Article

The Input Fallacy

Talia B. Gillis[†]

[†] Associate Professor, Columbia Law School. Previously circulated under the title “False Dreams of Algorithmic Fairness: The Case of Credit Pricing.” Winner of the 2022 Association of American Law Schools (AALS) Scholarly Paper Competition. For comments and suggestions, I thank Oren Bar-Gill, John Beshears, Doni Bloomfield, Sarith Felber, Noah Feldman, Jesse Fried, Becca Goldstein, Jon Gould, Deborah Hellman, Howell Jackson, Guha Krishnamurthi, David Laibson, Ela Leshem, Natasha Sarin, Joshua Simons, Holger Spamann, Jann Spiess, Rafi Stern, Cass Sunstein, Oren Tamir, Rory Van Loo and Crystal Yang. I also thank Jenny Li for excellent research assistance.

INTRODUCTION

Algorithms are everywhere on the rise. In a wide range of domains, from screening resumes[1] to determining criminal justice outcomes,[2] automated decision-making using advanced prediction technologies and big data has replaced human decision-making. Consumer credit, too, relies increasingly on machine learning algorithms[3] and nontraditional data.[4]

These technologies have improved efficiency and accuracy. But they have also generated concern about bias.[5] Bias, a term used to

[1] *See* Josh Bersin, *Big Data in Human Resources: Talent Analytics (People Analytics) Comes of Age*, FORBES (Feb. 17, 2013), http://www.forbes.com/sites/joshbersin/2013/02/17/bigdata-in-human-resources-talent-analytics-comes-of-age [https://perma.cc/FN3H-BMQS]; 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 [https://perma.cc/3RAJ-AP9H].

[2] *See* Ed Yong, *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 [https://perma.cc/JBG2-NZV9].

[3] *See infra* Part II.A.3.

[4] *See infra* Part II.A.1; *see also* 84 C.F.R. § 32420 (2019) (describing the recent Fair Lending Report of the Consumer Financial Protection Bureau (CFPB), released on June 28, 2019 and a symposium the Bureau held in which participants "discussed the role alternative data and modeling techniques can play in expanding access to traditional credit").

[5] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 677 (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.

describe unfairness to a vulnerable population or legally protected group,[6] can occur in algorithms for several reasons. It can result from training an algorithm with nonrepresentative data,[7] from predicting a human decision that is biased,[8] or from imperfectly measuring the outcome of interest.[9] One particular source of concern is that we might be using characteristics or "inputs" that are biased. This con-

148, 168 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 874 (2016); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2233, 2251 (2019); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395, 402 (2018); *see also* Megan Smith, DJ Patil & Cecilia Muñoz, *Big Risks, Big Opportunities: The Intersection of Big Data and Civil Rights*, WHITE HOUSE: BLOG (May 4, 2016), https://obamawhitehouse.archives.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights [https://perma.cc/2866-4LVN].

[6] *See* Mayson, *supra* note 5, 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 5, at 887 ("Similarly, data mining models built using biased, error-ridden, or unrepresentative data may be statistically biased.").

[7] *See* Hurley & Adebayo, *supra* note 5, at 178 ("If credit scorers rely on nonneutral 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 scorer's 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 Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003,* FED. TRADE COMM'N iv (Dec. 2012), https://www.ftc.gov/sites/default/files/documents/reports/section-319-fair-and-accurate-credit-transactions-act-2003-fifth-interim-federal-trade-commission/130211factareport.pdf [https://perma.cc/K2VB-PLM3]. 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*, REUTERS (Oct. 10, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G [https://perma.cc/T4NT-WUUD].

[8] *See, e.g.*, Bruckner, *supra* note 5, at 26 (discussing an example in which an algorithm was set up to predict admissions decisions using a training set that was created by biased admissions officers).

[9] 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 5, 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, Jon Kleinberg, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*, *in* PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 275, 278 (Ass'n for Computing Machinery ed., 2017), https://dl.acm.org/doi/10.1145/3097983.3098066 [https://perma.cc/N48T-BDQX] (developing a method to overcome the problem of selective labels).

cern is important in the credit context, on which this Article focuses. There, inputs might be biased because they reflect preexisting disadvantages or replicate biased measurements of borrower characteristics.[10] The role of law in addressing these concerns remains contested.

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. 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.[11] HUD's Proposed Rule on the Implementation of the Fair Housing Act's Disparate Impact Standard[12] was one of the first attempts in the United States and worldwide to create concrete rules to determine whether an algorithm violates fair lending law. This attempt ultimately failed, and the sections relating to algorithmic decisions were omitted from the Final Rule, published on September 24, 2020.[13] But HUD made clear that it "expects that there will be further development in the law in the emerging technology area of algorithms."[14] A recent interagency Request for Information on the use of artificial intelligence (AI) in finance also indicates a regulatory focus on algorithmic lending.[15] The request discusses the use of AI and alternative data in credit decisions and the challenges in establishing that algorithmic lending is consistent

[10] *See infra* Part I.B.

[11] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42,854 (proposed Aug. 19, 2019) (to be codified at 24 C.F.R. pt. 100).

[12] *Id.*

[13] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 85 Fed. Reg. 60,288 (Sept. 24, 2020) (codified at 24 C.F.R. pt. 100). The Proposed Rule was highly problematic. 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," HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. at 42,855 (quoting Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 576 U.S. 519, 520–33 (2015)), it is confused and contradictory and reflects a lack of basic understanding of the technology at play. *See* Lorena Rodriguez, *All Data Is Not Credit Data: Closing the Gap Between the Fair Housing Act and Algorithmic Decisionmaking in the Lending Industry*, 120 COLUM. L. REV. 1843, 1878–79 (2020) (arguing that HUD's proposed rule is inconsistent with the Supreme Court's decision in *Inclusive Communities*).

[14] HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 85 Fed. Reg. at 60,290.

[15] Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 86 Fed. Reg. 16,837 (Mar. 31, 2021).

with fair lending laws.[16] Algorithmic lending has also been flagged as a key domain for future regulation outside the United States. The European Union's proposal for the regulation of AI, revealed in April 2021, presents an ambitious attempt to regulate AI across many domains, and specifically designates creditworthiness assessments as a "high-risk" domain to be more heavily scrutinized and regulated.[17]

The stakes in developing such a law are high. The allocation of credit has been historically distorted by discriminatory policies and practices.[18] From redlining[19] to other forms of financial exclusion,[20] discriminatory practices have prevented racial minorities from being equal participants in credit markets.[21] Credit pricing always risks perpetuating this inequality because a lender's risk assessment is backward-looking, in that it considers the historical lending behavior of groups and individuals. And indeed, existing lending practices have often perpetuated historical injustices in that way, leaving millions of consumers without access to credit,[22] including a disproportionate number of Black consumers.[23] Algorithmic credit pricing

[16] *Id.* at 16,841 ("[I]t may be challenging to verify that a less transparent and explainable approach comports with fair lending laws.").

[17] *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF [https://perma.cc/H5MA-6W7V] [hereinafter European Union Proposal].

[18] *See, e.g.*, Harold Black, Robert L. Schweitzer & Lewis Mandell, *Discrimination in Mortgage Lending*, 68 AM. ECON. REV. 186, 189 (1978) ("[R]ace is an important determinant in the loan decision . . . ."); Helen F. Ladd, *Evidence on Discrimination in Mortgage Lending*, 12 J. ECON. PERSPS. 41, 45 (1998) ("In the past, mortgage lenders have clearly discriminated against some groups of borrowers and much of the discrimination was overtly part of their policy guidelines.").

[19] *See* Michael H. Schill & Susan M. Wachter, *The Spatial Bias of Federal Housing Law and Policy: Concentrated Poverty in Urban America*, 143 PA. L. REV. 1285, 1309 (1995) (showing that appraisal maps of the Federal Home Loan Bank Board determined that areas with even a small Black population receive the lowest rating).

[20] *See* MEHRSA BARADARAN, THE COLOR OF MONEY: BLACK BANKS AND THE RACIAL WEALTH GAP (2017) (documenting how the creation of Black banks further contributed to the wealth gap in the United States).

[21] Importantly, part of this exclusion is a result of unequal credit terms. *See* KEEANGA-YAMAHTTA TAYLOR, RACE FOR PROFIT: HOW BANKS AND THE REAL ESTATE INDUSTRY UNDERMINED BLACK HOMEOWNERSHIP 257 (2019).

[22] The Federal Deposit Insurance Corporation (FDIC) estimates that in 2019, 7.1 million households were unbanked. *See* FDIC, HOW AMERICA BANKS: HOUSEHOLD USE OF BANKING AND FINANCIAL SERVICES 12 (2019). My primary concern is the exclusion from non-predatory credit markets.

[23] *See* Rory Van Loo, *Making Innovation More Competitive: The Case of Fintech*, 65 UCLA L. REV. 232, 254 (2018) (discussing how Black and Latino households are

greatly improves the ability of lenders to assess credit risk.[24] This improvement carries both danger and promise for fair lending. The danger is that algorithms' greater ability to analyze and distinguish people based on past lending behavior will further replicate and even exacerbate past injustices.[25] The promise is that algorithms' improved accuracy in predicting creditworthiness will increase the availability of credit for formerly excluded consumers and disadvantaged groups.[26] The ambition of this Article is to help fair lending realize this promise.[27]

Against this backdrop I advance two arguments. First, I argue that the leading approaches to algorithmic discrimination are misguided, even on their own terms. These approaches commit what I call "the input fallacy" in that they hold on to the input-focused view of traditional fair lending, even though machine learning pricing makes this view obsolete. The input fallacy creates an algorithmic myth of colorblindness[28] by fostering the false hope that input exclu-

more than twice as likely to be unbanked compared to the national average).

