportunity to use rental payment data in addition to FICO scores, this could reduce the differences in predicted default. In this example, the use of algorithmic credit pricing could decrease disparity rather than increase disparity.

There is empirical evidence that the use of nontraditional data leads to decreased reliance on FICO scores. A recent paper written by researchers at the Philadelphia Federal Reserve found that the correlation between the credit ratings of LendingClub,[125] a Fintech lender, and FICO scores has decreased over time, indicated by the increased usage of nontraditional data.[126] This evidence is consistent with the story that the impact of the measurement bias of FICO scores is reduced via the use of nontraditional data. [127]

The Consumer Financial Protection Bureau recently discussed the

[124] Andreas Fuster et al., Predictably Unequal? (Rochester, NY Nov. 6, 2018).

[125] These ratings are called “rating grades” and are determined by LendingClub.

[126] *See* Julapa Jagtiani & Catharine Lemieux, *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the Lendingclub Consumer Platform* (2018), https://papers.ssrn.com/abstract=3178461.

[127] *See Id.*

potential benefit of alternative data and machine learning in expanding credit. Based on the finding than an algorithmic lender's model "approves 27% more applicants than the traditional model, and yields 16% lower average APRs for approved loans," it concluded that "some consumers who now cannot obtain favorably priced credit may see increased credit access or lower borrowing costs" as a result of the use of non-traditional data.[128]

The conclusion is that it is difficult to assess at the outset the exact consequences of the widespread changes occurring in the world of credit pricing. On the one hand, the use of advance prediction technologies means that only inputs that contribute to prediction accuracy are considered in pricing. However, the use of biased inputs might further increase disparities when algorithms are better able to differentiate among people. On the other hand, the expansion of input data could undo some of the harm of measurement bias. Thus, the fact that the use of algorithmic pricing could either increase or decrease disparities relative to classic credit pricing suggests that only experimentation or empirical investigation can determine the direction of the effect. This will be further explored in Part IV.

## III. APPROACHES TO ALGORITHMIC DISCRIMINATION

The changes taking place in the landscape of credit pricing could have far-reaching implications for how fair lending law applies to the algorithmic setting. In this Part, I focus on the principal approaches of how to apply discrimination law to the algorithmic context, including approaches of legal academics and policy makers, along with proposed regulation. Some of these approaches have not developed primarily with credit pricing in mind but are highly relevant to fair lending.

Disagreements over the scope and boundaries of discrimination law in the non-algorithmic context, discussed in Section 1.3, carry into the new world. For the intent-based theory of disparate impact, the focus is primarily on whether a lender uses a protected characteristic in pricing, even when this occurs in a facially neutral way. For the effect-based theory of disparate impact, the concern will be whether algorithmic credit pricing exacerbates or entrenches disadvantage. The specific interpretation of the burden-shifting framework may be informed by these theories, such as the stringency applied to the initial burden on the plaintiff and how narrowly to construe the "business justification."

[128] Patrice Ficklin and Paul Watkins, Consumer Financial Protection Bureau, An Update on Credit Access and the Bureau's First No-Action Letter (Aug. 6, 2019), https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/.

Although I cover a wide range of approaches that are based on different interpretations of the doctrine, a common thread is their outdated focus on input scrutiny. These approaches follow the logic of traditional discrimination law by focusing chiefly on what goes into the algorithm ("inputs"). These approaches are inadequate for three reasons. First, they often fail on their own terms by not fulfilling their own loose definition of fairness. Second, they sometimes cannot be practically implemented and are unsuitable for the machine learning setting. Finally, at times these approaches could restrict access to credit for vulnerable populations and further entrench disadvantage.

I cover four approaches, summarized in Table 1.[129] The first approach I discuss is the exclusion of protected characteristics, primarily as a method for negating a claim of intentional discrimination under the "disparate treatment" doctrine.[130] Information about a person's protected characteristic is embedded in other information about the individual, meaning that a protected characteristic can be "known" to an algorithm even when formally excluded. I demonstrate this by showing that "age" and "marital status" can be predicted fairly accurately from the HMDA data. Moreover, we should be wary of excluding protected characteristics if we care about outcome disparities. As I demonstrate through a simulated example, price disparities could in fact *decrease* when algorithms are "race aware."

The second approach I discuss expands the exclusion of inputs to proxies for protected characteristics.[131] This approach recognizes that other inputs may act as "proxies" for protected characteristics and therefore should be excluded too. The approach, however, is not feasible when there is no agreed-upon definition of a proxy and when complex interactions between variables are unidentifiable to the human eye. Even inputs that have traditionally been thought of as proxies for race, such as zip codes, may be less concerning than other ways in which a borrower's race can be recovered. Using HMDA data, I demonstrate that there is a greater ability to predict "race" from traditional credit pricing inputs in HMDA than from ZCTAs, the Census Bureau equivalent of zip codes.

The third approach I discuss takes the opposite perspective of restricting

[129] One approach I do not explicitly discuss is the approach of modifying input data. *See, e.g.*, Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389 (2018–2019) (arguing that we should modify the information algorithms are fed). These approaches often lack an articulation of the criteria they are meant to fulfill, making them difficult to judge. Moreover, they often focus on modifying the algorithm's training data which does not address problems that stem from actual population differences when the algorithm is applied.

[130] *See infra* Section 3.1.

[131] *See infra* Section 3.2.

algorithm inputs to only preapproved inputs.[132] This is unlike the first two approaches that allow all inputs other than certain forbidden inputs. I argue that if algorithms are restricted to traditional credit pricing inputs, such as income and credit scores, this approach risks entrenching disadvantage without the ability of big data to undo or mitigate the harm from "biased measurement" inputs. More generally, I argue that this approach could ultimately restrict access to credit. Using HMDA data, I demonstrate that risk prediction using fewer inputs decreases prediction accuracy. When lenders are less able to differentiate among borrowers based on their risk, the lenders are limited in their ability to price the lending risk accurately. Lenders will likely then increase the price of credit and cut back on total credit extended.

The last approach I discuss, the orthogonalization approach, is based on a statistical method to prevent inputs that correlate with protected characteristics from serving as proxies.[133] Variables that are correlated with protected characteristics can both provide information relevant to the prediction and function as well as proxies for protected characteristics. This approach therefore seeks to isolate the component of these variables that serves as a proxy for protected characteristics. I argue that this framework is inappropriate for the machine learning context. I provide a technical demonstration of how the approach does not work in the machine learning context because variable selection is unstable. Building on this demonstration, I argue that the machine learning algorithm's variable selection should not be interpreted as reflecting some true model of the impact of the characteristic.

The primary source of the shortcomings of these approaches is that they continue to scrutinize decision inputs, similar to traditional fair lending, when this strategy is no longer effective in the algorithmic context. Approaches to discrimination law in the algorithmic age continue to rely on an outdated paradigm of causality, which has been the primary focus of traditional fair lending: Disparate treatment centered on the question of whether a protected characteristic had a causal effect on a credit decision. Disparate impact required plaintiffs to show a causal connection between disparities and a policy. A defendant could then negate a claim of discrimination by showing that a policy had a causal relationship to a legitimate business interest.

Machine learning, however, is a world of correlation and not causation. When using a machine learning algorithm to predict an outcome, the focus is on the accuracy of the prediction, and this is the metric by which the success of the algorithm is judged. Therefore, effective approaches to discrimination law in the algorithmic setting cannot rely on traditional causal analysis.

[132] *See infra* Section 3.3.

[133] *See infra* Section 3.4.

*Table 1: Summary of approaches*

| **Approach** | **What the approach is trying to achieve?** | **Can the approach be implemented?** | **Is the approach effective?** | **Is the approach otherwise undesirable?** |
|---|---|---|---|---|
| *Excluding protected characteristic* (Section 3.1) | No direct consideration of race | Yes | Algorithm can use protected characteristics regardless (recovery of protected characteristics) | Exclusion of protected characteristic can increase disparities |
| *Excluding proxies for protected characteristic* (Section 3.2) | No consideration of race through proxies | Difficulty in defining and identifying proxies | Algorithm can recover protected characteristics better than classic proxies (like zip codes) | — |
| *Restricting inputs to pre-approved characteristics* (Section 3.3) | No consideration or race through proxies (and possibly avoid large impermissible disparities) | Challenging to determine which inputs are permissible | Classic inputs can continue to serve as proxies | The selection of pre-approved variables could entrench disadvantage. High cost to prediction accuracy. |
| *Orthogonalizing inputs* (Section 3.4) | Prevent omitted variable bias (use of proxies for race) | Yes | Machine learning variable selection is unstable | Small differences in noise could lead to large variations in treatment |

### *3.1 Excluding protected characteristics*

One approach to addressing the concerns highlighted in Part I is to require that algorithms not consider a protected characteristic directly by excluding the characteristics as an input. This means that prior to running the algorithm on the training set, a lender would exclude any protected characteristics from the inputs of the algorithm, even if they were available to the lender. Formally, the prediction is blind to a borrower's protected characteristic because any two people who are identical except for the input "race," for example, would have the same predicted default probability.[134]

The requirement to exclude protected characteristics is mainly discussed in the context of the disparate treatment doctrine. Disparate treatment focuses on the intentional discrimination or the direct classification on the basis of a

[134] See Kleinberg et al., *supra* note 36, at 27 for an articulation of this approach ("the algorithm might be engaging in disparate treatment – as, for example, if it considers race or gender and disadvantaged protected groups (perhaps because racial or gender characteristics turned out to be relevant to the prediction problem it is attempting to solve).") *See also* Sunstein, *supra* note 78, at 7 ("Importantly, the algorithm is made blind to race. Whether a defendant is African-American or Hispanic is not one of the factors that it considers in assessing flight risk."). In the context of employment discrimination, *see* Sullivan, *supra* note 7, at 405. In Sullivan's motivating example, "Arti" is an algorithm who determines whom to employ: "Arti doesn't have any "motives" which seems to mean that its using a prohibited criterion to select good employees can't be said to violate Title VII's disparate treatment prohibition". Ultimately Sullivan argues that Title VII is primarily concerned with the causal connection between a protected characteristic and a decision, and "motivation" is one way to establish causality.

protected characteristic. Therefore, the requirement that an algorithm exclude a protected characteristics is seen akin to avoiding the classification on the basis of a protected characteristic.[135]

What is particularly appealing about the exclusion approach is that in the automated setting, protected characteristics can formally be excluded, which is often not possible in the human context. In the human decision-making context, very often the protected characteristic, such as race, is observed. This has been a major challenge for discrimination law, as it is difficult to plausibly show that an observed characteristic was *not* taken into account.[136] In the context of algorithmic decision-making, companies can guarantee the formal exclusion of these characteristics when they define or delineate the features used by an algorithm. Enforcement of the prohibition is also more feasible as long as there is some documentation of the inputs used by the algorithm.

Despite the intuitive appeal of this approach, as I will argue below, it is ineffective in guaranteeing that a protected characteristic is not used to form a decision. Moreover, this approach might lead to undesirable outcomes, particularly if we also care about the disparities created by a pricing algorithm.

### 3.1.1 Ineffective exclusion

Information about a person's protected characteristic is embedded in other information about the individual, meaning that a protected characteristic can be "known" to an algorithm even when formally excluded. The ubiquity of correlations in big data combined with the flexibility of

[135] This assumed translation between inclusions of a protected characteristic and "discriminatory intent" is not obvious. *See* Aziz Z. Huq, *What Is Discriminatory Intent*, CORNELL L. REV. 1211, 1242 (2017–2018) for a discussion of the various interpretation of discriminatory intent in context of the Equal Protection Doctrine. Discriminatory intent has been interpreted as "motivation" and "animus," which are human attributes and seem irrelevant when considering an algorithm. The basis for attributing discriminatory intent to an algorithm seems to derive primarily from discriminatory intent as an impermissible proxy or anticlassification understanding of intent, articulated by Huq. In the algorithmic setting the mainstream position seems to be that disparate treatment would require the exclusions of protected characteristics. For a related discussion of whether statistical discrimination violates the Equal Protection doctrine, *see* Will Dobbie & Crystal Yang, *Equal Protection under Algorithms: A New Statistical and Legal Framework* 7 (2019).

[136] *See* Kleinberg et al., *supra* note 36, at 16 (arguing that a major challenge for discrimination law has always been detecting and establishing illicit motivations.). The problem is deeper than a mere evidentiary barrier in establishing discriminatory intent, given that people might suffer for implicit bias and are unaware of how a protected characteristic shapes their decision. *See* Samuel R. Bagenstos, *Implicit Bias's Failure*, BERKELEY J. EMP. & LAB. L. 37 (2018).

machine learning means it is much likelier that an algorithm can recover protected characteristics. It is hard for the human eye to disentangle these correlations and interactions between variables to identify when an algorithm is actually using a protected characteristic. Particularly with the use of nontraditional data, much more can be inferred about a person's protected characteristic, such as their gender, age, and race.

Requiring the exclusion of protected characteristics implicitly assumes that an algorithm might want to use a protected characteristic in forming a prediction. This means that a protected characteristic could be empirically relevant, in that it could provide information on default probability. In this respect, the use of an algorithm significantly alleviates the concern over the arbitrary use of protected characteristics, as an algorithm would not consider a protected characteristic unless it had informational value.

This wedge between what is empirically relevant for an accurate prediction to what is legally permissible to be considered in credit pricing is the core of the concern that a protected characteristic can be discovered through other information. It presents the central tension between a statistical technology that is focused on empirical accuracy, and the need to comply with legal restrictions that go beyond empirical relevance.

One reason an algorithm would consider a protected characteristic is that the characteristic correlates with some other unobservable characteristic that is of true interest.[137] In such a case, the protected characteristic is not of interest in and of itself, but rather it correlates with other factors that are related to the outcome that are imperfectly observed by the algorithm. For example, an algorithm may use "race" in predicting an outcome because it correlates with other characteristics that the algorithm cannot observe directly such as wealth or access to credit, which in turn affect default risk.[138] Economists often describe this situation as "statistical discrimination" because race is used to infer other information.[139]

[137] It is not possible to perfectly establish whether a characteristic is causal or not of an outcome. However, a common intuition is that race, to the extent that it is predictive of default, is not causal in of itself but because it serves as a proxy for another characteristic. However, race might be causal of some other characteristic that is causal of default. For example, animus could cause minority workers to be excluded from the labor force, and employment was causal of default. Nonetheless it is employment itself that is the object of interest and not race. Intuitively, therefore, an algorithm could use race as a proxy for employment.

[138] This example closely relates to the category of proxy discrimination that Prince and Schwarcz call "Omitted Variable Unintentional Proxy Discrimination". *See* Anya Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data* 23 (2019), https://papers.ssrn.com/abstract=3347959.

[139] Statistical discrimination is the use of protected characteristics to form accurate beliefs about unobservable characteristics. *See* Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972). Kenneth Arrow, *The Theory of*

Legally, "statistical discrimination" is likely to be prohibited by fair lending's disparate treatment doctrine.[140] The direct conditioning on a protected characteristic, even if it merely served as a proxy for another characteristic, is nonetheless pricing differently for protected groups. Moreover, the fact that the use of the protected characteristic can be supported empirically, in that it increases prediction accuracy, would not serve as a defense.[141]

Protected characteristics are also sometimes of direct interest. When a protected characteristic is causal or closely related to the outcome of interest, then an algorithm has a direct interest in recovering the characteristic.[142] The protected characteristic is not substituting for an unobservable variable. In these cases, the wedge between what is empirically relevant and legally permissible is the greatest.

Discrimination law often prohibits consideration of a characteristic that is of direct empirical relevance.[143] In the case of ECOA, the Act prohibits discrimination on grounds that are possibly causal of default.[144] ECOA prohibits discrimination based on age or whether a borrower receives his or her income from public assistance programs, as discussed above. It is plausible that these two factors affect a borrower's predicted future income and therefore closely relate to default risk. Yet, ECOA prohibits their consideration. Similarly, ECOA prohibits discrimination based on marital status, although this could affect the likelihood of default when a mortgage is

*Discrimination* (1973). There is more nuance to the types of correlations that an algorithm might want to discover than is presented here. See Prince & Schwarcz, *supra* note 135, for other types of examples. *See also* Deborah Hellman, *Measuring Algorithmic Fairness* (2019), https://papers.ssrn.com/abstract=3418528.