[24] *See* Part II.A.2.

[25] *See* Part II.C.

[26] For a critical perspective on the focus on access to credit among low-income consumers, see Abbye Atkinson, *Rethinking Credit as a Social Provision*, 71 STAN. L. REV. 1093, 1099 (2019), arguing that credit, which shifts consumption temporally, is only beneficial if income increases in the future. Today, however, "credit is fundamentally incompatible with the entrenched intergenerational poverty that plagues low-income Americans." *Id.* Although it is important to recognize that credit is not a panacea for all economic struggles, and particularly wage stagnation, and that credit is not beneficial to all consumers at all prices, affordable credit can still play an important role in the creation of wealth. *See* Mehrsa Baradaran, *Banking and the Social Contract*, 89 NOTRE DAME L. REV. 1283, 1336 ("Access to safe credit is crucial in allowing the poor to escape poverty."); *see also The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings,* FINREGLAB 32, 34 (Jul. 2019), https://finreglab.org/wp-content/uploads/2019/07/FRL_Research-Report_Final.pdf [https://perma.cc/N65Q-WQML] (showing that the use of cash flow data can allow prediction of default risk, providing an alternative or supplement to traditional credit scores and also that the cash-flow data were consistently predictive across demographic groups).

[27] *See, e.g.*, Richard R.W. Brooks, *Credit Past Due*, 106 COLUM. L. REV. 994, 999–1003 (2006) (arguing that poor communities are excluded from many credit markets because fringe credit lending is not reported to credit agencies). The persistence of discriminatory practices can be seen in other consumers domains. *See, e.g.*, Ian Ayres, *Fair Driving: Gender and Race Discrimination in Retail Car Negotiations*, 104 HARV. L. REV. 817, 819 (1991) (documenting discrimination in the sale of cars); Rory Van Loo, *A Tale of Two Debtors: Bankruptcy Disparities by Race*, 72 ALB. L. REV. 231, 232 (2009) (finding that Black debtors fare worse in bankruptcy).

[28] *See* David A. Strauss, *The Myth of Colorblindness*, 1986 SUP. CT. REV. 99, 113 (arguing that "race-consciousness, not color-blindness, is the basis of the prohibition against discrimination").

sion can create non-discriminatory algorithms. Moreover, when input-focused approaches exclude a broad set of inputs, they risk turning fair lending law into a weapon that entrenches the status quo and undermines the promise of algorithmic credit pricing to create a more inclusive credit market. Second, I argue that we should instead explore ways to expand and emphasize regulatory output analysis through empirical testing of algorithmic outcomes. This outcomes-based approach allows us to face up to the tradeoffs that algorithmic credit pricing necessarily entails. It enables us to weigh the danger of disparate credit allocation against the promise of increased credit access for marginalized groups.

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 rich Boston Fed Home Mortgage Disclosure Act (HMDA) dataset,[29] which contains information on mortgage applications, with simulated default rates disciplined by information on the loans.[30] My hypothetical lender uses a machine learning algorithm to predict default probability, which is then used to price credit for future borrowers. In this simulation exercise, the loan and borrower characteristics serve as the "inputs" to the credit decisions, while the predicted default probability is the "output."

My Article provides discussions of algorithmic bias with new structure and clarity by distinguishing among different types of input biases. Current discussions tend to overlook that even traditional credit pricing relies on borrower characteristics that reflected preexisting disadvantage ("biased world" inputs)[31] or were inaccurately measured ("biased measurement" inputs).[32] In the algorithmic con-

[29] *Home Mortgage Disclosure Act (HMDA) Data for New England*, FED. RSRV. BANK OF BOS. (Oct. 29, 2018), https://www.bostonfed.org/data/data-items/home-mortgage-disclosure-act-hmda-data-for-new-england.aspx [https://perma.cc/Q4Q4-XRPX].

[30] As explained in Part II.B 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.

[31] *See infra* Part I.B.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.

[32] *See infra* Part I.B.2. If, for example, credit scores only consider certain types of creditworthiness indicators, such as timely loan payments, but do not consider

text, as my empirical simulations will demonstrate, the use of biased inputs can increase disparities in some instances while actually decreasing them in others.

Fair lending law is the primary lens to determine whether disparities in traditional credit pricing amount to discrimination. 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.[33] The dominant method for determining whether lender pricing amounts to discrimination has been to scrutinize decision inputs.[34] This has been true not only for disparate treatment, but also—despite its name—for disparate impact.[35] And even though traditional credit pricing was based on few inputs and involved human discretion, scholars have tried to extend this dominant method of input scrutiny to the algorithmic context.[36]

This Article challenges three leading approaches to discrimination law in the algorithmic context that scrutinize inputs.[37] The first approach excludes protected characteristics, primarily as a method for negating a claim of intentional discrimination. Goldman Sachs, for instance, recently relied on such an approach when it responded to a complaint that a man received a credit line twenty times higher than

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* Bd. of Governors of the Fed. Rsrv. Sys., *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit*, FED. RSRV. S-2 (Aug. 2007), https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf [https://perma.cc/AQ8Z-RAVV] (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).

[33] *See infra* Part I.C.

[34] *See infra* Part I.C. For disparate treatment, 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 also focuses on analyzing decision inputs after an initial demonstration of the outcome disparities. 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.

[35] *See infra* Part I.C.

[36] *See* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459, 460 (2019); Hurley & Adebayo, *supra* note 5, at 183–84. This is also true of other areas of discrimination law. *See* Kim, *supra* note 5. *See generally* Barocas & Selbst, *supra* note 5, at 694.

[37] 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.

his wife[38] by arguing that it was not possible for Goldman Sachs to discriminate against her because its algorithms "do not know your gender" and do not make decisions "based on factors like gender."[39]

The problem with this first approach is that information about a person's protected characteristics 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,[40] from the other variables within the HMDA dataset.[41]

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.[42] The lender is aware, however, that older

[38] Neil Vigdor, *Apple Card Investigated After Gender Discrimination Complaints*, N.Y. TIMES (Nov. 10, 2019), https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html [https://perma.cc/99GH-3JWD].

[39] Shahien Nasiripour, Jennifer Surane & Sridhar Natarajan, *Apple Card's Gender-Bias Claims Look Familiar to Old-School Banks*, BLOOMBERG BUSINESSWEEK (Nov. 11, 2019), https://www.bloomberg.com/news/articles/2019-11-11/apple-card-s-ai-stumble-looks-familiar-to-old-school-banks [https://perma.cc/3NTJ-P9LS].

[40] *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).").

[41] 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 Part II.A.1.

[42] 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. 15 U.S.C. § 1691(a). However, 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* 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 12 C.F.R. § 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 12 C.F.R. § 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." 12 C.F.R. Pt. 1002(6)(b)(2), supp. I. This is because with algorithmic pricing, unlike expert based scoring, the weights are not pre-assigned to different characteristics. Similarly, one must be wary of interpreting the weight on "age" as the true and stable contribution of that variable to a prediction. *See* Kathryn P. Taylor, *Equal Credit for All—An*

borrowers are different from other borrowers.[43] They often have less documented credit history and tend to use cash more frequently.[44] And it is also aware of course that an older person is less likely to live long enough to repay their loan before dying. Imagine that the lender then applies a machine learning algorithm to predict borrower default risk from the borrower's Amazon purchase history. Given the close relationship between age and default risk, the algorithm will recover the borrower's age based on their purchase history, even though the lender formally excluded age from the algorithm. The exclusion of protected characteristics thus creates a meaningless façade of neutrality.

We should also be wary of excluding protected characteristics if we care about outcome disparities.[45] As I demonstrate through a simulated example, price disparities can actually decrease when algorithms are "race aware." This is because a characteristic may need to be interpreted differently for various racial groups.[46] Conversely, we may increase disparities when we exclude the race variable because we are imposing a similar interpretation of a characteristic for both white and non-white applicants.

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.[47] The approach, however, is

Analysis of the 1976 Amendments to the Equal Credit Opportunity Act, 22 ST. LOUIS U. L.J. 326, 338 (1978) ("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.

[43] A recent report by Deloitte shows how age is one of the most important factors in black box AI models of credit risk. *See Explain Artificial Intelligence for Credit Risk Management*, DELOITTE 4 (Apr. 2020), https://www2.deloitte.com/content/dam/Deloitte/fr/Documents/risk/Publications/deloitte_artificial-intelligence-credit-risk.pdf [https://perma.cc/6XJR-WFFL].

[44] *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).

[45] As discussed further in Part III, the exclusion of protected characteristics may be considered a fair procedure, regardless of 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). 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.

[46] *See infra* Part III.A.2.

[47] HUD Proposed Rule 2019 is an example of an attempt to formally incorporate this position. In HUD's circulated draft of its Proposed Rule, a lender can defend

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 we can recover a borrower's race. Using the HMDA data, I demonstrate that there is a greater ability to predict "race" from the 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, the combination of many consumer behaviors can still reveal borrower age.

The third approach I discuss restricts algorithm inputs to pre-approved features. It thus differs from the first two approaches, which allow all inputs other than certain forbidden features. Although this third approach may allow for greater control over what algorithms use to price credit, it does not guarantee a reduction in disparities. Moreover, it risks restricting access to credit by limiting an algorithm to traditional credit pricing inputs and further perpetuating the exclusion of consumers lacking formal credit histories, which are disproportionately racial minorities.[48]

The three approaches share a common fallacy. They scrutinize decision inputs, even though such scrutiny is no longer feasible or effective in the algorithmic context. In committing this input fallacy, they remain focused on two causal questions that lie at the heart of traditional fair lending: 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).[49] However, machine learning is a world of correlation and not causation.

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." *See* HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42,854, 42,862 (proposed Aug. 19, 2019) (to be codified at 24 C.F.R. pt. 100).

[48] *See* CFPB Off. of Rsch., *Data Point: Credit Invisibles*, CFPB 6 (May 2015), https://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf [https://perma.cc/8UNG-B5C9] ("Blacks and Hispanics are more likely than Whites or Asians to be credit invisible or to have unscored credit records.").

[49] *See, e.g.*, Sheila R. Foster, *Causation in Antidiscrimination Law: Beyond Intent Versus Impact*, 41 HOUS. L. REV. 1469, 1472 (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 Part IV.A.

Instead of continuing to commit the input fallacy, fair lending law must shift to outcome-focused analysis.[50] For when it is no longer possible to scrutinize inputs, outcome analysis provides the only way to evaluate whether a pricing method leads to impermissible disparities. This is true for the legal doctrine of disparate impact, which has always cared about outcomes, even when it did so by scrutinizing inputs.[51] And it is also true for disparate treatment, a doctrine that has historically been quite detached from disparate outcomes.[52] In the algorithmic context, both can no longer rely on input scrutiny but must analyze outcomes.

I end the Article by proposing a testing method that regulators should use to analyze the discriminatory effects of algorithmic pricing rules. My testing method applies a credit pricing rule to a dataset of hypothetical borrowers. Regulators can then examine the outcomes of the pricing rule to determine whether the pricing rule discriminates. This method of outcome-focused testing resembles the first stage of a disparate impact complaint in traditional fair lending law but adapts it to the machine learning context.

Because the criteria for determining discrimination continue to be disputed, I do not provide an exact test. Instead, I show that my testing method for algorithmic outcomes can answer meaningful questions. The first such question is whether the pricing rule treats borrowers who are "similarly situated" equally. The second question is whether the pricing rule increases or decreases disparities relative to some baseline, such as the non-algorithmic credit pricing method.

My outcome-focused test 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.[53] Similarly, we cannot reliably predict whether an algorithmic method will increase or decrease disparities by looking only at inputs.[54] By contrast, my

[50] 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. 1023, 1039 (2017) (reviewing FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015)) ("The focus on outcomes rather than how an algorithm operates seems especially useful as algorithms become increasingly complicated, even able to modify themselves.").

[51] *See infra* Part I.C.

[52] *See infra* Part I.C.

[53] *See infra* Part I.B.

[54] *See infra* Part I.B.

outcome-focused testing method can measure the actual effects of a credit pricing rule.[55] It can thus provide regulators with a workable and appropriate Regtech response to the Fintech industry, by deploying technology to fight discrimination.[56]

My critique of input-based approaches and my proposal of outcome-focused tests chart a course for discrimination law in an algorithmic world beyond just the context of fair lending. Both speak more broadly to the challenges of enforcing anti-discrimination law in algorithmic contexts, from employment to criminal justice.[57] The Article also contributes to discussions in the computer science and statistical literature on algorithmic fairness, by demonstrating how legal doctrine and regulatory realities should inform our evaluations of algorithmic decisions.[58]

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 a fundamental shift in fair lending law

[55] *See infra* Part IV.A.

[56] For an example of an attempt by the CFPB to regulate through technology and a discussion of how the CFPB leverages digital tools to attempt to help consumers find credit, see Rory Van Loo, *Rise of the Digital Regulator*, 66 DUKE L.J. 1267, 1304 (2017).

[57] Other papers discuss the application of discrimination in other areas of law. *See* Kim, *supra* note 5 (employment discrimination); Daniel Westreich & James Grimmelmann, *Incomprehensible Discrimination*, 7 CALIF. L. REV.: ONLINE 164 (Apr. 2017), (criminal justice discrimination); Allan G. King & Marko J. Mrkonich, "*Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 563 (2016) (employment discrimination). *See generally* Chander, *supra* note 50, at 1024 (describing the impact of algorithmic decision-making in numerous areas).

[58] *See, e.g.*, Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold & Richard Zemel, *Fairness Through Awareness*, *in* PROCEEDINGS OF THE 3RD INNOVATIONS IN THEORETICAL COMPUTER SCIENCE CONFERENCE 214, 214 (Ass'n for Computing Machinery ed., 2012) https://dl.acm.org/doi/10.1145/2090236.2090255 [https://perma.cc/T2U3-TNQ6] (discussing an individual fairness approach to algorithmic fairness); *see also* Sam Corbett-Davies & Sharad Goel, *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*, ARXIV, Aug. 14, 2018, at 1, https://arxiv.org/pdf/1808.00023 [https://perma.cc/4YEM-7747]. For a recent survey of the literature, see Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, ACM COMPUTING SURVS., Jul. 2022 at 1.

from input scrutiny to outcome analysis and develops an empirical method for outcome analysis.

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, 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 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 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.[59]

A. THE CREDIT PRICING DECISION

Credit contracts are often personalized,[60] meaning that lenders will determine the specific terms of the contract based on the charac-

[59] 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 generally* GARY S. BECKER, THE ECONOMICS OF DISCRIMINATION (2010). I focus on the problem of biased inputs, first, 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 Part II.C.

[60] 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,

teristics 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.[61]

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. Because credit contracts require an upfront transfer of money for a future promise of payments, lenders face challenges of asymmetric information as to the borrower's willingness and ability to repay a loan, adverse selection, and moral hazard.[62] One way to overcome these challenges is to price the risk through interest rates and other terms[63] and through an assessment of creditworthiness, which is essentially a prediction about future borrower behavior and finances. 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 based on a borrower's FICO score.[64] The borrower's income and future income are 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, such as the interest rate. The exact terms of the

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.

[61] 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, Andreas Fuster & Aurel Hizmo, *Paying Too Much? Price Dispersion in the US Mortgage Market* (Bd. of Governors of the Fed. Reserve Sys., Working Paper No. 2020-062, 2020), https://www.federalreserve.gov/econres/feds/paying-too-much-price
-dispersion-in-the-us-mortgage-market.htm [https://perma.cc/67TG-9HC2].

[62] *See* George Akerlof, *The Market for "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q. J. ECONS. 488, 488 (1970).

[63] A loan's interest rate is only one term through which to consider the cost of a loan. Many other fees, such as closing fees, also increase the cost of the loan.

[64] Developed in 1989 by Fair, Isaac and Company, the standard FICO score was meant to create a generic model that would allow for comparing the reports of the various credit reporting agencies. This standardized system for scoring consumers quickly became the industry standard credit score used today. *See* Shweta Arya, Catherine Eckel & Colin Wichman, 96 J. ECON. BEHAV. & ORG. 175, 175 (2013).

loan vary greatly across borrowers; therefore, there is a degree of personalization of the prices paid by borrowers.[65]

Credit terms are also personalized because they are partially determined by lender employees or brokers (jointly "loan officers") who have discretion.[66] 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 loan.[67]

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.[68]

[65] In the U.S., there is in fact significant variation in the cost of credit for different borrowers, expressed by the variation in credit terms such as the interest rate and fees of the loan. Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai & Angsar Walther, *Predictably Unequal? The Effects of Machine Learning on Credit Markets*, 76 J. FIN. (forthcoming 2022) (manuscript at 15, Table I) (on file with author) (providing a description of the variation in interest rates).

[66] This is because mortgage lenders often create borrower "bins" based on a limited set of characteristics in determining par rates. These bins are often not based on sophisticated risk predictions but rather reflect more coarse divisions between lenders. As I have discussed elsewhere, it is not clear how exactly loan officers decide the final terms of the loan. For example, it is largely unknown whether loan officers are concerned with assessing credit worthiness or trying to learn a borrower's willingness-to-pay. *See* Gillis & Spiess, *supra* note 36.

[67] 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 tied to a loan's interest rate. *See* Truth in Lending (Regulation Z), 75 Fed. Reg. 58,505–08 (proposed Sept. 24, 2010) (codified at 12 C.F.R. pt. 226). Even absent direct compensation for higher interest rates, more expensive loans are clearly more profitable for lenders and could ultimately affect loan officer compensation. *See* Howell E. Jackson & Laurie Burlingame, *Kickbacks or Compensation: The Case of Yield Spread Premiums*, 12 STAN. J.L. BUS. & FIN. 289, 289 (2007) (discussing how yield spread premiums lead to higher mortgage prices for consumers, which may fall disproportionately on the least sophisticated borrowers).

[68] One can, in theory, separate the "prediction" problem from the "decision" problem. *See, e.g.*, Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel & Aziz Huq, *Algorithmic Decision Making and the Cost of Fairness*, *in* PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING 797, 797 (Ass'n for Computing Machinery ed., 2017) https://dl.acm.org/doi/10.1145/3097983.3098095 [https://perma.cc/X7LJ-Q2BA].