[140] This might be not true under the interpretation of "discriminatory intent" that is concerned primarily with animus and not with classification. For a discussion of the different types of discriminatory intent in the context of Equal Protection, see: Huq, *supra* note 132, at 1249.

[141] See Section 1.3.

[142] This is closely related to what is often referred to as "rational discrimination" that often comes up in the context of disability insurance. *See* Samuel R. Bagenstos, *Rational Discrimination, Accomodation, and the Politics of (Disability) Civil Rights*, VA. L. REV. 825 (2003).

[143] *See* Prince & Schwarcz, *supra* note 135, at 4 (discussing the example of life insurance and genetic information). Clearly, genetic information is highly relevant to the cost of insuring an individual, and yet the insurer is forbidden from considering this information.

[144] Although not the mainstream view of ECOA, there is an interpretation that ECOA is only really meant to address "arbitrary" consideration of these factors. *See* Taylor, *supra* note 31.. For an economics perspective on this type of discrimination, *see* J. Aislinn Bohren et al., *Inaccurate Statistical Discrimination*, Working Paper 25935 (National Bureau of Economic Research), Jun. 2019 (proposing a new category of discrimination "inaccurate statistical discrimination," which is a type of statistical discrimination that is based on inaccurate beliefs).

underwater.[145]

To demonstrate the ability to recover a protected characteristic from other information, I use the Boston Fed HMDA data set to predict two protected characteristic, "age" and "marital status." Each time I exclude the protected characteristic while predicting this characteristic from the remaining variables.

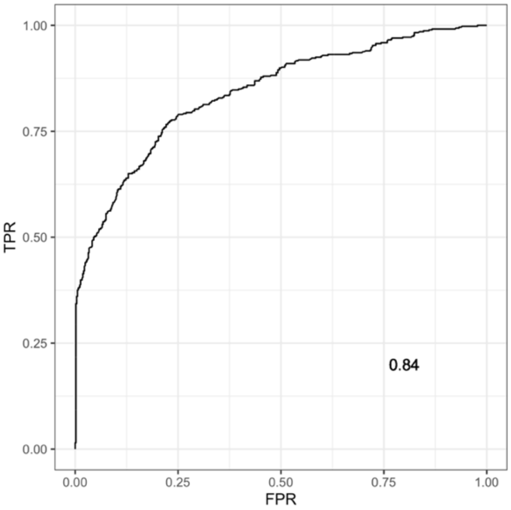


*Figure 4: ROC curve for prediction of borrower "age." The "age" variable in the Boston Fed HMDA dataset is not a continuous variable of age but rather an indicator of whether the applicant's age is above or below the median in the Boston Metropolitan Statistical Area. The ROC curve plots the true-positive-rate and false-negative-rate for different cut off rules. The number in the lower right corner is the Area Under Curve (AUC).*

Figure 4 demonstrates the ability to predict "age" with a high level of accuracy from the other HMDA dataset variables. Figure 17 in Appendix B shows similar analysis for predicting "marital status" from the other HMDA variables. The two figures are a representation of how accurately I was able to predict a borrower's age and marital status from the HMDA dataset in the form of a receiver operating curve (ROC) curve. The number on the bottom

[145] There are examples in other domains of discrimination law prohibiting the consideration of causal characteristics. For example, many states prohibit the consideration of gender in setting life and health insurance premiums. *See* Ronen Avraham et al., *Understanding Insurance Antidiscrimination Law*, 87 S. CAL. L. REV. 195 (2013–2014). However, gender is clearly empirically relevant to the cost of insuring an individual. Nonetheless, the law in these states prohibits insurers from charging different rates to men and women. In this case, gender is morally irrelevant even when it is empirically relevant. Another important example are laws that prohibit discrimination of costs of annuities based on gender, such as the EU Directive on insurance pricing. *See* Council Directive 2004/113/EC (Dec. 13, 2004) (covering insurance in general and not only annuities). A person's gender will highly affect the costs of providing an annuity, given that women often live longer than men.

right corner is the Area(s) Under Curve (AUC), which measures the prediction accuracy. Appendix B provides more details on how the ROC curve is plotted and how it should be interpreted. Intuitively, because the ROC curves are close to the upper left corner, and the AUC are high (0.84 for age and 0.9 for marital status), we are able to predict these protected characteristics with a high level of accuracy.

This prediction shows that the formal exclusion of a protected characteristic may be meaningless with respect to the ability of an algorithm to actually use the characteristics. In reality, the results above are a lower bound with what is feasible with big data and machine learning. As discussed above in Section 2.3.1., the variables in the Boston Fed HMDA dataset are primarily more traditional pricing variables and are unlikely to represent the richness of data available to algorithmic lenders. With nontraditional data, lenders can potentially recover protected characteristics with greater accuracy. Even if an algorithm does not seek to recover the information—that is, it never tries to derive race or marital status—such characteristics are available to it because they are so embedded into the rest of the data.

The ability to recover a protected characteristic from other information may arguably be less of a concern when the characteristic only serves as a proxy for true characteristic of interest. This is because the protected characteristic was never of interest in and of itself, and therefore a "blind" algorithm will search for proxies for the underlying characteristic of interest rather than attempt to recover the protected characteristic. Moreover, as the data scope and accuracy increase there is no need to use protected characteristics, even if the algorithm was not "blind."

The concern is likely to be much greater when considering protected characteristics of direct interest. In the case of a protected characteristic that is of direct interest, changes in data scope and accuracy may only mean that algorithms will have a better ability to learn and use protected characteristic, even when formally hidden. The wedge between what is empirically relevant and legally permissible never disappears. Eventually this could mean that there is no difference between a "blind" and "aware" algorithm, rendering the exclusion strategy meaningless.

### 3.1.2 Disparities may increase with exclusion

There is an additional reason to be wary of the exclusion of protected characteristics as a way to apply discrimination law in the algorithmic setting. Namely, if we care about price disparities, the inclusion of a protected characteristic, rather than the exclusion, could decrease disparities.[146]

[146] The extent to which disparate impact is concerned in directly reducing outcome disparities is discussed above in Section 1.4, and may depend on what type of reason is

When a characteristic should be interpreted differently for various racial groups, excluding "race" could increase disparities. This is because by excluding the race variable, we are imposing a similar interpretation of a characteristic for both white and non-white applicants. When there are many more whites in a training dataset, which is likely to be the case even in a representative dataset,[147] the prediction will be formed according to the weight attributed to the characteristics for whites. For example, even if the borrower's number of children is predictive of default only for white applicants and not non-white applicants, the algorithm will give the characteristic the same weight for all racial groups when "race" is excluded. This critique is consistent with the growing skepticism among scholars about the usefulness of the wholesale approach of excluding protected characteristics.[148]

Similarly, the inclusion of protected characteristics may also be important in mitigating the harms of "biased measurement" variables. Consider a hypothetical lender that predicts default from an input that suffers from measurement bias. In this example, "ability" is equally distributed across the population  and that higher "ability" people default less, perhaps because their earnings are higher.[149] The characteristic "ability" is not observed by the lender. Instead the lender has information about college attendance, which is correlated with "ability." Assume that racial minorities face discrimination in college applications and are therefore less likely to attend college. In this example, the input "college attendance" suffers from measurement bias because it is a noisier measurement of "ability" for racial minorities.

driving the disparities created by exclusion.

[147] In the 2000 HMDA dataset, for example, black applicants are less that 10% of all applications reported. *See* Federal Financial Institutions Examination Council (FFIEC), Reports – Nationwide Summary Statistics for 2000 HMDA Data, Factsheet (July 2001), Table 2, https://www.ffiec.gov/hmcrpr/hm00table2.pdf. It's important to note that the Boston HMDA is skewed to overrepresent minorities relative to their share amongst mortgage applicants.

[148] *See* Jon Kleinberg et al., *Algorithmic Fairness*, 108 AEA PAPERS AND PROCEEDINGS 22 (2018). *See also* Kim, *supra* note 7, at 904. ("Thus, a blanket prohibition on the explicit use of race or other prohibited characteristics does not avoid, and may even worsen, the discriminatory impact of relying on a data model"). *See also* Melissa Hamilton, *The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, AM. CRIM. L. REV. 1553 (2019) (considering the performance of the COMPAS risk assessment on Hispanics). *See* Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAVIORAL SCIENCES & THE LAW 145 (2019), for a discussion with respect to COMPAS and gender.

[149] This could be because higher ability borrowers are likely to have higher future earnings, and therefore have a lower risk prediction.

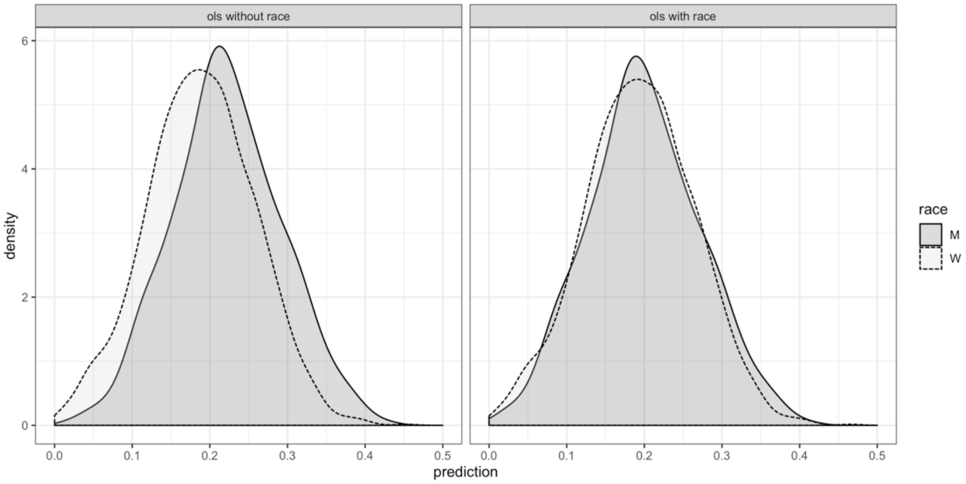


*Figure 5: Simulated example of default risk using a "race blind" algorithm (on the left) and a "race aware" algorithm (on the right). The graphs plot the distribution risk for white (W) and minority (M) borrowers using an OLS regression.*

Figure 5 shows that in my simulated example, predicting default risk only from college attendance results in non-white borrowers have a higher default probability.[150] This can be seen in the graph to the left in which the distribution for non-white and Hispanic borrowers ("M") is shifted to the right, meaning there are more borrowers with a higher default risk. When default prediction includes the race variable (graph on the right), the default risk of white and non-white is more similar. This is because a race-aware algorithm knows to treat "college attendance" differently for white versus non-white borrowers.

The conclusion is not that including protected characteristics always reduces disparity. In fact, this is unlikely to be true when a protected characteristic has a direct relationship to the outcome of interest.[151] Rather, the argument is that it is difficult to determine *a priori* what effect the inclusion of a protected characteristic might have. Therefore, we should be wary of treating the exclusion of protected characteristics as a means to reduce disparity.

It is questionable whether the inclusion of a protected characteristic for the purpose of reducing disparities, would be legal. As discussed above, many approaches to discrimination and algorithms assume that protected characteristics must be excluded.[152] Recently, however, several scholars have

[150] This example uses an OLS regression and not a machine learning algorithm. For the purposes of this highly stylized example, the OLS regression is sufficient.

[151] In Section 3.4. I present a case in which the lasso regression puts weight on the input "race," meaning that it predicts that white borrowers are less likely to default.

[152] *See* MacCarthy, *supra* note 74, at 72 ("these cases do suggest that the use of group

suggested that discrimination law's position on the consideration of a protected characteristic may be more nuanced. Deborah Hellman argues that separately considering which inputs are predictive of future criminal activity may not in fact constitute disparate treatment.[153]

### *3.2 Excluding proxies for protected characteristics*

A second approach to applying discrimination law to algorithmic pricing is expanding the prohibited inputs to also include "proxies" for protected characteristics. The discussion in the previous Section demonstrates that exclusion of the protected characteristic may be meaningless if an algorithm can use proxies for those characteristics. If there are proxies for protected characteristic, a natural response is to exclude these proxies as well. This second strategy can therefore be thought of as an expansion of the first strategy. In traditional fair lending, this strategy is sometimes adopted by excluding salient examples of proxies, such as zip codes.[154]

This approach in the algorithmic context was recently articulated by the Department of Housing and Urban Development (HUD) in a proposed rule from August 19, 2019. The Proposed Rule revises §100.500 of HUD's 2013 Rule[155] with respect to its interpretation of the disparate impact doctrine.

variables in algorithms would be subject to strict scrutiny, even if their purpose is to reduce group disparities.")

[153] See Hellman, *supra* note 136, at 38 for a discussion of the possibility of separately considering which inputs are predictive for different racial groups. Although Hellman discusses this proposal in the context of Constitutional discrimination, it is closely related to fair lending disparate impact. Hellman discusses a case in which "an algorithm might use race *within* the algorithm to determine what other traits would be used to predict the target variable" and suggests that perhaps strict scrutiny may not apply because a separate algorithm is treating the groups equally in the sense that "only relevant information is utilized" (*Id.* at 40). As a technical matter it is unlikely that the proposition is functionally different from including race in an algorithm, even if it might be treated differently legally.

In general, the fact that the explicit consideration of a protected characteristic can reduce disparities suggests a possible tension between disparate treatment and disparate impact. The tension between the requirement to ignore forbidden characteristics, yet assure that policies do not create a disparate impact, thereby requiring a consideration of people's forbidden characteristics, has recently been debated in the context of Ricci v. DeStefano, 557 U.S. 557 (2009) (in which a promotion test was invalidated by an employer because of the concern that promotion based on the test would trigger disparate impact). *See* Primus, *supra* note 77. *See also* Hellman, *supra* note 136, at 47; Kim, *supra* note 7, at 925.

[154] The use of a zip code in credit pricing could also trigger a claim of "redlining," in which a lender avoids extending credit to borrowers who live in neighborhoods with higher minority populations. *See* Alex Gano, *Disparate Impact and Mortgage Lending: A Beginner's Guide*, U. COLO. L. REV. 1109, 1136 (2017) (discussing redlining).

[155] Implementation of the Fair Housing Act′s Disparate Impact Standard, 84 Fed. Reg. 42854 § 100 (U.S. Aug. 19, 2019).

Section (c)(2) relates to a case in which a plaintiff is challenging a defendant's use of model with a discriminatory effect and lays out the defenses on which a defendant can rely. According to (c)(2)(i), a defendant can rebut a claim of discrimination by showing that "none of the factors used in the algorithm rely in any material part on factors which are substitutes or close proxies for protected classes under the Fair Housing Act."[156] Therefore a defendant can negate a claim's disparate impact by showing that a risk assessment algorithm excludes proxies for protected characteristics.

Several scholars have also proposed preventing algorithms from using variables that are highly correlated with a protected characteristic. For example, Hurley and Adebayo propose a model bill — the Fairness and Transparency in Credit Scoring Act – that contains this type of provision.[157] The model bill requires credit scores "not treat as significant any data points or combinations of data points that are highly correlated to immutable characteristics."[158]

### 3.2.1 What is a "proxy"?

The expansion of input exclusion, beyond protected characteristics themselves, requires a clear articulation of the criteria for exclusion. Prior work has suggested that a proxy be defined as an input that is 1) highly correlated with the protected characteristic[159] and/or 2) does not contain informational value beyond its use as a proxy.[160] Hurley and Adebayo focus

[156] According to (c)(2)(i), the defendant must also show that the model is predictive of credit risk or "other similar valid objective". Section (c)(2)(i) contains similar language but relates to showing that a third party has established that it does not rely on proxies or close substitutes. This is the third defense available to the defendant in the proposed rule. *See Id.* sec. 100 33.HUD Proposed Rule 2019, at 33.