Despite my focus on default probability, in reality, default prediction is rarely the only metric used to personalize credit contracts. Personalization could reflect wheth-

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. When a lender can accurately predict default, they can determine a cutoff for extending a loan or price risk accordingly.[69] Flat pricing, by contrast, can create significant harm because of the adverse selection of less creditworthy borrowers who will choose to pay the higher interest rate,[70] which can, in turn, lead to the drying up of credit markets altogether. Moreover, default and foreclosure are costly for both lenders and consumers.[71]

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 that even

er the loan is securitized or the purpose of the loan, as well as 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 suggesting that many borrowers overpay for mortgages because they do not shop around or negotiate for a better rate. *See* Bhutta et al., *supra* note 61. I focus on default prediction personalization since this is arguably the least controversial basis for personalization. *See, e.g.*, Robert Bartlett, Adair Morse, Richard Stanton & Nancy Wallace, *Consumer Lending Discrimination in the FinTech Era* 50 (Nat'l Bureau of Econ. Rsch., Working Paper No. 25,943, 2019).

The significance of price discrimination is likely to increase in the future. *See The Effects of Online Disclosure About Personalized Pricing on Consumers* 7 (Org. for Econ. Coop. and Dev., Working Paper No. 303, 2021) ("[T]he quantity of personal data held on online consumers, combined with the increasing prevalence of personalization in other domains (e.g., advertisement), means there is at least potential for online personalized pricing to become more commonplace and more sophisticated.").

[69] This is often referred to as "risk based pricing." *See* Robert Phillips, *Optimizing Prices for Consumer Credit*, 12 J. REVENUE & PRICING MGMT. 360, 365 (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*, 11 J.L. ECON. & POL'Y 33, 33 (2015).

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

[71] *See* John Gathergood, Benedict Guttman-Kenney & Stefan Hunt, *How Do Payday Loans Affect Borrowers? Evidence from the U.K. Market*, 32 REV. FIN. STUD. 496, 496 (2019) (showing that payday loans cause persistent increases in defaults and cause consumers to exceed their bank overdraft limits).

riskier borrowers can receive a loan at a certain cost.[72] This is particularly likely to be the case in the tails of default prediction, i.e., for people with a higher probability of default.

B. 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.[73] 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 I.C, fair lending prohibits discrimination on the basis of race, religion, sex, marital status and age, among other grounds.

1. Biased World

Lenders seeking to personalize credit terms to borrowers confront the problem that many of the factors used to determine individual risk are a product of pre-existing disadvantage or discrimination.[74] Although this is not the lender's fault, using these inputs exacerbates the effects of existing discrimination in a new domain. There is no consensus on whether the use of biased world inputs gives rise to discrimination claims.[75]

[72] *See* Liran Einav, Mark Jenkins & Jonathan Levin, *The Impact of Credit Scoring on Consumer Lending*, 44 RAND J. ECON. 249, 249 (2013) (showing that the adoption of automated credit scoring at a large auto finance company led to higher-risk applicant lending).

[73] The two Acts that determine the protected groups for fair lending are the ECOA, *see* 15 U.S.C. § 1691(a)(1)-(2), and the Fair Housing Act, *see* 42 U.S.C. §§ 3604–07.

[74] I use the term "discrimination" here to describe a reality in which a group is 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 5, at 691.

[75] *See infra* Part I.C. According to some theories of discrimination, the use of

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 United States.[76] These gaps may be a result of "pre-market factors," such as reduced access to higher education, or a result of labor market discrimination.[77] Similarly, higher rates of incarceration of racial

"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." Deborah Hellman, *Indirect Discrimination and the Duty to Avoid Compounding Injustice*, *in* FOUNDATIONS OF INDIRECT DISCRIMINATION LAW 107 (Hugh Collins & Tarunabh Khaitan eds., 2017).

[76] *See* Kayla Fontenot, Jessica Semega & Melissa Kollar, *Income and Poverty in the United States: 2017*, U. S. CENSUS BUREAU 6, 8–9 (Sept. 2018), https://www.census.gov/content/dam/Census/library/publications/2018/demo/p60-263.pdf [https://perma.cc/MGM2-SXXE]. Importantly, the Black-white wage gap has increased as wage inequality has risen from 2000 to 2018. *See* Elise Gould, *State of Working America: Wages 2018*, ECON. POL'Y INST. 4 (Feb. 20, 2019), https://www.epi.org/publication/state-of-american-wages-2018 [https://perma.cc/K2JR-Q992].

[77] 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 the borrower's future income is less stable. *See* Alicia H. Munnell, Geoffrey M. B. Tootell, Lynn E. Browne & James McEneaney, *Mortgage Lending in Boston: Interpreting HMDA Data*, 86 AM. ECON. REV. 25, 29 (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); Todd J. Zywicki & Joseph D. Adamson, *The Law and Economics of Subprime Lending*, U. COLO. L. REV. 1, 9 (2009) (arguing that Black workers with the same ability and education earn less than comparable white workers or have fewer employment opportunities).

While the racial wage gap in the labor market is well documented, interpreting this gap and the extent to which it reflects either taste-based or statistical discrimination has proven difficult. *See* Dan Black, Amelia Haviland, Seth Sanders & Lowell Taylor, *Why Do Minority Men Earn Less? A Study of Wage Differentials Among the Highly Educated*, 88 REV. ECON. & STAT. 300, 300 (2006) (finding substantial wage gaps between Black men and men of other races and discussing challenges in attributing gap to 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) (finding 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); Roland G. Fryer Jr., Devah Pager & Jörg 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, 991 (2004) minorities could have a negative impact on credit scores as well.[78] Levels of debt might also 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.[79] There is also evidence that credit card lenders may screen for minority consumers.[80]

When a lender uses variables that reflect pre-existing disadvantage, or a biased world, it compounds that disadvantage by carrying it into the new domain of lending. The biased 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 creating wealth, this discrepancy in credit pricing risks reinforcing wealth gaps in the United States.

2. Biased Measurement

Many inputs into a pricing decision partially reflect measurement bias, meaning that the way in which an input is defined or estimated is biased rather than the underlying characteristic. 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, meas-

(finding that white-sounding names triggered a callback rate that was 50% higher than that of equally qualified applicants with Black-sounding names); *see also* John M. Nunley, Adam Pugh, Nicholas Romero & R. Alan Seals, *Racial Discrimination in the Labor Market for Recent College Graduates: Evidence from a Field Experiment*, 15 B.E. J. ECON. ANALYSIS & POL'Y 1097, 1097 (2015).

[78] A recent paper documented the negative impact of incarceration on credit scores and income. *See* Abhay P. Aneja & Carlos F. Avenancio-León, *No Credit for Time Served? Incarceration and Credit-Driven Crime Cycles*, 109 AEA PAPERS AND PROC. 161 (2019). 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.

[79] *See* Oren Bar-Gill & 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, 241 (2011) (arguing that subprime lending was incontrovertibly steered toward minority communities); Creola Johnson, *The Magic of Group 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).

[80] *See* Andrea Freeman, *Payback: A Structural Analysis of the Credit Card Problem*, 55 ARIZ. L. REV. 151, 180–81 (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.").

urement, and estimation. For example, if we want to use a borrower's income, we must define what income is and how to calculate 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.[81] 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 "measurement bias."[82]

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.[83]

In Subsection III.A.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 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

[81] 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 42, at 339.

[82] 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. *See* Corbett-Davies & Goel, *supra* note 58 at 18 (arguing that label bias is the more severe bias).

[83] 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 5, at 876.

or borrowing from smaller and more local financial institutions, may also be predictive of default.[84]

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 I.C.[85]

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 why 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.

C. 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

[84] 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, Jillian McLaughlin & Marina Levy, *Big Data: A Big Disappointment for Scoring Consumer Creditworthiness*, NAT'L CONSUMER L. CTR. 12 (Mar. 14, 2014), https://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf [https://perma.cc/W4SK-NXD5]. According to the CFPB, Black and Latino consumers are more likely to be credit invisible, at rates of around 15% in comparison to 9% for whites. *See* CFPB Off. of Rsch., *supra* note 48, at 24–25.

In September 2021, Fannie Mae announced that it will begin considering timely rental payments in underwriting calculations. *See* Ron Lieber, *Always Pay the Rent? It May Help Your Mortgage Application*, N.Y. TIMES (Sept. 11, 2021), https://www.nytimes.com/2021/09/11/your-money/paying-rent-mortgage.html [https://perma.cc/42DS-FS5T].

[85] See Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 1, 28–29, 33–34 (2018) for the distinction between "group differences in the raw data" and biases for the "choice of predictors."

discuss how the different positions view the problem of biased inputs, but I do not adopt a particular interpretation.

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.[86] The FHA prohibits discrimination in the terms of credit based on race, color, religion, sex, disability, familial status, and national origins.[87] In 1974, Congress passed the Equal Credit Opportunity Act (ECOA), banning discrimination in all types of credit transactions. ECOA 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.[88]

ECOA and FHA cover both discrimination doctrines of "disparate treatment," dealing with the direct condition of a decision on a protected characteristic, often with the 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 rec-

[86] In 1988, the Fair Housing Amendments Act was passed, strengthening the mortgage lending provisions of the FHA. *See* Raymond H. Brescia, *Subprime Communities: Reverse Redlining, the Fair Housing Act and Emerging Litigation Regarding the Subprime Mortgage Crisis*, 2 ALB. GOV'T L. REV. 164, 180–81 (2009).