[157] Hurley & Adebayo, *supra* note 7, at 196.

[158] *Id.* at 206. The "immutable characteristics" that the provision is referring to are race, color, gender, sexual orientation, national origin, and age. There is a similar provision for marital status, religious beliefs or political affiliations.

[159] *See* Charles River Associates, *Evaluating the Fair Lending Risk of Credit Scoring Models* , 3 (2014), http://www.crai.com/sites/default/files/publications/FE- Insights-Fair-lending-risk-credit-scoring-models-0214.pdf ("Ostensibly neutral variables that predict credit risk may nevertheless present disparate impact risk on a prohibited basis if they are so highly correlated with a legally protected demographic characteristic that they effectively act as a substitute for that characteristic.")

[160] *See* FED. TRADE COMM'N, REPORT ON BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION?, 67 (Jan. 2016). Here I focus on two possible definitions of a proxy, however, this is not the only way to define a proxy we may want to exclude. A somewhat different approach to the expansion of excluded features, looks to other factors beyond correlation to a protected characteristic. For example, Sunstein in a recent paper has suggested: "Difficult problems are also presented if an algorithm uses a factor that is in some sense an outgrowth of discrimination". See Sunstein, *supra* note 78, at 8. In the context of credit pricing this

on variables that highly correlate with protected characteristics.[161] Other approaches require something beyond a correlation, such as requiring that the variable not contain much information relevant to the outcome of interest.

Identifying characteristics that contain little or no informational value beyond their use as a substitute for a protected characteristic,[162] is difficult to implement in practice. The problem is that we do not have a good understanding of the "model" of default and which variables are causal of default. Even if we knew the true model of default, we would not necessarily know how other variables relate to those causal variables. In some cases, intuition is used to replace empirical understanding of how variables relate to default, by attempting to tell a plausible story of whether an input that correlates with race does or does not contain information related to default, beyond its use as a proxy.[163] However, even zip codes, which have become the archetype of a proxy for race, are likely to contain informational value relevant to default risk.[164]

A further difficulty is that many variables can both be an indicator of a protected characteristic but also independently contain information relevant to the outcome of interest. In most cases, we are not able to isolate the

would mean excluding many of the fundamental features used to price credit, even today, such as credit scores and wealth. This approach is likely to be based on the argument that disparate impact grows out of the duty to avoid compounding injustice as argued by Deborah Hellman. *See* Hellman, *supra* note 52, *available at* https://papers.ssrn.com/abstract=3033864. While Hellman's focus is the justification of the disparate impact and indirect discrimination doctrines as a moral wrong, it also suggests guidelines as to what conduct should be prohibited. In our case, if variables used by an algorithm were themselves a result of discrimination, even if this discrimination did not originate in any action on the part of the lender, the active use of these variables in setting the price of credit could be a compounding of the initial injustice towards minority groups.

[161] *See* Hurley & Adebayo, *supra* note 7, at 200. ("The FaTCSA addresses the potential problem of proxy-based discrimination by prohibiting the use of models that "treat as significant any data points or combinations of data points that are highly correlated" to sensitive characteristics and affiliations.")

[162] *See* Prince & Schwarcz, *supra* note 135, at 4 (defining "proxy discrimination" as the use of an input that not only correlates with a protected characteristic but "that the usefulness to the discriminator of the facially-neutral practice derives, at least in part, from the very fact that it produces a disparate impact").

[163] *See* Yu et al., *supra* note 62, at 29 (providing s an example of this type of intuitive argument). According to the NCLC, to rely on a business necessary justification, the lender would need to show the connection between the input and credit risk. For example, "[t]here is an understandable connection between timely repayment of past obligations and the likelihood of timely repayment of future obligations, so a "demonstrable relationship" argument can be easily made." *See Id.*

[164] For example, the real estate fluctuations in a particular area. *See* Erik Hurst et al., *Regional Redistribution through the US Mortgage Market*, 106 AM. ECON. REV. 2982 (2016) (documenting large regional variation in default risk, despite the uniform pricing of Government Sponsored Enterprises (GSEs) across regions.)

component of a variable that is merely a proxy for a protected characteristic and the component that contains independent information. In Section 3.4, I discuss a statistical approach that seeks to isolate the proxy component from variables that correlate with protected characteristics. I argue, however, that this method is inappropriate for the machine learning context.

### 3.2.2 Identifying proxies

Focusing on proxies for protected characteristics defined as inputs that highly correlate with those characteristics, is also unlikely to guarantee that protected characteristics are not used by an algorithm. This is because in the big data context, considering how individual inputs correlate with protected characteristics does not fully capture the complex interactions among inputs. Therefore, expanding the excluded characteristics to inputs that correlate with protected characteristic will only have a limited effect in reducing disparities, if any at all.

Figure 6 shows how an algorithm may produce different risk predictions for white and non-white borrowers even when excluding inputs that highly correlate with race.[165] The graph on the left shows the distribution of default risk for white and non-white borrowers when the algorithm does not use "race," and the graph on the right shows the default risk when the algorithm excludes "race" and the ten variables that correlated most with "race." One way to consider the disparities between the groups is by the gap between the vertical lines, which are the median predictions for white and non-white borrowers. Although the difference in median risk prediction for white and non-white borrowers is lower in the graph on the right,[166] the disparities between the groups continue to persist. This is because the individual correlations of variables with a protected characteristic do not capture the full range of how variables correlate and interact.[167]

[165] This figure is similar to the figure produced in Gillis & Spiess, *supra* note 25, at 469. One important difference is that this figure does not contain a separate distribution for black and non-white Hispanic borrowers but rather collapses them into one category of non-white borrowers.

[166] This is partially because the distributions have altogether been condensed as a result of the use of fewer variables to distinguish between borrowers. See Section 2.3.1.

[167] It is important to note that this demonstration is somewhat of a lower bound of how information on protected characteristics is embedded in other inputs with big data. As already mentioned the number of variables and types of data used in the simulation example are similar to more traditional credit pricing since it does not include non-traditional data such as consumer purchasing and payment behavior. When the amount of data and type of data expands, this problem is likely to be more severe given the complex relationship between different characteristics and the ubiquity of correlations.

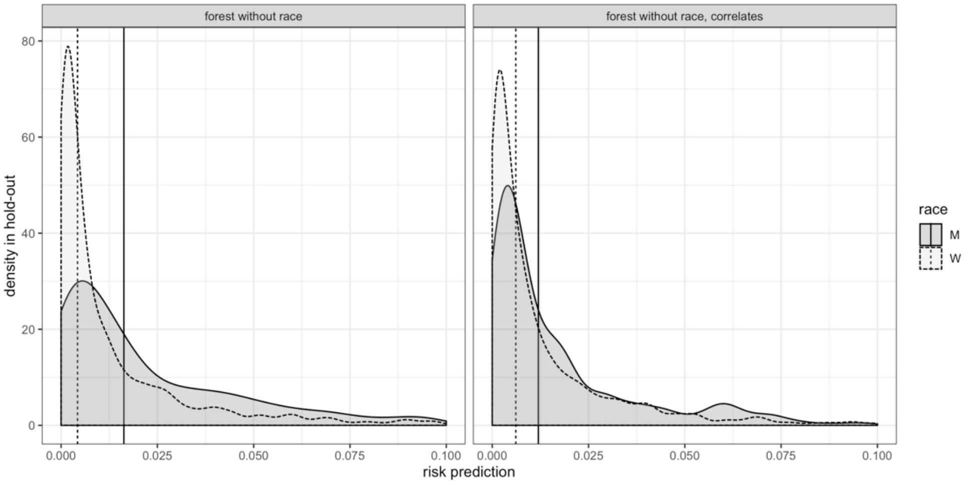


*Figure 6: Distribution of risk predictions across groups for different inputs. The graph on the left shows the risk predictions when using all HMDA inputs other than race, plotted separately for the non-Hispanic white (W) and non-white (M) borrowers in the holdout group. The graph on the right shows the risk predictions when using HMDA inputs other than race and ten variables with the highest correlation to race. Also in this graph, the predictions are plotted separately for white and non-white borrowers. The vertical lines are the median risk prediction for each racial group. The ZCTA populations are reweighted to account for the oversampling of black borrowers in the Boston Fed HMDA dataset.*

Furthermore, classic examples of "proxies," such as zip codes, may be less indicative of race than other variables used by lenders. To demonstrate this, I consider how accurately I am able to predict whether a borrower is black from the Boston Fed HMDA dataset, which contains mostly classic variables used by lenders. I then compare this to how accurately I am able to predict whether a borrower is black from Zip Code Tabulation Areas (ZCTAs), the Census equivalent of zip codes,[168] for the Boston Metropolitan Statistical Area.[169]

[168] The reason that the Census uses ZCTAs and not zip codes is that zip codes often cross state, county, census tract, and census block group and therefore could not be used as a defined area in the Census.

[169] This is the geography that the HMDA dataset is based on. The populations in the ZCTAs have been reweighted to reflect the over sampling of blacks in the Boston Fed HMDA dataset. See the description of who was included in the Boston Fed HMDA dataset in: Munnell et al., *supra* note 54, at 26.

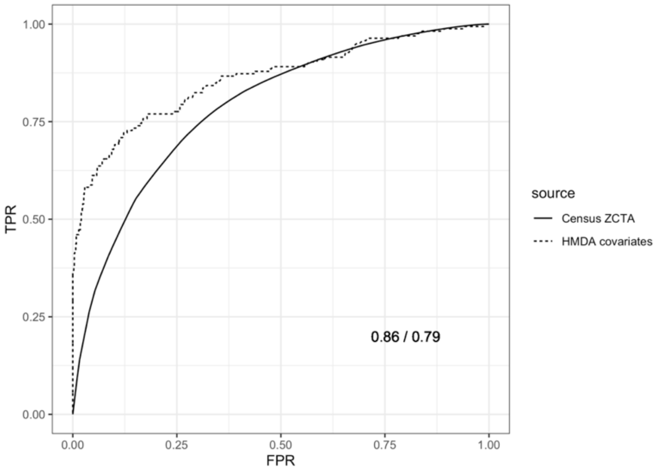


*Figure 7: ROC curve for prediction of "black" using HMDA covariates and using Census ZCTA*

Figure 7 shows that the prediction of whether a borrower is black is more accurate using the HMDA dataset than ZCTAs. For nearly all the distribution, the curve of the HMDA covariates is above the Census ZCTA curve. This means that for nearly any cut-off rule with respect to predicting whether a borrower is black, the HMDA covariates produce a more accurate prediction (meaning that the "true positive rate" is higher and the "false positive rate" is lower. See Appendix B). This can be also seen by comparing the area under a curve for the Census ZCTA (0.86) and the HMDA covariates (0.79). The example demonstrates how common intuitions about which variables serve as proxies might be misleading. If what we are truly interested in is the ability to recover a person's protected characteristics, intuitive judgments are insufficient to determine which features to exclude. Features that intuitively feel like proxies might correlate less than might features that do not feel like proxies.

The final reason to be wary of this exclusion approach is that most inputs used to price credit, even in the traditional context, correlate with a protected characteristic.[170] Restricting the use of variables that correlate with protected characteristics reduces lender's ability to accurately predict default risk and personalize pricing accordingly.[171] The recent HUD Proposed Rule shows that not appreciating the prevalence of correlations can create confused and incoherent policy. According to the Proposed Rule, after a defendant shows they did not use a protected characteristic a plaintiff can undermine a defense of a model by showing "that the defendant's analysis is somehow flawed, such as by showing that a factor used in the model is *correlated with a*

[170] *See supra* Section 1.2.

[171] *See supra* Section 1.1.

*protected class*."[172] This would likely mean that a lender using a risk model would never be able to rely on the defense laid out in the Proposed Rule.

In summary, while the attempt to exclude proxies in addition to protected characteristics is intuitively appealing, there are practical challenges endemic to defining and detecting proxies. Correlation to a protected characteristic does not fully capture the extent to which variables can be used as a substitute for a protected characteristic. Moreover, variables that correlate with race form the core of even traditional credit pricing. Finally, input exclusion comes at the price of prediction accuracy, which may hurt vulnerable populations.

### *3.3 Restricting algorithm to predetermined set of variables*

A third approach of restricting the inputs of an algorithm to inputs that are pre-approved. This approach was recently proposed by Prince and Schwarcz in the context of insurance: "Instead of allowing use of any variable not barred, as in the traditional antidiscrimination model, actors can only use pre-approved variables."[173] This approach is similar to the first two in that it limits the inputs into an algorithm. However, instead of focusing on excluding variables that are impermissible, this approach seeks to define what variables are permissible.

Predetermining permissible variables could either be implemented by a regulator or using an internal screening process by the lender to decide which variables can be used. Approaches requiring that lenders show that inputs are "relevant" or "causal" to the outcome are likely to amount to a form of predetermining permissible inputs.[174] If lenders must show that a lending decision relies on inputs that are logically related to outcome, they will need to exclude other variables. Predetermining which variables are related to the outcome will allow lenders to meet this burden.[175]

[172] Implementation of the Fair Housing Act′s Disparate Impact Standard, 84 Fed. Reg. 42854 § 100 (U.S. Aug. 19, 2019), at 20-21.

[173] *See* Prince & Schwarcz, *supra* note 135, at 54.

[174] This is another one of the proposals in *Id.* at 59 ("One possible solution is to require those employing algorithms to establish the potential causal connections between the variables utilized and the desired outcome.")

[175] *See* Grimmelmann & Westreich, *supra* note 39, at 176 ("Where a model has a disparate impact, our test in effect requires an employer to explain why its model is not just a mathematically sophisticated proxy for a protected characteristic.") *See also* Kim, *supra* note 7, at 921 ("The existence of a statistical correlation should not be sufficient. Instead, because the employer's justification for using an algorithm amounts to a claim that it actually predicts something relevant to the job, the employer should carry the burden of demonstrating that statistical bias does not plague the underlying model.")

### 3.3.1 Entrenching disadvantage

The main challenge with this approach is defining which variables are permissible. This depends on what the restriction is meant to achieve. If the goal was simply to only use inputs relevant to default in that they predict default, there is no reason to restrict input at all. The algorithm, rather than a human, is likely to be the best judge of whether an input predicts default. Therefore, to the extent that what we are trying to achieve is a more accurate prediction, restricting inputs to a predetermined list is a misguided strategy.

Limiting the algorithm to characteristics that are used in traditional credit pricing, such as FICO scores or a borrower's income, undermines the benefits of big data and machine learning in extending access to credit. The use of nontraditional data can expand credit to people without sufficient credit history, so excluding this data maintains their status as "credit invisibles."[176] Moreover, when FICO scores, for example, only measure certain indicators of the likelihood of meeting obligations on time, big data can mitigate this "bias measurement" by expanding the data used to predict default. By restricting algorithms to classic characteristics, these benefits cannot be captured, potentially entrenching disadvantage for certain populations.

An alternative approach is to allow for the use of characteristics that are not classic credit pricing variables but to restrict inputs to variables that are closely related to models of repayment. This is the approach advanced by Bartlett et al. who argue that in determining what variables are legitimate, "one can write down a life-cycle model in which cash flow for repayments emerge from the current borrowing position (debt), cost of borrowing (credit score), income (in levels, growth, and risk), wealth, and regular expense levels (cost of living measures)."[177] When a variable correlates with a protected characteristic, it can only be used to the extent that it relates to the life-cycle structural model of debt repayment. Similar to the approach in Section 3.2.2., the position advanced by Bartlett et al. seeks to prevent an algorithm from using proxies for protected characteristics.

The success of this approach relies on human intuitions to accurately determine how inputs relate to the "life-cycle" model of repayment. In reality, we do not always directly observe the variables of the structural model of repayment and rely instead on noisy substitutes for the variables of the model. With high-dimensional data wherein correlations are ubiquitous, we can lose

[176] *See* Patrice Ficklin and Paul Watkins, *Consumer Financial Protection Bureau, An Update on Credit Access and the Bureau's First No-Action Letter* (Aug. 6, 2019), https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/.