[87] 42 U.S.C. § 3604 (2018) ("To discriminate against any person in the terms, conditions, or privileges of sale or rental of a dwelling, or in the *provision of services or facilities in connection therewith*, because of race, color, religion, sex, familial status, or national origin." (emphasis added)).

[88] There are other laws that have additional provisions relating to credit pricing discrimination that are not the focus of this Article. The Community Reinvestment Act of 1977 (CRA) 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. Pub. L. No. 95-128, 91 Stat. 1111 (codified as amended at 12 U.S.C. §§ 2901–08). The CRA does not give a right to private action but rather instructs the relevant supervisory agency on how to ensure that institutions are serving the lending needs of their community. Another federal law related to credit pricing discrimination is the Home Mortgage Disclosure Act of 1975 (HDMA). Pub. L. No. 94-200, 89 Stat. 1124 (codified at 12 U.S.C. §§ 2801–11), 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.

ognized 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*,[89] confirming the position of eleven appellate courts and various federal agencies, including HUD, the agency primarily responsible for enforcing the FHA.[90] 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.[91]

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.[92] The doctrine can 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."[93] Disparate treatment iden-

[89] Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 576 U.S. 519, 546–47 (2015).

[90] *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.").

[91] *See, e.g.*, Ramirez v. GreenPoint Mortgage Funding, Inc., 633 F. Supp. 2d 922, 926–27 (N.D. Cal. 2008); *CFPB Consumer Laws and Regulations: Equal Credit Opportunity Act*, CFPB 1 (June 2013), https://files.consumerfinance.gov/f/201306_cfpb_laws-and-regulations_ecoa-combined-june-2013.pdf [https://perma.cc/GBE4-ZCWC] ("The ECOA has two principal theories of liability: disparate treatment and disparate impact."). During the Trump Administration, the CFPB proposed abandoning disparate impact liability under the ECOA. *See* Press Release, CFPB, Statement of the Bureau of Consumer Financial Protection on Enactment of SJ Res 57 (May 21, 2018) https://www.consumerfinance.gov/about-us/newsroom/statement-bureau-consumer-financial-protection-enactment-sj-res-57 [https://perma.cc/8MJS-63Q7] (stating that the CFPB will reexamine its guidance on disparate impact liability under the ECOA). 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*, 61 BUS. L. 829, 829 (2006).

[92] In the employment discrimination context, see Sullivan, *supra* note 5, at 408, suggesting that one way to read Title VII is that it "embraces a causal view of what we call disparate treatment".

[93] 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, e.g.*, Laufman v. Oakley Bldg. & Loan Co., 408 F. Supp. 489, 491 (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 pro-

tifies cases in which a protected characteristic directly influenced a credit decision and is therefore 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 disparate effect. A disparate impact case typically follows the burden-shifting framework that was developed primarily in the Title VII employment discrimination context.[94] At the first step of the framework, the plaintiff must make a prima facie showing of a disparate outcome for a protected group.[95] This requires the plaintiff to identify the specific conduct or policy that led to the disparate outcome. 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 the disparity.[96] 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 over whether the case law and regulatory actions are consistent with those foundations. One of the most important disagreements is over the extent to which disparate impact

tected 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. *See generally* Helen F. Ladd, *Evidence on Discrimination in Mortgage Lending*, 12 J. ECON. PERSPS. 41 (1998).

[94] Disparate impact first entered U.S. law in the 1971 breakthrough case *Griggs v. Duke Power Co.*, in which hiring requirements of a high school diploma and an aptitude test were challenged. 401 U.S. 424, 431–32, 436 (1971). A formal burden shifting framework was articulated in the subsequent employment decision *Albermarle Paper Co. v. Moody*, and this was articulated into the three-step burden-shifting approach that is applied today. 422 U.S. 405, 425 (1975). This burden-shifting framework was formalized into the language of Title VII in § 703(k), added by the Civil Rights Act of 1991. Similar language exists in HUD's 2013 Disparate Impact Rule. *See, e.g.*, HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 78 Fed. Reg. 11,460 (Feb. 15, 2013) (codified at 24 C.F.R. pt. 100); Regulation B, 12 C.F.R. § 202.6 n.2 (discussing the relevance of Title VII for interpreting fair lending disparate impact); *see also* Equal Credit Opportunity Act, 41 Fed. Reg. 29,870, 29,874 (July 20, 1976) ("Congress intended certain judicial decisions enunciating this 'effects test' from the employment area to be applied in the credit area.").

[95] *See Albermarle Paper Co*, 422 U.S. at 425.

[96] 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*, 167 U. PA. L. REV. 1375 (2019) (providing a detailed discussion of this burden-shifting framework in the context of employment discrimination).

is meant to address cases that are more about effect than intent.[97] 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.[98] This account emphasizes disparate impact's ability to unearth cases in which there is a discriminatory motive that is hard to prove.[99]

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."[100] This second theory has also characterized disparate impact as "disturbing in itself, in the sense that a practice that produces such

[97] For an articulation of these disagreements see Richard A. Primus, *Equal Protection and Disparate Impact: Round Three*, 117 HARV. L. REV. 494, 520 (2003). There are other debates around disparate impact, or "indirect discrimination," a similar doctrine in Europe and many other countries. *See generally* Mark MacCarthy, *Standards of Fairness for Disparate Impact Assessment of Big Data Algorithms*, 48 CUMB. L. REV. 67 (2018) (discussing the extent to which disparate impact represents a moral wrong or should be considered discrimination altogether).

[98] *See* Michael Selmi, *Was the Disparate Impact Theory a Mistake*, 53 UCLA L. REV. 701, 708 (2006) (tracing the origins and implementation of disparate impact in the context of Title VII to argue that it may have limited a more expansive theory of intent under disparate treatment theory); *see also* Nicholas O. Stephanopoulos, *Disparate Impact, Unified Law*, 128 YALE L.J. 1566 (2019) (discussing this theory in the context of voting discrimination).

[99] *See* Primus, *supra* note 97, 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 ("anti-classification") and scholars who assert that discrimination law targets practices that disadvantage groups or perpetuate disadvantage ("anti-subordination"). *See, e.g.*, Jack M. Balkin & Reva B. Siegel, *American Civil Rights Tradition: Anticlassification or Antisubordination?*, 58 U. MIA. L. REV. 9 (2003). Balkin and Siegel's primary focus is on the Equal Protection Clause, however, they point out that an anti-classification reading of Title VII disparate impact would view the doctrine as primarily concerned with implicit disparate treatment. *Id.* at 22.

[100] Primus, *supra* note 97, at 518. Primus provides a more detailed discussion of the different possible motives of Title VII disparate impact. *See id.* at 518–36; *see also* Stephanopoulos, *supra* note 98, at 1604 (discussing the view that the purpose of 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 racial hierarchies of American society."); Samuel R. Bagenstos, *Disparate Impact and the Role of Classification and Motivation in Equal Protection Law After* Inclusive Communities, 101 CORNELL L. REV. 1115, 1132 (2016); Richard Primus*, The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1352 (2010) ("Disparate impact doctrine was widely understood as a means of redressing unjust but persistent racial disadvantage in the workplace . . . .").

an impact helps entrench something like a caste system."[101] On this theory of disparate impact, which I call the "effect-based" theory, intent is irrelevant for more than just evidentiary reasons.[102]

Under both theories, the need to establish causal connections between "policies" and "outcomes" is at the heart of disparate impact. Under either theory, the plaintiff must establish a causal link between the policy and disparate outcome to make a prima facie claim of disparate impact.[103] The stringency of this requirement determines how broad or limited a disparate impact claim can be.

[101] Cass R. Sunstein, *Algorithms, Correcting Biases*, 86 SOC. RSCH. 449, 506 (2019).

[102] 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 100 (arguing that *Griggs* is consistent with both understandings of disparate impact). In the context of fair lending, disparate impact cases have been vague when arguing that loan officer discretion leads to higher rates for minority borrowers. *See* Ian Ayres, Gary Klein & Jeffrey West, *The Rise and (Potential) Fall of Disparate Impact Lending Litigation*, *in* EVIDENCE AND INNOVATION IN HOUSING LAW AND POLICY 231 (Lee Anne Fennell & Benjamin J. Keys eds., 2017). Schwemm and Taren argue that these cases may be considered hybrid impact/intention cases. *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 n.171 (2010). The conduct being scrutinized is discretion provided to brokers (arguably a neutral practice), but discretion may allow brokers to intentionally discriminate against minorities. *See id.* For a discussion of how disparate impact's limited effect in practice is linked to its difficulty to relate to employer "fault," see Michael Selmi, *Indirect Discrimination and the Anti-Discrimination Mandate*, *in* PHILOSOPHICAL FOUNDATIONS OF DISCRIMINATION LAW 257 (Deborah Hellman & Sophia Moreau eds., 2013). *See also* Robert Bartlett, Adair Morse, Nancy Wallace & Richard Stanton, *Algorithmic Discrimination and Input Accountability Under the Civil Rights Acts* 19 (Aug. 1, 2020) (unpublished article) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3674665 [https://perma.cc/4K45-82MB] (discussing the "business justification").

[103] According to the HUD rules implementing the Fair Housing Act's Discriminatory Effects Standard, the "plaintiff has the burden of proving that a challenged practice causes a discriminatory effect." Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11,469 (2013); *see also* Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18,269 (Apr. 15, 1994) ("The existence of a disparate impact may be established through review of how a particular practice, policy or standard operates with respect to those who are affected by it."). 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.

Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 576 U.S. 519, 541 (2015). For a discussion of whether and how *Inclusive Communities* differs from the HUD joint policy, see Schwemm, *supra* note 90. Similarly, for the OCC to find that credit score meets can be justified by a business necessity the variable causing the

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 with the inputs into a decision. Although the prima facie case requires an analysis of the effects or outcomes of a policy, the focus quickly shifts to what inputs created the disparity and whether they relate to a legitimate business justification.[104]

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.[105] 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 its actions.[106] 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.[107]

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

disparity must have "an understandable relationship to an individual applicant's creditworthiness." Westreich & Grimmelmann, *supra* note 57, at 175.

[104] *See Inclusive Communities*, 576 U.S. at 524 (citing Ricci v. DeStefano, 557 U.S. 557, 577 (2009)) ("[A] plaintiff bringing a disparate impact claim challenges practices that have a 'disproportionately adverse effect on minorities' and are otherwise unjustified by a legitimate rationale.").

[105] *See supra* Part I.C (distinguishing disparate treatment, which involves direct conditioning, from disparate impact, in which a facially neutral rule has an impermissible disparate effect).

[106] 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.

[107] *See infra* Part III.

multiple employers and sources, and measure income that is less consistent or formal.[108]

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.[109] 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.[110] 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 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 three 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.[111] I end my discussion with approaches that further exclude inputs that correlate with protected characteristics, in line with the effects-based theory of disparate impact.[112]

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.[113]

[108] 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 85.

[109] *See supra* note 102.

[110] *See supra* note 31 and accompanying text.

[111] *See infra* Part III.A.

[112] *See infra* Parts III.B & III.C.

[113] In his article, *Equal Protection and Disparate Impact: Round Three*, Richard

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.[114] I then present the central methodology of this Article, which is a simulation exercise in which a hypothetical lender uses machine learning to price credit.[115] Building on the simulation exercise, the Part ends by discussing what those changes mean for pricing based on biased inputs and for the application of fair lending law.[116] My conclusion is that algorithmic pricing could, in some cases, exacerbate the problem of biased inputs but, in other cases, mitigate the harm.

A. WHAT IS CHANGING?

Changes in how people receive credit are 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.[117] In this Article, I focus on technology-driven changes in the pricing of credit.[118] I discuss three aspects of artificial intelligence (AI) that are reshaping the personalization of credit pricing, namely the use of non-traditional data, advanced prediction technologies, and automated lending decisions.[119] Many lenders have incorporated a ver-

Primus further discusses how case law and statutory language do not fully support any one theory of disparate impact. *See* Primus, *supra* note 97, 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.").

[114] *See infra* Part II.A

[115] *See infra* Part II.B.

[116] *See infra* Part II.C.

[117] "Fintech" covers a large range of financial activity, including payment and trading systems, and not just the use of technology to automate credit approval and pricing.

[118] 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 a mortgage.

[119] 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

sion 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,[120] credit card lending, and personal loans.[121] In addition, many traditional lenders are using the services of third parties that engage in alternative ways of predicting creditworthiness and pricing credit.[122]

The Fintech market share in borrowing services is significant and increasing. According to one estimate, 82% of lenders report us-

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.

[120] *See* Becky Yerak, *AI Helps Auto-Loan Company Handle Industry's Trickiest Turn*, WALL ST. J. (Jan. 3, 2019), https://www.wsj.com/articles/ai-helps-auto-loan-company-handle-industrys-trickiest-turn-11546516801 [https://perma.cc/37GM-FPM3] (using 2,700 characteristics instead of the few it was using before). Other companies have embraced this type of lending, for example, Synchrony Financial and Ford Motor Credit Co. *Id.*

[121] For example, Upstart 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. *See* UPSTART, https://www.upstart.com [https://perma.cc/TH5E-69Z8]. Another company, Lendbuzz, targets populations that may not have easy access to credit, such as foreign students who are less likely to have US credit histories. *See* LENDBUZZ, https://lendbuzz.com [https://perma.cc/PPV3-RVX7]. The alternative lender, Crest Financial, for example, uses the software of DataRobot for underwriting decisions. *See* Alyssa Schroer, *AI and the Bottom Line: 20 Examples of Artificial Intelligence in Finance*, BUILT IN (July 30, 2021), https://builtin.com/artificial-intelligence/ai-finance-banking-applications-companies [https://perma.cc/2G6W-LRXY].

[122] *See* AnnaMaria Andriotis, *Shopping at Discount Stores Could Help Get You a Loan*, WALL ST. J. (Mar. 4, 2019), https://www.wsj.com/articles/use-a-landline-that-could-help-you-get-a-loan-from-discover-11551695400 [https://perma.cc/W4LU-P4GM]; *see also* UNDERWRITE.AI, https://www.underwrite.ai [https://perma.cc/4Q6S-5SCW]; *KreditTech*, STARTUS, https://www.startus.cc/company/kreditech [https://perma.cc/PP7R-73DZ] ("100% of smartphone or computer owners generate data by anything they do with that device (be it social media, surfing, ecommerce purchases, financial transactions, etc.). Our proprietary algorithm factors in 20,000 data points, which are constantly changing based on newly identified patterns.") (quoting the financial services KreditTech offered before it went out of business); Tom Groenfeldt, *Lenddo Creates Credit Scores Using Social Media*, FORBES (Jan. 29, 2015), http://www.forbes.com/sites/tomgroenfeldt/2015/01/29/lenddo-creates-credit-scores-using-social-media [https://perma.cc/L78C-H2JC] ("Lenddo is finding a lot of interest in its lending application from outside of banking.").

ing nontraditional and alternative data in lending decisions.[123] 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.[124]

1. Nontraditional Data

The first change taking place in the world of credit is the expansion of credit decision "inputs" to nontraditional data. Whereas traditional lending relied on relatively few defined characteristics, lenders are increasingly using new data and additional borrower characteristics to assess creditworthiness. Among them are data on payment and consumer behavior, social media behavior, and digital footprints,[125] as well as information on education, such as the school attended and degree attained,[126] and GPA and SAT scores.[127] Such educational information intuitively relates to a borrower's future income and is particularly valuable for young borrowers who have yet to

[123] *See Alternative Data Across the Loan Life Cycle: How FinTech and Other Lenders Use It and Why*, AITE 11 (2018), https://www.experian.com/assets/consumer-information/reports/Experian_Aite_AltDataReport_Final_120418.pdf [https://perma.cc/PJ2D-CDEC]; *see also Stability Implications from FinTech: Supervisory and Regulatory Issues that Merit Authorities' Attention,* FIN. STABILITY BD. 35 (2017), https://www.fsb.org/wp-content/uploads/R270617.pdf [https://perma.cc/BXT4-LM8U] ("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.").

[124] *See Mortgage Lender Sentiment Survey: How Will Artificial Intelligence Shape Mortgage Lending,* FANNIE MAE (Oct. 4, 2018), https://www.fanniemae.com/sites/g/files/koqyhd191/files/migrated-files/resources/file/research/mlss/pdf/mlss-artificial-intelligence-100418.pdf [https://perma.cc/BLF3-2V33]. 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.

[125] See Request for Information Regarding Use of Alternative Data and Modeling Techniques in the Credit Process, 82 Fed. Reg. 11,183 (Feb. 21, 2017), for a definition of traditional data. See also Hurley & Adebayo, *supra* note 5, at 162–68, 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. *See* Kim, *supra* note 5, at 861 (employment decisions context).

[126] *See supra* note 121.

[127] *See, e.g.*, UPSTART, https://www.upstart.com [https://perma.cc/C64E-7DVY]. This is information that was is typically not considered in credit scoring, such as FICO scores. *See* Hurley & Adebayo, *supra* note 5, at 156.

build up a credit history and who typically have difficulty obtaining certain types of loans.[128]

Credit scores have traditionally only used loan payments to large and established financial institutions to determine creditworthiness.[129] But now, lenders are increasingly using information on timely payment of utility bills and rent payments as indicators of creditworthiness for people without credit history.[130] Similarly, data on phone bills and short-term loans, which were often not included in credit files, are now used by Fintech lenders.[131] Companies with rich information on consumer behavior, such as Alibaba, are using this information to create alternative credit scores.[132]

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.[133] These digital footprints predict default slightly better than traditional credit bureau scores, suggesting that the digital footprints hold information

[128] 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. *See* Hurley & Adebayo, *supra*, note 5 at 166.

[129] *See, e.g.*, *id.* at 156 (describing how FICO principally considers loan payment history, among other factors in determining creditworthiness).

[130] *See supra* note 31. 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.

[131] *See* Chris Brummer & Yesha Yadav, *Fintech and the Innovation Trilemma*, 107 GEO. L.J. 235, 267 (2019) (explaining that fintech lenders collect cellphone records for insight into customers); *Alternative Data Across the Loan Life Cycle: How FinTech and Other Lenders Use It and Why*, *supra* note 123, at 8 (explaining that Fintech lenders use data on short-term installment loans to make lending decisions).