[177] *See* Bartlett et al., *supra* note 45, at 26.

any direct sense of how inputs relate to the structural model.[178] For example, a person's wealth is typically unknown, and so in order to infer wealth, we may need to rely on proxies or correlates of wealth. As the complexity of the structural model and the list of inputs that can be used to infer the variables in the structural model increase, the dependence on human intuition in determining what variables feel related to a characteristic in the model, such as wealth, becomes particularly weak.

Furthermore, there is little reason to believe that we know the true structural model of repayment. Structural models are useful when engaging in empirical research and need to estimate the effect of different changes on an outcome of interest. They also provide discipline in interpreting empirical results. However, they are a far cry from a true reflection of the actual causal relationships that exist in the world. For example, the literature on micro-financing in development economics points to a number of factors that might affect default rates, not captured by Bartlett et al.'s "life-cycle" model. For example, public repayment of loans may lead to lower default rates due to reputation concern.[179] If we rely on a structural model to dictate what can and cannot be used as an input, it is a problem when the structural model is incomplete. This is particularly worrisome if the mode is more incomplete for protected groups.[180]

### 3.3.2 High cost to prediction accuracy

More generally, limiting the inputs an algorithm can use to form a prediction of default could lead to less accurate predictions, the main benefit of machine learning pricing. This increase in accuracy can stem from the various changes machine learning pricing brings about. First, the focus on an automated system of prediction versus human prediction could add accuracy to the prediction.[181] Second, machine learning versus other statistical

[178] Bartlett et al., avoid this problem by focusing on a context in which mortgage lenders do not face default risk so that differential pricing cannot be explained by default prediction altogether. *See* Bartlett et al., *supra* note 45.

[179] *See, e.g.*, Abhijit Vinayak Banerjee, *Microcredit Under the Microscope: What Have We Learned in the Past Two Decades, and What Do We Need to Know?*, 5 ANN. REV. ECON. 487 (2013) (discussion of the various theories and empirical evidence on microlending).

[180] For example, suppose creditworthiness is affected by social attitudes to foreclosure, but that these social norms were not part of the structural model of repayment. If minorities are more likely to belong to such communities you could be excluding measures of creditworthiness more reflective of minority communities, creating a bias against minorities. Essentially in trying to restrict the bias concerns of big data, this type of restriction may in fact increase bias.

[181] *See* Lkhagvadorj Munkhdalai et al., *An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments*, 11 SUSTAINABILITY 699 (2019).(comparing a human-expert based model of prediction, FICO, with a machine learning prediction, finding

methods, such as linear regressions, allows for greater flexibility in forming a prediction versus a linear regression, for example, and also increases accuracy.[182] Finally, the expansion of the type and number of inputs considered by an algorithm could increase the accuracy.

To demonstrate how accuracy can change when reducing the inputs of an algorithm, I return to my hypothetical lender. I compare two algorithms, one that uses the full set of inputs (other than race) and another that is only limited to a small subset of variables.[183]

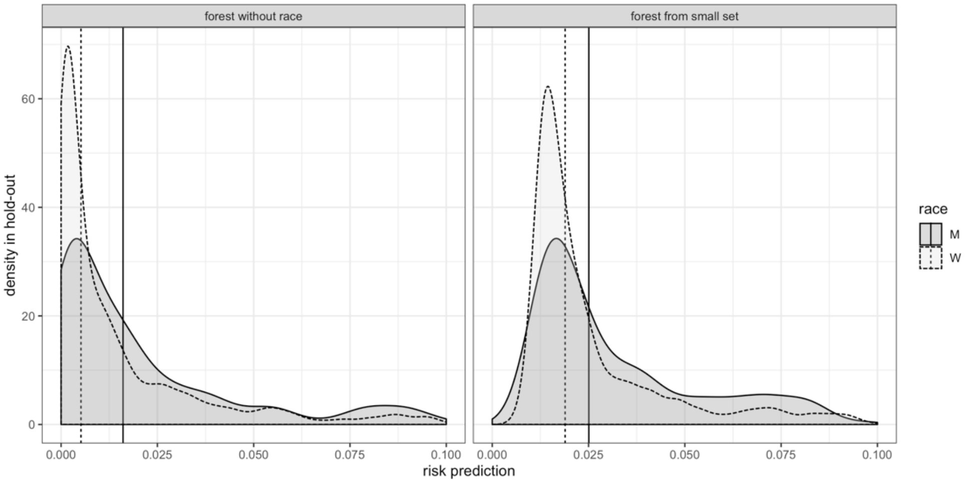


*Figure 8: Distribution of risk predictions. The graph on the left shows the risk predictions using a random forest with the full set of inputs (other than race). The graph on the right shows the risk predictions using a random forest with a small set of more traditional credit inputs. Both graphs seprate the risk prediction for non-Hispanic white borrowers (W) and non-white borrowers (M). The vertical lines are the median for each group of borrowers.*

Figure 8 shows that when using a smaller set of inputs, the risk distribution changes. The risk distribution becomes more condensed when predicting from a smaller set of inputs (graph on the right). This is because using fewer variables means that there are less variables to distinguish between people so that the distribution is more concentrated around the mean.

To demonstrate the change in prediction accuracy, I plot the receiving operator characteristic (ROC) curve corresponding to the two distributions in Figure 8.[184]

that the non-human expert prediction is superior in predicting default). In the context of bail decisions, see Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237 (2018).

[182] *See supra* Section 2.3.

[183] I use one possible subset, which includes some variables that are typically used to price credit today – income, debt-to-income ratio and characteristics of the loan.

[184] *See* Appendix B on ROC curves.

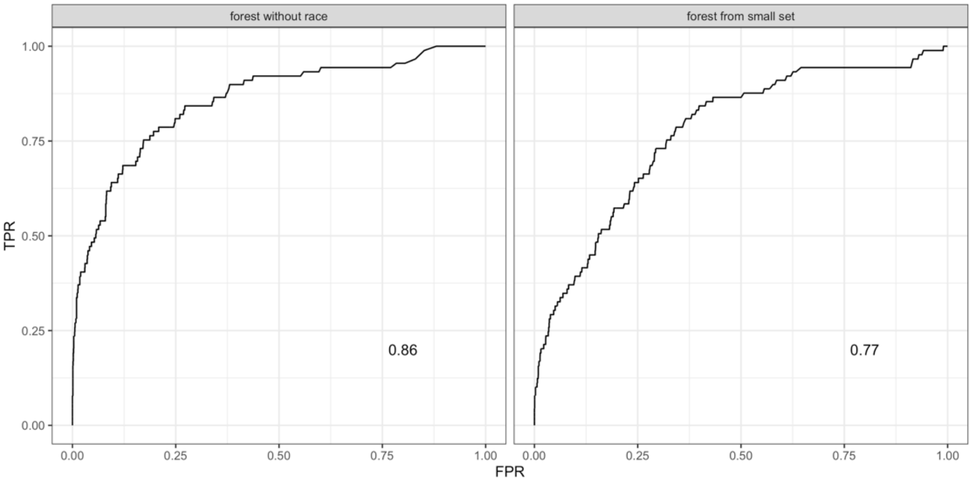


*Figure 9: ROC curves corresponding to risk distributions in Figure 8. The ROC curve on the left shows the accuracy of the risk predictions using a random forest with the full set of inputs (other than race). The graph on the right shows the accuracy of the risk predictions using a random forest with a small set of more traditional credit inputs. The number on the bottom right corner is the Area Under Curve (AUC).*

Figure 9 shows that the prediction based on the larger set of inputs is more accurate. This can be seen from the curve in the left graph being closer to the upper left corner and from the AUC in the lower right corner being higher for the prediction using the full set of inputs.

The potential tradeoff between different notions of fairness and accuracy has been previously noted and is also relevant when trying to limit the inputs into an algorithm.[185] However, as argued in the previous sections, the proposal to limit an algorithm to inputs that seem intuitively relevant to the outcome of interest face further challenges, as the decision could be arbitrary and even undermine the benefits of big data in mitigating the harm of measurement bias. Therefore, the restriction of inputs may not even be a case of trading off accuracy for fairness but could in fact reduce accuracy *and* fairness. Moreover, as discussed in Section 1.1, reduced accuracy could hurt vulnerable borrowers who are excluded altogether from credit markets when the lender cannot accurately price risk.[186]

[185] *See* Corbett-Davies et al., *supra* note 45. Also see: Geoff Pleiss et al., *On Fairness and Calibration*, ARXIV:1709.02012 [CS, STAT] (2017). *See also* Prince & Schwarcz, *supra* note 135, at 53 ("Given that the goal of algorithms is to ferret out the most efficient predictors of a specified outcome, any changes to this system will naturally introduce inefficiencies. However, society cannot just put blinders on and argue that algorithms should be allowed to be as efficient as possible without intervention; the absence of intervention will result in proxy discrimination. If, for efficiency's sake, no solution is adopted, it must be acknowledged that this comes at the expense of the goals of anti-discrimination laws.")

[186] The need to be sensitive on who bears the burden of more or less accurate predictions has been discussed by Huq in the context of criminal justice. Huq argues that racial equity

### *3.4 Orthogonalizing inputs*

A fourth approach to applying discrimination law to the algorithmic context focuses on statistical methods of "orthogonalizing" inputs. Orthogonalization in this context means that inputs that correlate with a protected characteristic do not serve as proxies for protected characteristics. This approach, recently proposed by Prince & Schwarcz[187] and fully developed by Dobbie & Yang[188] attempts to transform the inputs into an algorithm in a way that reduces or eliminates bias. Although neither of these proposals deals directly with fair lending, each one's analysis is highly relevant for fair lending.

This approach is meant to address problems that arise when a variable that correlates with a protected characteristic plays a dual role. On the one hand, a variable that correlates with race, for example, may provide important information for the outcome of interest, such as when default is predicted using a borrower's income. On the other hand, a variable that correlates with race could also serves as a proxy for race.

An important contribution of this approach is that it highlights the limits of excluding a protected characteristic. Dobbie & Yang demonstrate how when a protected characteristic is excluded, the coefficients of inputs that correlate with the protected characteristic will partially reflect the omitted protected characteristic.[189] Therefore, the inputs that correlate with the protected characteristics serve as "proxies."

This statistical approach separates between the "training" and "screening" stages of an algorithm.[190] Focusing on the example of race, in the training stage, the algorithm is "race aware" in the sense that the algorithm uses "race" as one of its inputs. This produces an estimate of the weight given to race in forming the prediction. However, in the screening stage, meaning the stage in which the prediction is applied to a particular person, the algorithm is unaware of a borrower's race. This means that even though "race" was used

requires to consider who bears the cost of algorithmic errors in determining how to apply notions of fairness. See Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2018–2019).

[187] *See* Prince & Schwarcz, *supra* note 135, at 57. They propose this method in the context of insurance.

[188] *See* Dobbie & Yang, *supra* note 132. Dobbie &Yang discuss their proposal in the context of the Equal Protection. This paper, as well as Prince and Schwarcz's paper, base their proposals on an economics paper from 2011. *See* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AMERICAN ECONOMIC JOURNAL 206 (2011).

[189] This is typically considered "omitted variable bias".

[190] See discussion of this separation in Kleinberg et al., *supra* note 36, at 20–21.

to train the algorithm, formally there is no differential treatment based on race.[191] Intuitively, this method alleviates concerns over the use of proxies because it is able to subtract the pure effect of "race" on the prediction.[192]

How would this framework apply in the context of machine learning? In their paper, Dobbie & Yang apply this method to an OLS, or linear, regression and do not demonstrate their method in the machine learning context.[193] Prince & Schwarcz provide a general discussion of the method in the context of artificial intelligence.[194] In order to evaluate the method in the machine learning setting, I use a lasso (least absolute shrinkage and selection operator),[195] which like an OLS regression produces a function of variables $x$ and estimators $\beta$. Therefore, we may be able to apply this method by substituting "mean race" ($\bar{X}^{race}$) when the lasso selects the race variable.

### 3.4.1 Instability of machine learning selection

Applying the orthogonalization method to the machine learning context creates practical and conceptual difficulties. Practically, the variable selection

[191] In spirit, this method is similar to an approach in the computer science and statistics literature known as "disparate learning processes" (DLP). In DLP, protected characteristics are used in the training stage but not during what I call the "screening" stage. The primary purpose of DLP is reduce disparate outcomes across groups without formally treating people differently based on a protected characteristic. For a skeptic view of this approach see: Zachary Lipton et al., *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 31 8125 (S. Bengio et al. eds., Curran Associates, Inc. 2018).

[192] Formally, Dobbie & Yang consider a case in which we are trying to predict outcome $y_i$ for individual $i$, where there are three types of inputs. There is the protected characteristic, such as race, $X_i^{race}$, inputs that correlate with race $X_i^{corr}$,[192] and inputs that do not correlate with race, $X_i^{noncorr}$. Focusing on a linear regression, at the training stage, the following regression is estimated:

$$y_i = \beta_0 + \beta_1 X_i^{noncorr} + \beta_2 X_i^{corr} + \beta_3 X_i^{race} + \epsilon_i$$

Estimating the model above produces coefficients $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$. Applying this estimated function to predict default for future borrowers is likely to trigger "disparate treatment," as it treats borrowers differently on the basis of race.[192] Therefore, when applying this model to future borrowers, race is set to the mean race ($\bar{X}^{race}$) for all individuals, meaning that the model does not formally distinguish between different racial groups. This formally satisfies the requirement of not discriminating on the basis of race while not allowing correlates to serve as proxies for race.

[193] Pope & Sydnor also focus on the application to an OLS regression but also expand the analysis to a probit regression. *See* Pope & Sydnor, *supra* note 185, at 215.

[194] *See* Prince & Schwarcz, *supra* note 135, at 57 ("For a model produced by an AI, accomplishing this requires including in the training data information on legally prohibited characteristics, such as the race or health status of individuals in the training population.")

[195] The objective of the lasso is to minimize the sum of squares between the true outcome and predicted outcome (like a linear regression), subject to regularization that restricts the magnitude of coefficients.

of the lasso is unstable, and even small amounts of noise lead to different variable selection.[196] Conceptually, a lasso algorithm is not meant to estimate a model, as with an OLS regression, so that it is incorrect to interpret the weights of different variables as reflecting some underlying model, as the orthogonalization method does.[197]

To demonstrate the practical challenges in applying this method to machine learning, I show that whether an algorithm selects and puts weight on race is unstable. To create 10 comparable datasets with slightly different noise, I randomly draw 2,000 observations from my full dataset 10 times. Because these 10 datasets are randomly drawn from the same full dataset, they should roughly be the same, although they are unlikely to be identical. I then fit a lasso regression to each of the 10 training datasets: I let the algorithm choose which of the many characteristics to include in the model. As pointed out above, the advantage of using a lasso regression is that its output looks quite similar to an OLS output in that it produces a function with the variables used to form a prediction and the weights of each variable.

Despite being drawn from the same population, it is not the case that the random sampling leads to identical algorithmic decisions. Figure 10 plots the weights on the variable "race."[198] Each column represents a different random draw from the data set. We can see that for seven of the draws, the variable "race" is not selected at all. In draw 4 the "race" variable receives a very small and negative weight, and in draw 5 and 9, "race" receives a larger negative weight of around $-0.0059$ and $-0.0082$, respectively. This means that in draw 5 and 9 (and to some extent draw 4), white borrowers are predicted to have a lower default risk.

[196] Pope & Sydnor's paper also presented results on how accuracy is largely maintained using this method. *See* Pope & Sydnor, *supra* note 185. For a discussion of how this analysis may not hold in the machine-learning context, *see* Kristen M. Altenburger & Daniel E. Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 JOURNAL OF INSTITUTIONAL AND THEORETICAL ECONOMICS (JITE) 98 (2019).