[132] Sesame Credit, developed by Alipay, uses big data to monitor people's buying habits and social circles. *See, e.g.*, John Gapper, *Alibaba's Social Credit Rating Is a Risky Game*, FIN. TIMES (Feb. 20, 2018), https://www.ft.com/content/99165d7a-1646-11e8-9376-4a6390addb44 [https://perma.cc/3SMT-LZS6]. PayU, developed by LazyPay, also develops its own model including customer interaction with apps and spending behaviors to determine creditworthiness. *See* Nikhar Aggarwal, *Here's How PayU Leverages Data Science to Manage Customer's Credit Line*, ETCIO (Dec. 24, 2020), https://cio.economictimes.indiatimes.com/news/next-gen-technologies/heres-how-payu-leverages-data-science-to-manage-customers-credit-line/79933707 [https://perma.cc/876K-EBJD].

[133] *See* Tobias Berg, Valentin Burg, Ana Gombovic & Manju Puri, *On the Rise of FinTechs: Credit Scoring Using Digital Footprints*, 33 REV. FIN. STUD. 2845, 2850 (2019).

that is not contained in credit scores.[134] 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-term loan.[135]

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 creditworthiness, third parties are using this information to provide lenders with alternative or additional data on borrowers.[136] 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.[137] 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,"[138] 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.[139]

[134] *See id.* at 2868. 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.

[135] For example, Afterpay allows shoppers to pay in four installments with zero interest. *See How It Works,* AFTERPAY, https://www.afterpay.com/how-it-works [https://perma.cc/V7HQ-ELGZ]. Klarna offers payment in installments or payment in 30 days with zero interest or 6–36-month financing. *See How It Works,* KLARNA, https://www.klarna.com/us/what-is-klarna [https://perma.cc/FC38-S2R5].

[136] *See* Brummer & Yadav, *supra* note 131, at 265 (describing how data used for loans "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 [https://perma.cc/ZF63-SRA4].

[137] *See* Van Loo, *supra* note 23, at 254 (2018) (concluding that fintechs could expand access to credit through new data sources and other innovations for assessing creditworthiness).

[138] *See* CFPB Off. of Rsch., *supra* note 48, at 6 ("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* Hurley & Adebayo, *supra* note 5, at 155.

[139] *See also* Bruckner, *supra* note 5, at 18.

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 increasingly uses 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.[140] In the case of machine learning, the algorithm itself determines which inputs to use and what weights to assign them in reaching an accurate prediction.[141]

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.[142] 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.[143] 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.[144]

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 has received considerable attention in both academic and policy circles and has been the motivation behind

[140] In the case of FICO scores, the "model assigns a numeric value for each of these five variables, and then applies a pre-determined weight (in percentage terms) to each of these input values and averages them to arrive at a final credit score." *See* Hurley & Adebayo, *supra* note 5, at 162.

[141] For example, Zest AI 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. In that sense, it is using a different prediction technology but more traditional data. *See* ZESTAI, https://zest.ai [https://perma.cc/K3HU-HDNF].

[142] *See* GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE & ROBERT TIBSHIRANI, AN INTRODUCTION TO STATISTICAL LEARNING 239 (2013).

[143] *See id.* at 27.

[144] *See id.* at 238.

legislation that attempts to mitigate the harms that stem from uninterpretable algorithms.[145]

3. Automation

Another important trend in credit lending is the automation of credit pricing—meaning the reduction of human involvement and discretion in setting prices.[146] 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,[147] the ultimate loan terms 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.[148] In addition, it may avoid errors in human judgment with respect to evaluating creditworthiness.[149] Typically, the literature refers to algorithms as "black boxes" and opaque.[150] However, it is harder to imagine a decision-making process that is more of a "black box" than human decision-

[145] See discussions on the right to an explanation in 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. 18, 65–67 (2017); Lilian Edwards & Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16 IEEE SEC. & PRIV. 46 (2018).

[146] *See, e.g.*, Rocket Mortgage from Quicken Loans, which uses a complete end-to-end online mortgage application and approval process. ROCKET MORTGAGE, https://www.rocketmortgage.com [https://perma.cc/RK46-64CM].

[147] 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.

[148] *See* Andreas Fuster, Matthew Plosser, Philipp Schnabl & James Vickery, *The Role of Technology in Mortgage Lending*, 32 REV. FIN. STUD. 1854, 1895 (2019) (describing how technological diffusion will speed up mortgage origination and lead to more efficient refinancing decisions).

[149] A recent paper demonstrates how loan officers that have discretion may make worse decisions when busy, for example. *See* Dennis Campbell, Maria Loumioti & Regina Wittenberg-Moerman, *Making Sense of Soft Information: Interpretation Bias and Loan Quality*, 68 J. ACCT. & ECON. 1 (2019).

[150] *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY (2015).

making.[151] Automation brings an added level of transparency, which provides important regulatory opportunities, as discussed in Part IV.

B. Simulation Exercise – Hypothetical "New World" Credit Lender

To consider the implications of these changes 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.[152]

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,"[153] allows for a meaningful analysis of the legal and methodological challenges in analyzing algorithmic decision rules in a stylized setting.

My hypothetical lender uses loan information reported by mortgage lenders under the Home Mortgage Disclosure Act (HMDA)[154] 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.[155]

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 "random forest," in which the machine learning algorithm makes the prediction using decision trees,[156] or a "lasso regression," another common machine learning algorithm in which the algorithm

[151] *See, e.g.*, Aaron Chou, *What's in the Black Box: Balancing Financial Inclusion and Privacy in Digital Consumer Lending*, 69 DUKE L.J. 1183 (2020).

[152] New lenders or lenders seeking to improve predictions might rely on third parties that collect information on consumer and payment behaviors.

[153] *See generally* Gillis & Spiess, *supra* note 36 (using a simulation exercise based on real-world mortgage data to illustrate the authors' arguments).

[154] 12 U.S.C. § 2803(a)(1).

[155] 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.

[156] *See* Leo Breiman, *Random Forests*, 45 MACHINE LEARNING 5, 6 (2001) (defining "random forests").

selects the variables it deems most important for the prediction.[157] The algorithm is trained on a sample of 2000 clients, with more than 40 variables each (many of which are categorical, taking on a fixed number of possible values).[158] 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.

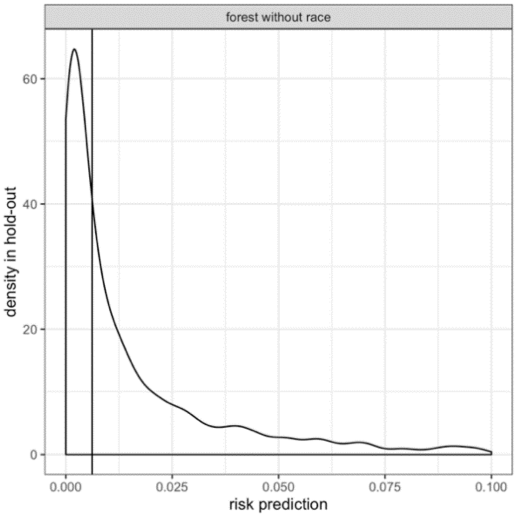

Figure 1: Distribution of predicted risk. The graph shows the distribution risk for all borrowers in the holdout set of 2,000 borrowers. The graph is cutoff at 10%, meaning that only borrowers with a default risk of less than 10% are plotted. The vertical line is the median borrower (of the full sample, 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

[157] 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.

[158] The 40 variables include more types of variables than mortgage originators typically use in setting the "par-rate" in traditional lending, though it does not include many of the nontraditional data discussed above in Part II.A.1 due to data limitations. For a full description of the variables in the Boston Fed HMDA dataset see Munnell et al., *supra* note 77.

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.[159] Borrowers who fall on the right side of the distribution are more likely to default and therefore will pay a higher interest rate.[160]

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.[161] 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 (on the right).

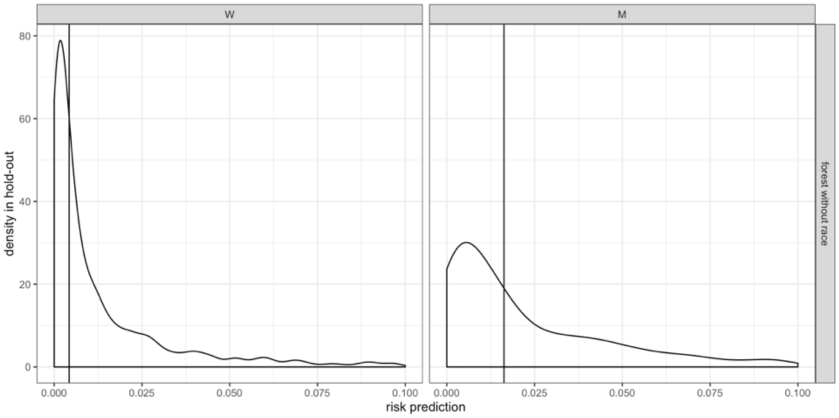

Figure 2: Distribution of default risk for white (W) and minority (M) applicants. Both graphs are cut off at 10% default risk. The vertical line plots the median default risk for the full sample.

[159] *See* Phillips, *supra* note 69 ("[Lenders believe] a riskier customer should pay higher prices in order to compensate for the higher probability of default and the associated cost to the lender").

[160] 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.

[161] 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. Below, in Part III.A, 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 that a higher proportion of white borrowers 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 Part III, I will use this simulation to demonstrate the shortcomings of current approaches to algorithmic discrimination.