[197] Prince and Schwarcz discuss other challenges for this approach, which I do not discuss in detail. For example, they suggest that an insurer might be more inclined to discriminate if they learn from this method that a protected characteristic is predictive. They also suggest that because this approach requires the explicit consideration of protected characteristics, it might trigger discrimination law. It could also give rise to a constitutional challenge because it requires companies to consider protected characteristics. *See* Prince & Schwarcz, *supra* note 135, at 57.

[198] The "race" variable is a dummy variable. It is 1 when a borrower is white and 0 otherwise.

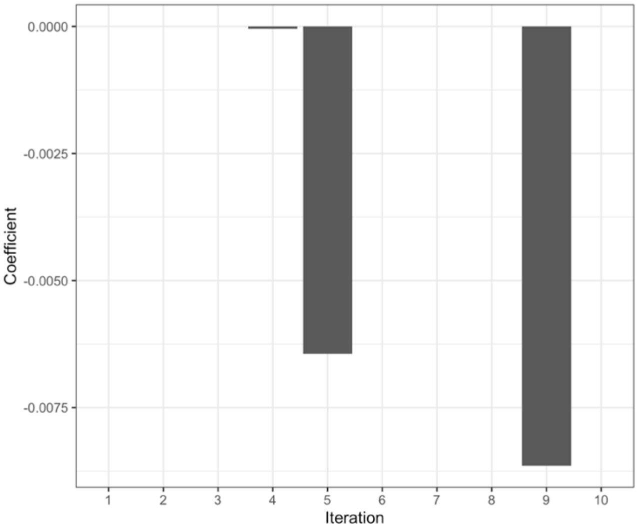


*Figure 10: Weight on "race" variable. Each of the 10 columns represents a different random draw of 2,000 observations from the full data, on which I fit a lasso regression. The columns plot the weight of the lasso regression on the race variable. There is a non-zero weight on the race variable for only 3 out 10.*

Importantly, despite the different weights put on “race” in draw 5 and 9, relative to the other draws, the overall predictions appear qualitatively similar. The top row of Figure 11 shows the distribution of default predictions by group for whites and non-whites for draws 4 and 5. The distribution for whites and non-whites is not identical across the two draws; however, they are qualitatively quite similar with respect to their pricing properties across groups. Therefore, although the prediction functions look very different, the underlying data, and the way in which they were constructed, and the resulting price distributions are all similar.

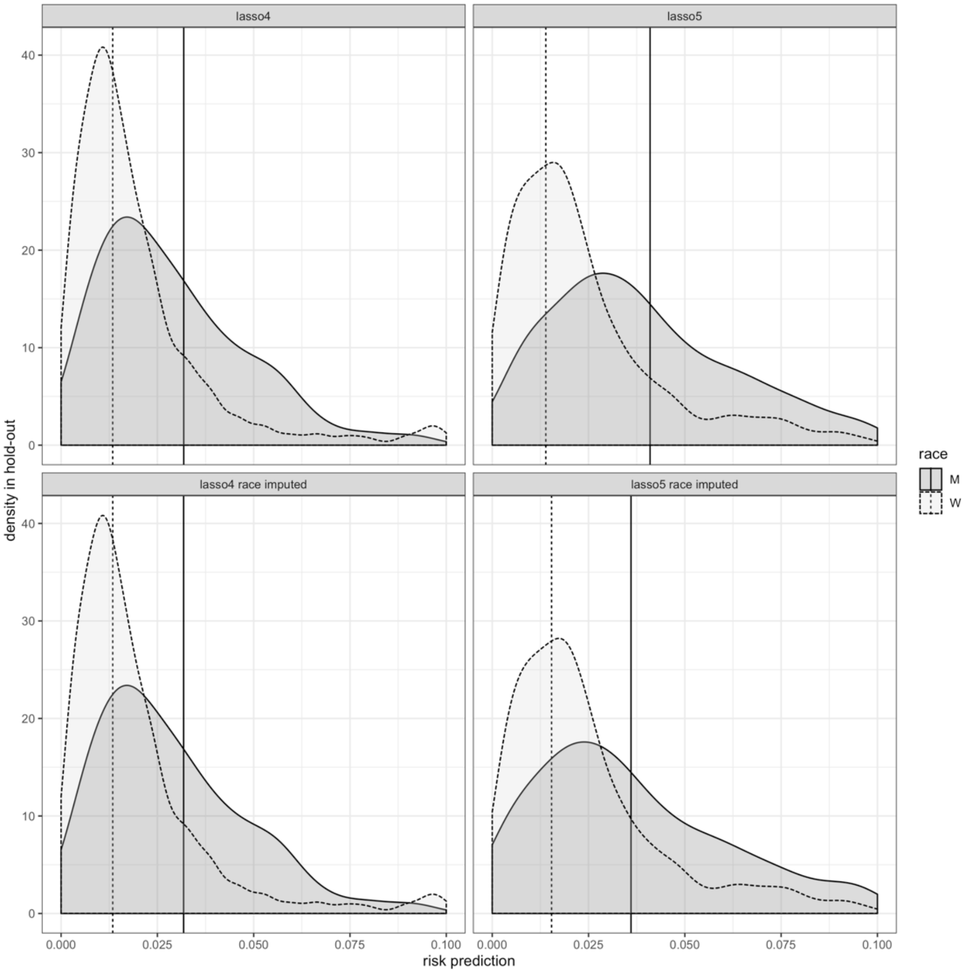


*Figure 11: The top row shows the distribution of the risk prediction for white (W) and non-white (M) borrowers, from a lasso regression using the all inputs. The graph on the top left corner is the prediction from random draw 4 from the dataset, and the graph on the top right corner is the prediction from random draw 5 from the dataset. The bottom row shows the prediction when "race" is substituted by "mean race" (0.8). The vertical lines represent the median for whites and non-whites.*

Applying the orthogonalization method would lead to different results depending on the draw. To see this, I applied the "mean race" to all borrowers[199] and plotted the default distribution in the bottom row of Figure 11. For draw 4, the left column of Figure 11, the top and bottom row are nearly identical, and the median for white and non-white borrowers does not change. This is because in draw 4, the variable "race" has a weight that is close to 0. For draw 5, on the right, the orthogonalization reduces the disparities between whites and non-whites. This can be seen by the fact that

[199] There are many more white borrowers in my dataset than non-white borrowers, so that $\bar{X}^{race} = 0.8$.

the vertical lines, representing the median for white and non-white borrowers, are closer to one another on the bottom graph relative to the top graph, for draw 5.

The conclusion of this exercise is that even though the training datasets of draws 4 and 5 are very similar, the lasso regression made different choices with respect to the weight on "race." The orthogonalization method, which uses the coefficient or weight on "race" for the screening stage, will therefore yield different results based on the random draw. Appendix C broadens this example by demonstrating how, more generally, there is instability with respect to the variables selected by the lasso regression.

#### 3.4.2 Why machine learning is different

The technical exercise in the previous section reveals a more substantive limitation in how to interpret machine learning algorithms. In a standard regression analysis, the coefficients represent some estimation of the impact of the independent variables on the predicted dependent variables. The fact that regression coefficients are often stable, even when slightly adding noise to the dataset, reflects the regression's estimation of an underlying model.

This is not the case with machine learning. Although the lasso regression output function looks similar to the output of an OLS regression, it should be interpreted differently. Machine learning is constructed to optimize the prediction accuracy. Therefore, the fact that even small amounts of noise in the data can change the variables that are selected by the algorithm in forming the prediction may not matter as long as the prediction accuracy is somewhat stable. 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.

The orthogonalization method goes wrong in the machine learning context because it essentially involves lying to the algorithm. The method asks the algorithm to optimize the prediction when it has access to race, only to restrict this access when applying the prediction function. This may not be a problem when the prediction was based on estimating the model, therefore isolating the effect of race on the prediction. However, when using a machine learning algorithm, the use of race is instrumental in optimizing the prediction accuracy and is not a substantive evaluation of its contribution to the prediction.

### *3.5 Required shift from causation to correlation*

Translating traditional discrimination law to the algorithmic context requires more than small tweaks suggested by approaches that continue to focus on the credit pricing inputs and their causal relationship to the differential treatment on protected groups. Instead we must recognize that in the machine learning context, we cannot identify causal relationships and cannot consider variables selected by an algorithm in forming a prediction as estimating a model.[200]

As discussed in Section 1.3, fair lending law has traditionally focused on causal questions. Disparate treatment is concerned with whether a protected characteristic affected a decision and therefore seeks the causal connection between a protected characteristic and a lending decision. Disparate impact is also concerned with causal connections. A disparate impact claim is initiated through establishing the causal link between a specific input or policy and a disparate outcome. Similarly, a defense to a disparate impact claim must rely on the causal connection between an input and a legitimate business goal.

In considering discrimination law in the algorithmic context, not only do many scholars continue to apply a causal framework,[201] scholars as well as

[200] *See* Martin J. Katz, *The Fundamental Incoherence of Title VII: Making Sense of Causation in Disparate Treatment Law*, 94 GEO. L .J. 489, 552 (2006) (discussing the causation requirement in anti-discrimination laws); Sheila R. Foster, *Causation in Antidiscrimination Law: Beyond Intent versus Impact*, 41 HOUS. L. REV. 1469, 1472 (2005) ("The prohibition against discrimination is a prohibition against making decisions or taking actions on account of, or because of, a status characteristic singled out for protection by our civil rights laws or constitutional traditions (which generally include race, gender, nationality, religion, disability, and age).")

[201] *See* MacCarthy, *supra* note 74, at 83 ("The law is generally clear that there must be a nexus or causal connection between some element of institutional practices and the disparate outcome for there to be a finding of illegal discrimination.") *Also see* Grimmelmann & Westreich, *supra* note 39, at 170 ("We believe that where a plaintiff has identified a disparate impact, the defendant's burden to show a business necessity requires it to show not just that its model's scores are not just *correlated* with job performance but *explain* it."). The extent to which this might be true under fair lending is unclear, given the strict criteria for a "business necessity" defense. *See* King & Mrkonich, *supra* note 39, at 571. Law's causal inquiry should be distinguished from a social science understanding of causality. Legal causal analysis does not rely on presenting rigorous empirical identification of causal relationships. Instead, claims of causality focused on intuitive understanding of how factors and inputs are related to the outcomes of policies or how they related to a legitimate business end and were justified accordingly. For example, see the NCLC's description of this intuitive type of argumentation ("There is an understandable connection between timely repayment of past obligations and the likelihood of timely repayment of future obligations, so a "demonstrable relationship" argument can be easily made"). Yu et al., *supra* note 62, at 29. In the context of employment discrimination, Kim argues that in traditional forms of testing for a job, employers determined which skills are related to job performance and then used tests for those attributes. This is consistent with an intuitive model of causality between inputs (attributes) and outputs (job performance). *See* Kim, *supra* note 7, at 874. Also see page 881 ("If, however, the variables are merely correlated and not

regulators focus on causal relationships. The recently proposed HUD disparate impact rule suggests that defendants can negate a claim of disparate impact if they "break down the model piece-by-piece and demonstrate how each factor considered could not be the cause of the disparate impact."[202] HUD's articulation of the defense relies on the ability to isolate inputs and separately evaluate their causal relationship to a disparate outcome.

These causal relationships break down in a machine learning world. The relationships that an algorithm uses to form a prediction reflect correlations in the data and not a causal connection to the outcome of interest. When an algorithm considers whether a borrower has an android phone to predict his or her creditworthiness, for example, it is not telling us about the causal relationship between phone type and default. A person who buys a new phone is unlikely to alter their actual risk of default. Rather, the basis for an algorithm to use a borrower's phone type could be its correlation with a variable that is causally related to default, such as income, or some other type of association.[203]

The traditional focus on causality is also problematic because in machine learning, the algorithm's selection of inputs is unstable, as demonstrated in the previous Section.[204] The differences between traditional regression analysis and machine learning are not only important for social scientists but also for legal analysis. In considering a claim of disparate impact, we want to know which inputs are driving disparities; however, with machine learning, we should not interpret the selection of variables as an indication of which input drives disparities.[205] If the inputs used to form a prediction are unstable, this will also make it difficult for the human eye to identify when "proxies" are being used for a protected characteristic.

causally related, there is no necessary connection between them, and the correlation may not hold in the future.")

[202] *See* HUD Proposed Rule 2019, at 20 (explanation of the proposed rule) . It is unclear what exactly HUD means by this defense, especially in light of the following sentence, according to which a plaintiff can undermine the defense if they show that "the defendant's analysis is somehow flawed, such as by showing that a factor used in the model is correlated with a protected class." *Id.*, at 20-21.

[203] Another example is that the time it takes a person to fill in out online application may be predictive of default risk. *See* Berg et al., *supra* note 102. We cannot know whether this is because it relates to a person's protected characteristic or not. All we know is that it is predictive.

[204] In most social sciences the focus is on parameter estimation, meaning to produce functions that produce meaning estimates of the way inputs (independent variables) relate to what is being predicted (dependent variable). For further discussion see: Sendhil Mullainathan & Jann Spiess, *Machine Learning: An Applied Econometric Approach*, 31 J. ECON. PERSP. 87 (2017).

[205] *See* Gillis & Spiess, *supra* note 25.

## IV. TOWARD A SOLUTION

Given the unsuitability of input-based approaches in the algorithmic setting, there is a need to rethink how to analyze discrimination in this new context. This is true for both disparate treatment and disparate impact. For disparate treatment, we have no reliable way to detect proxies for protected characteristics. For disparate impact, we need new tools to evaluate the effects of algorithmic pricing that are appropriate for machine learning, as restricting variables upstream can have a limited or surprising effect on the disparities downstream.

The shortcomings of current approaches mean that fair lending law must make the necessary, yet uncomfortable, shift to outcome-focused analysis. Discrimination law has always resisted focusing solely on the outcomes or effects of a policy as a way of identifying discrimination. However, when credibly scrutinizing inputs is not an option, downstream analysis provides important opportunities. [206]

[206] One possibility, not fully addressed in this paper, is that discrimination law altogether is no longer the appropriate legal framework to address concerns in the algorithmic context. Several academics and policy makers have argued that the unique challenges of the algorithmic fairness require an alternative framework to discrimination. Some have argued that the way we need to address algorithmic challenges is closer to the framework of affirmative action than discrimination. *See* Cynthia Dwork et al., *Fairness Through Awareness*, PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214 (ITCS '12, ACM New York, NY, USA 2012) (proposing "fair affirmative action.") Chander makes a similar argument with respect to affirmative action as the appropriate framework through which to deal with unfair outcomes as a result of biased inputs is through affirmative action. *See* Chander, *supra* note 38, at 1040. Huq's recent proposal to evaluate algorithmic criminal justice measured based on their effect on racial stratification is also an output-based framework because it looks to the benefits and costs of the criminal justice measures. *See* Huq, *supra* note 183, at 1128. Huq discusses the limitations of discrimination law in the context of the Constitutional Equal Protection, and therefore focuses on the disparate treatment doctrine. For broader discussions of the appropriateness of discrimination law, *see* Anna Lauren Hoffmann, *Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse*, 22 INFORMATION, COMMUNICATION & SOCIETY 900 (2019). Arguing that discrimination law is an insufficient framework to address the structural concerns that arise as a result of big data and algorithmic decision-making.) My focus here is on how to identify credit pricing that could raise red flags in light of the concerns highlighted in Section 2.3. Whether the correct way to address these concerns is through discrimination law or another legal framework is not an issue wherein I argue for any particular position. However, it is important to keep in mind that the basic challenges of "biased world" and "biased measurement" have always been a challenge for discrimination law, even prior to the algorithmic context.

Others have suggested that the concerns of algorithmic fairness be addressed not through mechanisms that directly regulate conduct but through creating appropriate frameworks that allow further private or public scrutiny. For example, one major approach discussed in the literature is transparency. See for example FRANK PASQUALE, THE BLACK BOX SOCIETY:

One challenge in developing an outcomes-based test is that there are currently widespread and far-reaching disagreements over the theoretical foundations and boundaries of the discrimination doctrine. The exact details about the implementation of the test rely on a clear definition of what discrimination law aims to achieve. I do not adopt a particular position on these disagreements. Instead I highlight how outcome analysis can be used to answer important questions that often are of interest to discrimination law. The framework I lay out is flexible enough to accommodate varying views of discrimination law.

This section begins with a discussion of how to create an outcome-based framework for discrimination analysis. In this framework, a regulator applies the pricing rule to a designated dataset to analyze the properties of the pricing rule. I highlight two particularly meaningful questions a regulator can address using this framework. First, a regulator can ask whether borrowers who are "similarly situated" are treated the same. Second, a regulator can analyze whether the pricing rule increases or decreases disparities relative to some baseline, such as the pricing rule used prior to the utilization of an algorithm. I end the section by discussing how this framework takes advantage of technological change to enhance its toolkit. This type of outcome-focused testing brings to the forefront the demonstration of disparities, which is formally part of first stage of a disparate impact complaint in traditional fair lending law. My proposed testing framework develops this type of analysis and adapts it to the machine learning context.

A full discussion of the various ways this test can be structured and implemented is beyond the scope of this paper. Future work will consider the various challenges in designing and implementing the test along with the incentives the test could create for lenders. Importantly, it will demonstrate how the test can be adjusted based on different normative theories of discrimination.

### *4.1 Testing outcomes in three steps*

The outcome testing I propose requires regulators to apply a lender's pricing rule to a dataset of hypothetical borrowers and then examine the properties of the outcome. The framework can therefore be split into three stages. At the first stage, the lender determines what inputs and which algorithm to use to predict default and price accordingly.[207] At the second

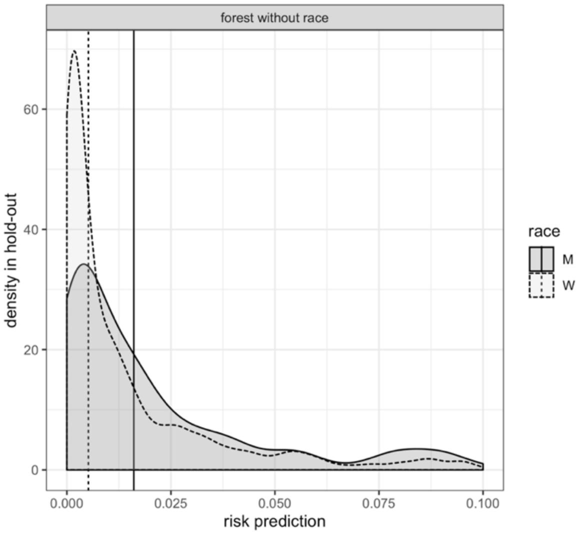
THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015). For skepticism over whether transparency or privacy can address fairness concerns, see Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy, and It's Not Fair*, STAN. L. REV. ONLINE 35 (2013–2014).

[207] An alternative analysis that the regulator could conduct would be to compare the binary decision of lenders of whether to extend or deny a loan application. In fact, HMDA

stage, the regulator then takes that prediction or pricing rule and applies it to a dataset of people to see the distribution of prices the rule produces. Finally, the regulator evaluates this outcome to determine whether the disparities created by the pricing rule amount to discriminatory conduct. I use the example of race as a protected characteristic, but the analysis is generalizable to other protected characteristics.

The first stage of the test is the pricing rule developed by the lender, which does not directly involve the regulator. What is unique about the machine learning context is that a pricing rule exists even before specific borrowers receive loans. In traditional credit pricing, little is known before actual prices were given to real borrowers. In the algorithmic context, because the process is fully automated, regulators can analyze prices in an *ex ante* manner, before the algorithm is applied to price credit.

In the second stage, the pricing rule is applied to a dataset, containing real or hypothetical people and their characteristics. Elsewhere I have argued that it is difficult to analyze a prediction function in the abstract.[208] Rather, the prediction function should be applied to a group of borrowers in order to examine its properties. For data scientists, this is typically the holdout set, meaning a subset of the data on which the algorithm is not trained but is instead used to assess the accuracy of the prediction. A regulator could be strategic in selecting which population to apply a pricing rule to, by not sharing the dataset with the lender in advance.[209]



is primarily focused on understanding whether this lender decision varies by race.

[208] *See* Gillis & Spiess, *supra* note 25, at 473.

[209] Future work will discuss the various factors a regulator should consider in selecting the population used to analyze a lender's pricing rule.

*Figure 12: Distribution of risk predictions. This graph separately plots the distribution of risk predictions for non-Hispanic white (W) and non-white (M) borrowers.*

Figure 12 plots an example of what the regulator's initial analysis would look like. In this example, a regulator takes a lender's pricing rule and applies it to a dataset held by the regulator. One way to think of the dataset used by the regulator is that it represents a group of hypothetical borrowers for which we want to learn the price this group would be charged for a loan. Figure 12 then plots the distribution of prices separately for white and non-white borrowers.[210]

The raw disparities are rarely of interest in and of themselves, so that in the third stage of the test, the regulator needs to determine whether disparities created by a pricing rule amount to discrimination.[211] For the reasons discussed in the previous Section, the criteria used to determine whether pricing disparities amount to discrimination needs to be formulated without reference to the inputs used. The exact criteria to be used in outcome analysis cannot be defined without a clear definition of what discrimination law, and disparate impact in particular, are meant to achieve.

[210] The credit price is not the only outcome metric that is of interest to a regulator. The regulator could use a similar method to analyze a lender's binary decision of whether to extend a loan. Using the lender's algorithmic rule in determining whether to reject a loan, the regulator could apply this rule to its hypothetical dataset of lenders. The focus could then be on analyzing disparities with respect to error rates. Much of the algorithmic fairness literature has focused on fairness definitions that are types of "classification parity" meaning they consider whether a measure of classification error is equal across groups. *See* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV:1808.00023 [CS] 5 (2018). Corbett-Davies and Goel define this category as any definition that can be calculated from a confusion matrix, which tabulates the joint distributions of a certain decision and outcomes by group. See: Richard Berk et al., *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOCIOLOGICAL METHODS & RESEARCH 0049124118782533, 3 (2018). Two of these measures, the "true positive rate" (TPR) and "false positive rate" (FPR), discussed in Appendix B, to provide a way to measure the prediction accuracy. An ancillary literature has focused on the documenting how the various classification errors can often not simultaneously be satisfied. See discussion is Jon Kleinberg et al., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV:1609.05807 [CS, STAT] (2016); Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017); Berk et al., *supra*.

Some recent legal literature has also focused on these types of outcomes. *See* Hellman, *supra* note 136; MacCarthy, *supra* note 74, at 88.

[211] Outcome analysis has always been a part of a disparate impact claim for the purposes of the *prima facie* case, but was rarely the determining factor. A typical disparate impact claim begins with a demonstration of outcome disparities. This showing of disparities is rarely sufficient in of itself even for the first stage of a case, since a plaintiff is also required to isolate the particular policy or input that lead to the disparity. Despite the role outcome analysis plays in a disparate impact case, there is little guidance on how exactly to conduct this analysis.

A full discussion of the different theories of discrimination and how to develop the closest equivalent outcome-based tests to those theories, is beyond the scope of this paper. Instead, the focus of this Part is on demonstrating how outcome analysis can answer meaningful questions related to discrimination. I focus on two questions that can be analyzed using outcome-based analysis. The first question is whether borrowers who are "similarly situated" are treated the same. The second question is whether the pricing rule increases or decreases disparities relative to some baseline.

#### 4.1.1 Comparing borrowers who are similarly situated

An important question for discrimination law is whether borrowers who are similarly situated are treated the same.[212] In traditional disparate impact case, this is required as part of the *prima facie* case. Discrimination law has long recognized that there are differences that are a legitimate bases over which to distinguish between borrowers.[213] Despite the significance of the definition of who is "similarly situated" under traditional fair lending, there is little guidance on this question.[214]

[212] This requirement originates in the seminal Title VII case, McDonnell Douglas Corp. v. Green, 411 U.S. 792 (1973)). Some courts were willing to extended the "McDonell Douglass standard" to the credit context. *See* Robert G. Schwemm, *Introduction to Mortgage Lending Discrimination Law A Fair Lending Symposium: Litigating a Mortgage Lending Case*, J. MARSHALL L. REV. 317, 329 (1994–1995)*,* (summarizing fair lending cases and the requirement that the plaintiff had to establish that "the defendant approved loans for white applicants with qualifications similar to the plaintiff's"). *See also* Simms v. First Gibraltar Bank, 83 F.3d 1546 , 1558 (5th Cir.1996). *See* Judge Posner in Latimore v. Citibank, 151 F.3d 712, 713 (7th Cir. 1998), for a more skeptical view of the application of the "similarly situated" requirement to the credit context. In general, the notion of "similarly situated" has been somewhat controversial over the years, including in the context of employment discrimination. For further discussion, see Suzanne B. Goldberg, *Discrimination by Comparison*, YALE L. J. 728 (2010–2011); Ernest F. III Lidge, *The Courts' Misuse of the Similarly Situated Concept in Employment Discrimination Law*, MO. L. REV. 831 (2002).

[213] According to the Supreme Court in *Inclusive Communities*, even the prima facie case of the plaintiff cannot rely only a showing of disparities, See Texas Dep't of Hous. and Cmty. Affairs v. Inclusive Cmty. Project, Inc., 135 S. Ct. 2507 (U.S. 2015). ("In a similar vein, 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.")

[214] Despite the role outcome analysis plays in a disparate impact case, particularly in the prima facie case of a claimant or plaintiff, there is little guidance on how exactly to conduct this analysis. See discussion in Giovanna Shay, *Similarly Situated*, GEO. MASON L. REV. 581, 583 (2010–2011) ("Although the phrase 'similarly situated' is a familiar component of equal protection case law, it has not received much scholarly attention. Constitutional law scholars have focused more on other aspects of the doctrine.") For an example of what outcome analysis might look like in the, see the expert opinion discussed in Ayres et al., *supra* note 79, at 238. The effect of race was considered by using a regression with various controls although the paper does not directly discuss which controls are appropriate to

A new interpretation to this old question can be used as an outcome-based test in the algorithmic setting. In world in which there is no credible way to determine at the outset whether a protected characteristic is being used to price, the closest alternative would be to ask- are the prices different for protect groups, controlling for the legitimate grounds for differentiation? In a sense this is reverse engineering the basic classification question of whether borrowers are distinguished based on the protected characteristic. Only the unexplained component of price disparity would then be the basis of discrimination and not the raw disparities alone.

In the algorithmic context, we can consider a set of characteristics which determines who is similarly situated. Any differences that are explained by this set of characteristics are not deemed to be impermissible discrimination.[215] This set can intuitively be understood as adding control variables into a regression in that they explain differences between people.[216] The size and scope of the similarly situated set are likely to have a significant effect on whether there is a finding of impermissible disparity. As this set expands, more of the raw differences are accounted for by the preexisting differences of protected groups.

It is important to note that who is similarly situated is essentially a normative question and not an empirical one, as it reflects who we believe should be treated similarly.[217] The difference between the empirical question of who is the same versus who should be treated the same becomes particularly apparent when considering that fair lending law prohibits discrimination based on protected characteristics, even if they are directly

include.

There is some ambiguity over whether the requirement to demonstrate that a member of a protected group was treated differently to someone “similarly situated” is part of the first or third stage of the burden-shifting framework. *See* Goldberg, *supra* note 209, at 746–47 (discussing this ambiguity in demonstrating "comparable"); *See also* Schwemm, *supra* note 209, at 328–29. Furthermore it is unclear whether the requirement is part of a disparate treatment case as well as a disparate impact case.

[215] There are some similarities between my framework the framework proposed by Dwork et al., *supra* note 203. Their approach is based on a similarity metric between individuals who are treated fairly if the classifier ensures similar outcomes for similar individuals.

[216] This is similar to the analysis discussed in Ayres et al., *supra* note 79. The expert report discussed in that paper presented different linear regression models, which included different variables as controls to consider whether there was still a significant coefficient on “race” after adding the controls.

[217] Note that the similarly situated set is separate from the set of characteristics that is *predictive* of the outcome. If all the characteristics that are predictive of an outcome were included in the similarly situated set, then by definition, the algorithmic credit pricing does not create impermissible disparity. Adopting such a definition of the similarly situated test puts us back into the world in which once the protected characteristic is excluded, discrimination law is no longer relevant, a position discussed in detail in Part II.

related to default. As discussed above, age and marital status may change a borrower's default risk, yet these characteristics cannot be used to distinguish between people.

Testing for disparities among the "similarly situated" may seem as a return to input-based approaches, as it relies on the selection of the legitimate bases for differentiation. If the test requires selecting normatively relevant criteria for distinction then it may seem similar to restricting an algorithm to pre-approved inputs. However, this test differs from restricting an algorithm's input to the characteristics in the similarly situated set in several ways. Restricting an algorithm's inputs to the similarly situated set would definitely be sufficient for this test, but it is not necessary to do so. This is because there may be many inputs that increase prediction accuracy while not creating significant disparities. This is especially important in the case of characteristics that would help increase access to credit for protected groups but are unlikely be included in the similarly situated set, such as timely rental payments. Moreover, a regulator may set the tolerance level such that some disparity is permissible, when using inputs beyond the "similarly situated" set.

In general, creating a test that relies on similarly situated characteristics makes the tradeoff between accuracy and other policy goals explicit, rather than the opaqueness of restricting to inputs that intuitively seem relevant to default. It also means that this set can be adjusted and tested rather than the inability to learn and adapt that results from input restriction. Nonetheless, the disadvantage of this approach is its reliance on a normatively determined set, which may be problematic particularly if the set includes characteristics that may themselves be the source of disadvantage, such as credit scores.

### 4.1.2 Considering incremental change

Another meaningful way to consider the disparities created by algorithmic pricing is to do so relative to some baseline, such as traditional credit pricing. Rather than consider the absolute levels of disparities created by a pricing rule, like in figure 12, the focus is on how these disparities compare to traditional credit pricing rules. Similarly, a regulator could compare the prices produced under the use of traditional lending variables versus a new data available to a lender, such as consumer and payment behavior.

An incremental approach to disparities recognizes that credit is priced in a "biased world" but also seeks to prevent algorithmic pricing from exacerbating preexisting disadvantage.[218] When personalized pricing relies

[218] *See* Berk et al., *supra* note 207, at 29. ("At the same time, the benchmark is current practice. By that standard, even small steps, imperfect as they may be, can in principle lead

on biased inputs, it is unlikely to ever produce pricing that is not disparate for protected groups. This type of test is therefore more appropriate for the effect-based interpretation of disparate impact,[219] as it seeks to balance both the concern for further entrenching disadvantage and the interests of lenders and importance of functioning credit markets.

Furthermore, as discussed in Section 2.3, the use of nontraditional datasets could in fact mitigate the harms of biased measurement, which would reduce disparities among groups. Accurate pricing could also expand access to credit, which could also benefit vulnerable groups. The conclusion is that there is a need for an empirical test for determining whether there is harm to protected groups stemming from changes in credit pricing rather than from the general use of biased inputs in credit decisions. This approach therefore avoids holding algorithms to a standard that is far harsher than current standards of fair lending are, which may end up overlooking the potential of algorithmic pricing to help consumers.

This type of incremental analysis is suggested by a recent update published by the Consumer Financial Protection Bureau (CFPB). The background for this update is a No-Action Letter that the CPFB provided an algorithmic lender, Upstart, in 2017.[220] In its update on the No-Action Letter, from August 6, 2019, the CFPB reported results from Upstart's analysis "comparing outcomes from its underwriting and pricing model (tested model) against outcomes from a hypothetical model that uses traditional application and credit file variables and does not employ machine learning (traditional model)."[221] The focus of Upstart's analysis was therefore the incremental change in moving from traditional credit pricing to algorithmic credit pricing.

to meaningful improvements in criminal justice decisions. They just need to be accurately characterized.")

[219] See Section 1.3.

[220] Consumer Financial Protection Bureau, *No-Action Letter* (Sep. 14, 2017), https://www.consumerfinance.gov/documents/5462/201709_cfpb_upstart-no-action-letter.pdf. This was the first and only No-Action letter that the CFPB has provided. For the general policy, see Bureau of Consumer Financial Protection Policy on No-Action Letters, 81 FR 8686 (Feb. 22, 2016). On August 6, 2019, the CFPB provided an update on the No-Action letter, discussing how Upstart had expanded access to credit. *See* Patrice Ficklin and Paul Watkins, Consumer Financial Protection Bureau, An Update on Credit Access and the Bureau's First No-Action Letter (Aug. 6, 2019). In its update, the CFPB suggests that Upstart conducted this type of incremental analysis ("Pursuant to the No-Action Letter, [Upstart] provides the Bureau with information comparing outcomes from its underwriting and pricing model (tested model) against outcomes from a hypothetical model that uses traditional application and credit file variables and does not employ machine learning (traditional model).")

[221] Patrice Ficklin and Paul Watkins, Consumer Financial Protection Bureau, An Update on Credit Access and the Bureau's First No-Action Letter (Aug. 6, 2019), https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/.

It is this type of analysis that should form the core of fair lending analysis.

HUD's recent Proposed Rule suggests a similar approach to determining a lender's defense in a disparate impact case. According to the proposal, when relying on an algorithm, a lender can show that "use of the model is standard in the industry."[222] The Proposed Rule therefore recognizes that one of the ways to establish that an algorithm is not discriminatory is by reference to some baseline, in this case the "industry standard," rather than some absolute level of disparity.

In summary, outcome-based testing could provide important information about two questions that are meaningful to discrimination analysis, namely whether similarly situated borrowers are treated the same and whether a change in pricing increases or decreases disparity. Many questions regarding the implementation of the test will be discussed in future work. These issues include how to deal with a case in which a new pricing rule decreases disparities for one protected class (e.g., African Americans) but increases disparities for another (e.g., women) or how regulators can balance the need to clear rules that allow for certainty with the flexibility of the incremental change approach.

### *4.2 Regtech response to Fintech*

Regulators need to develop tools that will allow them to respond effectively to changes in the credit pricing world. Credit pricing is becoming more complex, both with respect to the decision inputs and how those inputs are used to produce predictions and pricing rules. This environment is becoming increasingly difficult to oversee, as regulators need to supervise an evolving technological environment. Past regulatory focus on analyzing inputs was to a large extent feasible because of the limited complexity of credit pricing decisions.

The move to a more technologically complex environment can create important opportunities for regulators. This is underappreciated by many scholars who focus solely on the challenges for regulators in gaining competency in new domains. However, machine learning pricing brings new types of transparency that can also create new regulatory tools. In the case of credit pricing, the greatest change is that much is known about credit pricing even before a pricing rule is applied to new borrowers. In the traditional credit pricing context, regulators respond to materialized prices, meaning actually prices charged to actual borrowers.

In the machine learning pricing context regulators can analyze pricing rules before they are applied to real borrowers, creating the potential for *ex ante* testing. As argued throughout this Article, the effects of changes in credit

[222] HUD Proposed Rule 2019, at 42859.

markets on disparities between groups is unclear and cannot be adequately studied from a theoretical perspective. This means that only a testing method can provide information on the actual effects of pricing rules.

This approach also provides more certainty to lenders. Lenders that wish to depart from traditional credit pricing currently face a very uncertain regulatory landscape. Although much of the writing about discrimination and artificial intelligence focuses on preventing intentional discrimination, the reality for many lenders is that they are unsure of how to comply with discrimination law. A testing approach might be especially valuable for lenders that wish to embrace a new technology or use a novel dataset but are concerned with the legal uncertainty that often accompanies technological change. The outcomes-based testing approach provides a method to resolve some of that uncertainty, by analyzing the theoretical effects of a pricing rule.

## CONCLUSION

Risk-based pricing is about differentiating borrowers. Big data and machine learning enhance the ability to differentiate, increasing the tension with fair lending law that limits differentiation of borrowers on protected grounds. Traditional fair lending law sought to constrain pricing practices by scrutinizing inputs. This approach was developed in a world in which pricing relied on few inputs, depended on human expertise and used loan officers to set the final terms of credit contracts. Modern underwriting is increasingly relying on nontraditional inputs and advanced prediction technologies, challenging existing discrimination doctrine.

Legislators and regulators face a difficult puzzle in crafting regulation that retains the benefits of algorithmic credit pricing while limiting its potential to hurt protected groups. In May 2019, the House Financial Services Committee established a task force on financial technology to "examine the current legal framework for fintech, how fintech is used in lending, and how consumers engage with fintech," along with a second task force on artificial intelligence.[223] The CFPB, in its July 2019 fair lending report, highlighted the Bureau's interest in "ways that alternative data and modeling may expand access to credit" while also seeking to understand the risks of these models.[224] Finally, the CFPB announcement from August 6, 2019, endorsed the view that big data and machine learning lenders could comply with fair lending if they demonstrate that their lending practices do not increase disparities.[225]

[223] *See* U.S. House Committee on Financial Services, Press Release (May 9, 2019), https://financialservices.house.gov/news/documentsingle.aspx?DocumentID=403739.

[224] Fair Lending Report of the Bureau of Consumer Financial Protection, June 2019, 84 Fed. Reg. 32420, 32423 (July 8, 2019).

[225] Patrice Ficklin and Paul Watkins, Consumer Financial Protection Bureau, An Update

Fair lending is likely to be a central battleground on which the boundaries of algorithmic fairness and discrimination will be fought.

My aim in this Article has been to show that currently-favored proposals to resolve the tension between old law and new realities are not a promising one. Current approaches are inadequate because they continue to scrutinize decision inputs, similar to traditional fair lending, when this strategy is no longer feasible or effective in the algorithmic context. This input-scrutiny perspective is adopted both by opponents of a broad disparate impact standard, such as the Trump administration's HUD, and proponents of a broad standard. Algorithmic decision-making, however, requires a fundamental shift away from analysis that seeks to reveal causal connections between inputs and outcomes.

I propose that fair lending shift its gaze downstream to the outputs of an algorithm. Regulators should develop tests for considering when the outcomes an algorithm creates are impermissible, based on regulatory policy goals. Regulators can begin by asking meaningful questions that can be answered by examining algorithmic outcomes, such as whether similarly situated borrowers are treated differently or whether the move from traditional pricing to algorithmic pricing has increased disparities. This type of test is particularly important when it is not possible to determine at the outset whether a change in prediction technology or input variables will increase or decrease disparities. An empirically driven and experimental approach means not only that regulators keep up with the fintech industry but also that they embrace technological advancement to improve regulation.

The conclusions go beyond credit pricing. They are important for other domains in which scholars and lawmakers are struggling to apply discrimination law to the algorithmic setting, such as criminal justice and employment. It is time for discrimination law to fully recognize the challenges related to algorithmic decision-making while embracing its opportunities.

on Credit Access and the Bureau's First No-Action Letter (Aug. 6, 2019), https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter/.

# APPENDICES

## APPENDIX A: SIMULATION DATA[1]

As discussed in the main Article, I demonstrate my main points in a stylized simulation exercise that is calibrated to real data. Specifically, I consider a lender who prices mortgages based on an algorithmic prediction of their default risk, in order to consider the implications of using biased inputs in the algorithmic setting and to evaluate leading approach to the application of discrimination law to algorithmic decision-making. I also use the simulation exercise and to present my regulatory framework.

The simulation demonstration is based on real mortgage application data from the Boston HMDA data set. From this dataset a simulation model relates applicant and mortgage characteristics to the probability of default. Because mortgage defaults are not observed in this dataset, but are an essential aspect of the simulation demonstration, the default probabilities from loan approvals can be imputed and calibrated to overall default rates. As an important restriction of the analysis, I cannot make any statements about actual defaults in this data but rather demonstrate methodological points under this hypothesized model of default.

In general, HMDA requires mortgage lenders to disclose loan-level information on mortgage applications and whether they were granted or denied. A modified version of HMDA data is publicly available and includes basic data on the loan and the applicant, including demographic information such as race. I specifically use the Boston Fed HMDA dataset,[2] which is based on a follow-up survey conducted by the Boston Fed to supplement the data in HMDA on loans made in 1990 with additional information on financial, employment, and property characteristics.[3]

Despite the dataset's being nearly 30 years old, it is a uniquely rich dataset and therefore useful to consider a lender using machine learning predictions to set loan prices. The dataset contains information on the finances of the borrower, such as total debt-to-income ratio, the applicant's credit and borrowing history, whether the applicant is self-employed, and whether the borrower was denied private mortgage insurance. The dataset also contains information on the loan, such as whether the property is a multi-family home, whether the loan has a fixed interest rate, and the term of the loan. The most significant advantage of using HMDA data is that they contain

[1] (Adapted from "Big Data and Discrimination")

[2] For a description of how the Boston Fed created this unique dataset and a discussion of their findings, see: Munnell et al., *supra* note 54.

[3] HMDA does not contain information about credit histories, debt burdens, loan-to-value ratios among other factors. See: *Id.* at 25. Other important details about the Boston Fed data. They also used census data on neighborhood characteristics.

demographic characteristics, such as borrower race, gender, age, and marital status along with various neighborhood characteristics.[4] The lender can be considered a "big data" lender because this type of lender uses many variables (around 40) relative to the number of observations (around 3,000). Unfortunately, due to data limitations, this lender does not include many of the types of the nontraditional discussed in Section II.2.1.1; however, the types of variables are slightly broader than what is typically used by mortgage originator in setting the "par-rate" in traditional lending.[5]

The Boston Fed HMDA only contains information available at the time of the loan application and therefore does not contain information on the performance of the loan, such as whether a borrower defaulted on the loan. Based on the HMDA data alone, one could not run a default prediction exercise because the training data need to contain labels, meaning the outcome that the machine learning algorithm is trained to predict. To overcome this difficulty for the purposes of this exercise, I construct a model based on the dataset that links rejection approval rates to loan default.[6]

From this dataset a simulation model relates applicant and mortgage characteristics to the probability of default. Because mortgage defaults are not observed in this dataset, but are an essential aspect of the simulation demonstration, the default probabilities from loan approvals can be imputed and calibrated to overall default rates. As an important restriction of the analysis, I cannot make any statements about actual defaults in this data but rather demonstrate methodological points under this hypothesized model of default.

Specifically, a ridge-penalized logistic regression model of loan approval is fitted on approximately fifty characteristics of the loan and the borrower (including demographics, geographic information, and credit history), excluding race and ethnicity, which is then recalibrated such that the default rate among those approved for the loan matches the rate reported in a recent paper that uses the matched HMDA-McDash dataset.[7] As a result, for every individual in the Boston HMDA dataset, a probability of default is obtained.

The samples are drawn from the simulation population as follows. First, a bootstrap sample is drawn, without replacement, from the full Boston HMDA dataset. Second, for every individual in the bootstrap sample, whether that individual defaults is simulated based on the default probability

[4] Such as the appreciation of housing properties in the neighborhood.

[5] For a full description of the variables in the Boston Fed HMDA dataset see Munnell et al., *supra* note 54.

[6] The methodology is similar to the that discussed in the Online Appendix in Gillis & Spiess, *supra* note 25. Further details are provided in Appendix A.

[7] Fuster et al., *supra* note 20.

implied by the calibrated simulation model. As a result, default indicators along with individual characteristics for each individual in the sample are obtained.

In the simulation demonstration, the firm constructs a prediction of default based on a training sample of two thousand consumers drawn randomly. The firm utilizes a machine-learning algorithm that uses this data to produce a prediction function that relates available consumer characteristics (potentially including race) to the predicted probability of default. The properties of a given prediction rule on a new sample of two thousand consumers is then assessed.

As an example of an algorithm that produces such a prediction rule, the firm could run a simple logistic regression in their training sample that produces a prediction function of the form:

$$\begin{aligned}&\text{predicted probability of default}\\&\qquad= \text{logistic}(\alpha + \beta_1 \text{characteristic}_1\\&\qquad+ \beta_2 \text{characteristic}_2 + ...)\end{aligned}$$

where the characteristics could be the applicant's income or credit score. While the machine-learning algorithms considered in this Article also produce functions that relate characteristics to the probability of default, they typically take more complex forms that allow, among other things, for interactions between two or more characteristics to affect the predicted probability and are thus better suited to represent richer, possibly nonlinear relationships between characteristics and default. Some of these algorithms build on top of another simple prediction function, namely a decision (or regression) tree. The decision tree decides at every node, based on the value of one of the characteristics, whether to go left or right (for example, if income is below some threshold, go left, otherwise right), before arriving at a terminal node that returns a prediction of the probability of default of all individuals with the relevant characteristics. An example of a decision tree is given in Figure 1. Using this decision tree, the firm would predict that an individual who obtained mortgage insurance (top level, go left) but has a debt-to-income ratio of above 75 percent will have an 80 percent probability of default.

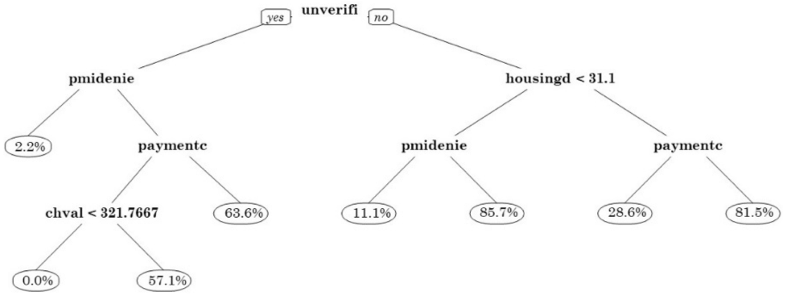


*Figure 13: A decision tree that predicts default probability on simulated data*

In order to analyze default predictions by group—for which the primary focus is on racial/ethnic groups in the simulation demonstration— I consider their distribution in the new ("holdout") sample of two thousand consumers drawn from the population. For most of the paper I use a rule obtain from a random forest machine-learning algorithm, which is a collection of many decision trees that are averaged.

## APPENDIX B: UNDERSTANDING ROC CURVES

The Receiver Operating Characteristic (ROC) curve is a way of capturing prediction accuracy, by focusing on the binary classification of borrowers. The algorithm used in the Article produces the default risk for each borrower. The predicted default risk can then be used by the lender to determine whether they believe a borrower is likely to default or not. For example, a lender can determine a cutoff of 30% default risk, so that all borrowers with a risk above 30% are deemed "defaulters" and all those below are "non-defaulters".

This cutoff will naturally produce some errors. There will a group of borrowers who were classified as "defaulters" but end up repaying the loan and not defaulting (type I error).[8] Conversely, there will be a group of borrowers that were classified as "non-defaulters" that end up defaulting (type II error). There is a tradeoff between the size of each of these error groups and minimizing the size of one group will increase the size of the other group. For example, raising the cutoff to 60% will decrease the type I error and increase the type II error. The more accurate a prediction, the smaller the tradeoff between these two types errors.

The ROC curve captures the intuition that a more accurate prediction requires less of a tradeoff between different types of errors. On the one hand it considers the True Positive Rate (TPR) which is the number of people who

[8] In reality this is often not observed. This is because the outcome of a loan is only known if an applicant actually receives a loan. I therefore treat these examples as the error rates that are observed in the holdout set.

were correctly classified as "positive" relative to the total number of people classified as "positive":

$$TPR = \frac{True\ Positive}{All\ Positive}$$

In our case a "positive" event is when a borrower defaults on the loan, so that the "true positive" is all the borrowers that the algorithm predicted would default on their loan and indeed they did. On the other hand, the ROC curve, considers the False Positive Rate (FPR) which is the number of people falsely classified as "positive" relative to the total number of people classified as "positive":

$$FPR = \frac{False\ Positive}{All\ Positive}$$

The ROC curve plots the TPR for every level of FPR. It therefore can be considered as a measure of the accuracy of the prediction. The closer the curve is to the top left corner, the more accurate the prediction. When the curve lies on the diagonal 45º line, it means that the prediction contains no information beyond random assignment.

Figure 2 shows the ROC curve for the risk prediction function that was produced using all variables other than race. The ROC curve is plotted separately for white and non-white borrowers. The Figure shows that the prediction for white borrowers is more accurate for nearly every classification cutoff.

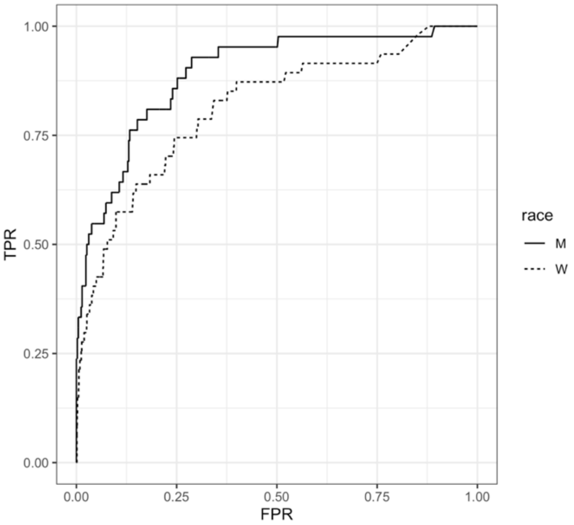


*Figure 14: ROC curve for risk prediction using all inputs (other than race), plotted separately for whites (W) and non-whites (M) borrowers.*

One common metric used to measure the prediction accuracy is the Area Under Curve (AUC). The AUC is a number from 0.5 (perfectly random prediction) to 1 (perfectly predictive). When comparing two prediction functions, the AUC is an useful metric to describe overall relative accuracy.

Another example is Figure 15, which plots the ROC curve for the prediction of "marital status" from other HMDA dataset variables. Figure 15 shows that a borrower's marital status can be predicted fairly accurately using the other HMDA variables.

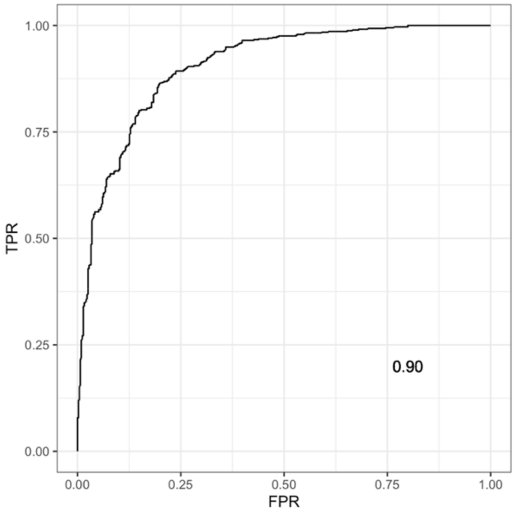


*Figure 15: ROC curve for prediction of "marital status." The "marital status" variable is a dummy variable equal to 1 if the applicant is married and 0 if the applicant is unmarried or separated in the Boston Fed HMDA dataset. The ROC curve plots the true-positive-rate and false-negative-rate for different cut off rules. The number in the lower right corner is the Area Under Curve (AUC).*

## APPENDIX C: INSTABILITY OF SELECTED VARIABLE

In this Appendix, I demonstrate how the variables selected by a machine-learning algorithm can be unstable. This instability has several implications. First, it means that we should be wary in putting weight on the fact that a variable was selected, since this selection could be fairly random. It also means we should not put too much weight on the fact that a variable was not selected by the algorithm. Ultimately, this analysis shows that we are unable to easily isolate the effect a variable has on the predicted outcome. I demonstrate this through an example similar to the one discussed in "Big Data and Discrimination".[9]

In a standard regression analysis, the coefficients obtained represent some

[9] *See* Gillis & Spiess, *supra* note 25. See Mullainathan & Spiess, *supra* note 201, for further discussion of the difference between estimation and prediction.

estimation of the impact of the independent variables on the predicted dependent variables. Although one must exercise caution in interpreting these coefficients as bearing a causal relationship to the dependent variable, they at the very least represent the weight they play in the prediction of the dependent variable. In most cases, adding some noise to the dataset used for the regression analysis should not significantly change the estimates. In other words, the estimates obtained are pretty stable and allow for some estimation of the underlying model of how the independent variables relate to the dependent variable.

This is not the case with machine-learning. With machine-learning even slight differences in the training set, or small amounts of noise in the data can vastly change the variables that are selected by the algorithm in forming the prediction. To create two comparable datasets with slightly different noise I randomly draw 2,000 observations from the full dataset ("training dataset 1") and then again randomly draw 2,000 observations from the full dataset ("training dataset 2"). Because these two datasets are randomly drawn from the same full dataset, they should roughly be the same, although they are unlikely to identical.

I then fit a logistic lasso regression to each of the two training datasets. The algorithm selects which of the many characteristics to include in the model. The advantage of using a logistic lasso regression is that its output looks quite similar to a regression output, in that it produces a function with the variables used to form a prediction and the weights of each variable.

Both training datasets originate from the same population and therefore we may expect that both training datasets 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.

Despite being drawn from the same population, it is not the case that the random sampling leads to identical algorithmic decisions. The specific representation of the prediction functions and which variables are used in the final decision rule vary considerably between training dataset 1 and training dataset 2. A graphic representation of this instability can be found in Figure 16. This Figure records which characteristics were included in the logistic lasso regressions we ran on the two draws. 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 the two draws, despite originating from similar data.

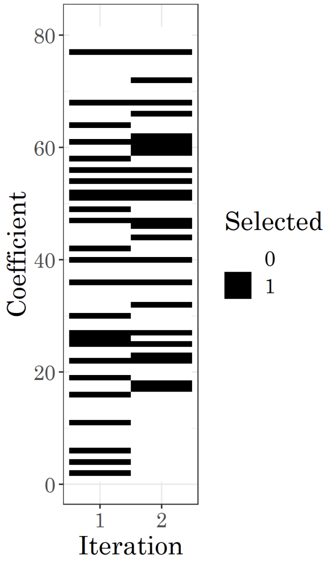


*Figure 16: Representation of the variables selected by the logistic regression in each of the 2 draws. Each column is one of the draws and each row is one of the variables. Black bars represent variables that are selected.*

Importantly, despite the two functions looking vastly different, their overall predictions indeed appear qualitatively similar. Figure 17 shows the distribution of default predictions by group for whites and non-whites, 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.

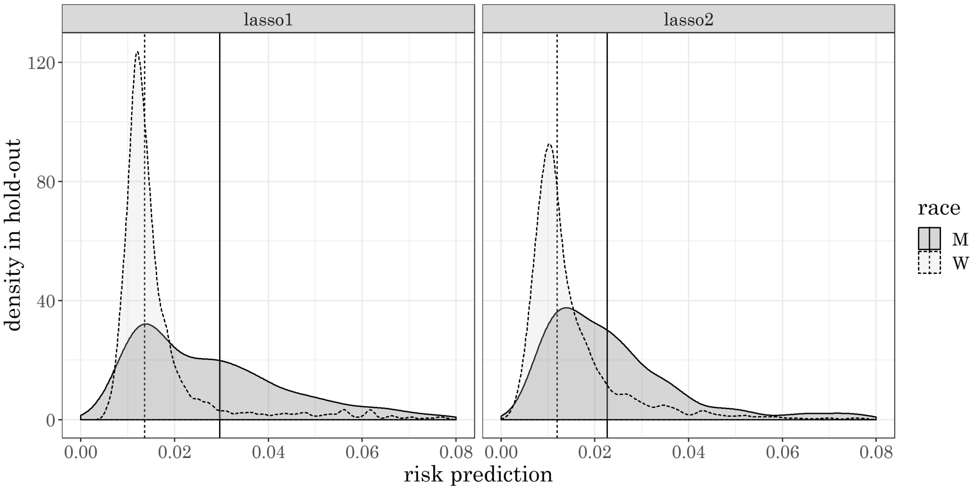


*Figure 17: Distribution of default risk using logistic lasso regression, plotted separately for white borrowers (W) and non-white borrowers (M). The two graphs show the distribution of the function based on two different draws from the full dataset. The vertical lines are the mean prediction by race.*

In addition, the accuracy of the prediction is quite similar when using the two training datasets. Figure 18 shows the ROC curves for the prediction using training dataset 1 and training dataset 2. While not identical across all levels of true positive rates and false positive rates, they roughly provide the same prediction accuracy. This makes sense because the predictions are based on the draws from the same underlying population and should provide similar information with respect to the ability to predict out of sample.

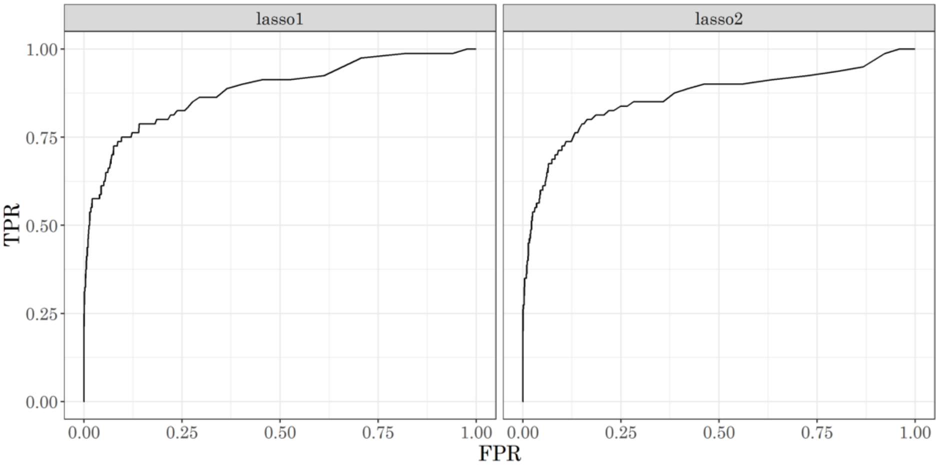


*Figure 18: ROC curves for risk prediction using training dataset 1 (left graph) and training dataset 2 (right graph).*

The conclusion from this exercise is that there is limited expressiveness of the variables an algorithm uses and so we should be wary by putting weight on this selection.

This conclusion is at odds with the orthogonalization approach, which uses the selection of the protected variable and its weight to orthogonalize correlated variables. First, the orthogonalization approach relies on the selection of a protected characteristic to argue that its omission would lead to omitted-variable-bias. The analysis above, however, suggests that an algorithm may or may not select a protected characteristic without this directly relating the underlying model. Second, the orthogonalization approach relies on recovering the "true" effect of a protected characteristic on the prediction in the full model. The analysis, above, however, suggests that the weight the lasso regression puts on a variable should not interpreted as some truth with respect to the significance it plays in the prediction of the outcome.

Finally, it is important to note that even when the prediction and accuracy stay stable across draws, this does not mean that each individual is treated the same. Figure 19 compares the risk prediction under each of the two functions, showing the variation for many borrowers.

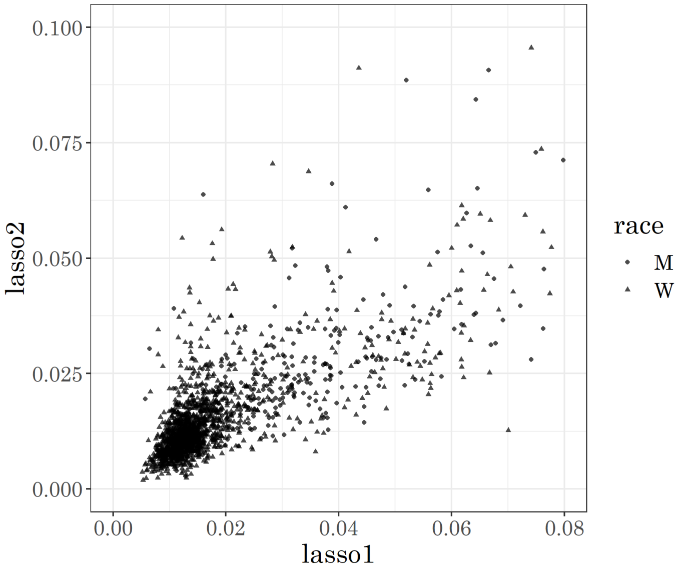


*Figure 19: Risk prediction under logistic lasso regression using training dataset 1 and training dataset 2. Each dot represents a different borrower and the predicted risk under either function.*

## APPENDIX D: IMPLEMENTING A TEST FOR SIMILARLY SITUATED

To implement the similarly situated test, we would need to first define which variables are part of the "similarly situated" set and then measure disparities controlling for this set. One way to do this is to predict default risk from the "similarly situated" set only and then compare this distribution to the distribution from the full set. This would be a comparison between two distributions, so the regulator would need to set the metric used to compare the distributions and the "tolerance level," meaning how much the two distributions are allowed to differ.[10] Setting this tolerance level is one way a regulator could adjust the test depending on the weight it gives to accuracy versus fairness.[11]

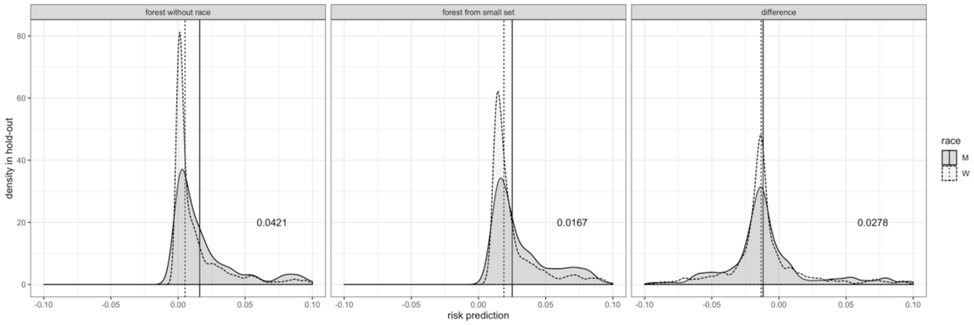


*Figure 20: Disparities controlling for "similarly situated" borrowers. The graph on the left is the risk prediction using all variables other than race, plotted separately for white and non-whites (the same as Figure 13). The graph in the middle is the risk prediction using "similarly situated" variables, plotted separately for white and non-whites. The graph on the right is the residual of the graph on the left and the middle graph, plotted separately for whites and non-whites. The numbers in the lower right corner are the "Wasserstein distances," meaning the difference between the distribution for whites and non-whites.*

Figure 20 shows an example of how a regulator can consider whether similarly situated borrowers are treated the same.[12] The graph on the right

[10] On possible metric is the "Wasserstein distance" metric, which measures the distance between two distributions. It could also be that the regulator wishes to focus only on part of the distribution.

[11] *See* Corbett-Davies et al., *supra* note 45, at 6 ("satisfying common definitions of fairness means one must in theory sacrifice some degree of public safety"). A full discussion of the question of how to define who is similarly situated is beyond the scope of this Article. However, it is important to note that the decision of what to include in this set is likely to be crucial in whether disparities amount to discrimination. First, the size of this set essentially determines the extent to which observed differences are used to explain raw disparities. Second, even though there is a natural tendency to include classic credit pricing variables in the similarly situated set, their inclusion could lead us to overlook the most significant discrimination.

[12] In this example, I defined the similarly situated set to include some variables that are typically used to price credit today – income, debt-to-income ratio, and characteristics of the

shows the residual from the prediction of the full set of inputs (graph on the left) and the prediction from the "similarly situated" set (graph in the middle). Intuitively, the graph on the right is the difference between white and non-whites, controlling for "similarly situated" characteristics.

loan. This is just one example of how the regulator could define this set. How large the set is, and what variables are included, will affect the difference of the residual.