C. 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 "biased world" problem because it increases the variance in predictions and may expand the number of "biased world" inputs through its use of nontraditional data.[162] Algorithmic pricing also allows for a greater ability to recover protected characteristics, as will be discussed in detail in Part III.[163] On the other hand, the algorithmic context could mitigate the harms of "biased measurement" by providing an increased amount of information on individuals.[164]

1. Biased World Inputs in Algorithmic Pricing

The first way in which the move to the new world of credit pricing can 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.[165] If algorithmic credit pricing differentiates be-

[162] *See supra* Part I.B.1 (discussing the "biased world" problem).

[163] This has drawn significant scholarly and policy attention. *See, e.g.*, Hurley & Adebayo, *supra* note 5 (discussing algorithmic pricing, its problems, and proposing policy solutions).

[164] *See id.* at 151–52 (explaining that some commentators argue that the increased amount of information about individuals in the complex algorithms companies use actually benefit underserved consumers).

[165] *See* Jennifer Miller, *Is an Algorithm Less Racist Than a Loan Officer?*, N.Y. TIMES (Sept. 18, 2020), https://www.nytimes.com/2020/09/18/business/digital-mortgages.html [https://perma.cc/2TAU-326C] ("[B]roadening the data set could introduce more bias."); Christopher K. Odinet, *The New Data of Student Debt*, 92 S. CAL. L. REV. 1617, 1670 (2019) (describing how new input variables such as educa-

tween people along dimensions that correlate with race, then clearly the outcome disparities will increase.[166]

Another way in which machine learning pricing can increase the disparities of credit prices is through 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, thus creating more granular predictions. Differences among individuals are then 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.[167] Accordingly, 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 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 Ordinary Least Squares (OLS) regression to predict default with a small subset of the variables available in the Boston Fed HMDA dataset.[168] For the machine learning prediction, I

tion-based data may increase disparities in credit lending).

[166] *See* Hurley & Adebayo, *supra* note 5, at 167 (describing Facebook's proposed credit-scoring tool as an example of how algorithmic pricing may perpetuate or even intensify existing biases). Another concern is that as the number of inputs increases, so will the number of inaccurate inputs. *See* Yu et al., *supra* note 84, at 4 ("Expanding the number of data points also introduces the risk that inaccuracies will play a greater role in determining creditworthiness."); *see also* Robert B. Avery, Paul S. Calem & Glenn B. Canner, *Credit Report Accuracy and Access to Credit*, 90 FED. RES. BULL. 297 (2004) (examining the possible effects of data limitations in consumer credit reports, including inaccuracies, on consumers). In general, the accuracy of the data used to price credit (and score consumers) is highly regulated. *See generally* Fair Credit Reporting Act of 1970, 15 U.S.C. § 1681 and Fair and Accurate Credit Transactions Act of 2003, Pub. L. No. 108-159, 117 Stat. 1952 (codified as amended in scattered sections of 15 U.S.C.).

[167] It is not clear that an Ordinary Least Squares (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).

[168] *See* Munnell et al., *supra* note 77, at 28–30 (discussing the variables used in the Boston Fed HMDA dataset). The four variables used for this example are: "housingdti" (housing expenses relative to income), "totaldti" (total debt payment obliga-

use a random forest with the full set of HMDA variables, other than "race." Therefore, the simple function and the machine learning function differ along two dimensions: the number of variables and the prediction technology.

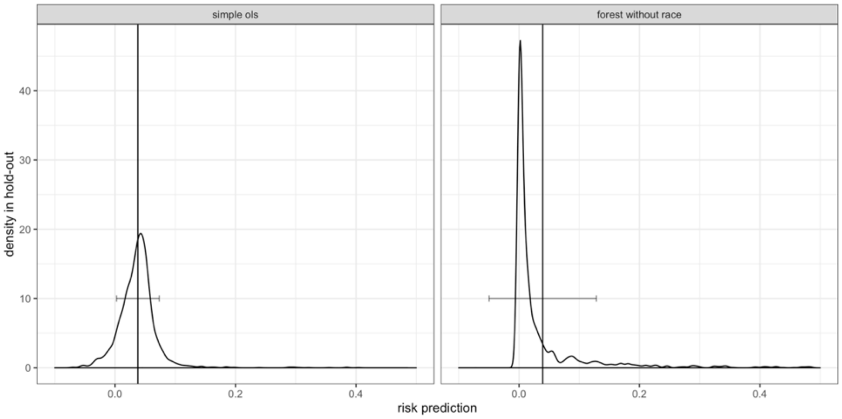

Figure 3: Increase of spread with machine learning. For the graph on the left, an OLS regression was used to predict default with the indepent variables "housingdti," "totaldti," "fixedadjustable," "loanterm." The prediction function was then applied to a "hold out" set. The graphs show the distribution 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 a 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.[169]

tions relative to income), "fixedadjustable" (fixed or adjustable loan term), and "loanterm" (length of loan term).

[169] In reality, these two effects are also likely to be combined because, when confronted with big data, classic regression analysis leads to overfitting—

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 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.[170] As will be discussed in further detail below, the more accurate prediction may allow 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 groups that were previously completely excluded from credit markets, albeit at a higher price than to safer borrowers.[171]

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 follow from the use of machine learning can increase the credit pricing disparities.[172]

At the same time, the use of big data and advance prediction technologies can also lead to decreased reliance on a biased proxy.

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.

[170] Fuster et al., *supra* note 65 (manuscript at 3–6) (explaining that the authors' theoretical work does not employ the exact same variables and machine learning that real-world lenders use, implying that studies like this cannot exactly replicate the effects of increased personalization in lending algorithms in the real world).

[171] *See id.* (manuscript at 6) (explaining the authors' finding that the machine learning model is predicted to provide an increase in number of borrowers access to credit, marginally reducing disparity in acceptance rates across race and ethnic groups, but also predicted increased interest rate disparity across the different groups, with Black and Hispanic borrowers' interest rates increasing).

[172] *See generally* Odinet, *supra* note 165 at 1674–80 (explaining that newly added variables to credit pricing algorithms can increase credit pricing disparities).

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.[173] If big data provides lenders with the opportunity to use rental payment data in addition to FICO scores, this can reduce the differences in predicted default. The use of algorithmic credit pricing could thus decrease 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,[174] a Fintech lender, and FICO scores has decreased over time, due to LendingClub's increased use of nontraditional data.[175] This evidence is consistent with the idea that the use of nontraditional data reduces the impact of the measurement bias of FICO scores.[176]

The Consumer Financial Protection Bureau has also recently discussed the potential benefit of alternative data and machine learning in expanding credit. Based on the finding that an algorithmic lender's model "approves 27% more applicants than the traditional

[173] *See* Hurley & Adebayo, *supra* note 5, at 162 (explaining that the basic FICO score only considers an individual's loan and credit "payment history, outstanding debt, length of credit history, pursuit of new credit, and debt-to-credit ratio in determining credit score").

[174] These ratings are called "rating grades" and are determined by LendingClub. Julapa Jagtiani & Catharine Lemieux, *The Roles of Alternative Data and Machine Learning in Fintech Lending: Evidence from the LendingClub Consumer Platform*, 48 FIN. MGMT. 1009, 1010 (2019).

[175] *See id.* ("Our results demonstrate that the correlation between the borrowers' FICO scores . . . and the rating grades assigned by LendingClub have dramatically declined over the years indicating an increased usage of alternative data in the internal rating process.").

[176] *See id.* (explaining this decrease in correlation between LendingClub rating grades and FICO scores can be explained, in part, by LendingClub's use of nontraditional data, including utility or rent payments, other recurring transactions, electronic records of deposit and withdrawal transactions, insurance claims, credit card transactions, a consumer's occupation or details about their education, their use of mobile phones and related activities, internet footprints, online shopping habits, and investment choices). There is also evidence that cash-flow data can more accurately assess creditworthiness than credit scores, and in some cases act as a supplement for credit scores. *See The Use of Cash-Flow Data in Underwriting Credit: Empirical Research Findings*, *supra* note 26, at 3 ("[C]ash-flow variables and scores tested were predictive of credit risk and loan performance across the heterogenous set of providers, populations, and products studied. Standing alone, the cash-flow metrics generally performed as well as traditional credit scores, which suggests that cash-flow variables and scores can provide meaningful predictive power among populations and products similar to those studied where traditional credit history is not available or reliable.").

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 nontraditional data.[177]

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. The use of advance prediction technologies means that only inputs that contribute to prediction accuracy are considered in pricing. Because algorithms are better able to differentiate among people, biased world inputs might further increase disparities. At the same time, the expansion of input data might undo some of the harm of measurement bias. The fact that the use of algorithmic pricing might 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 I.C, 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.[178] For the effect-based theory of disparate impact, the concern will be whether algorithmic credit pricing exacerbates or entrenches disadvantage.[179] The specific interpretation of the burden-shifting framework may be informed by these theories, such as the stringency ap-

[177] Patrice Alexander Ficklin & Paul Watkins*, An Update on Credit Access and the Bureau's First No-Action Letter*, CFPB (Aug. 6, 2019), https://www.consumerfinance.gov/about-us/blog/update-credit-access-and-no-action-letter [https://perma.cc/TD8G-PXAE].

[178]*. See generally supra* note 98 and accompanying text (describing the intent-based theory of disparate impact).

[179]*. See generally supra* note 102 and accompanying text (describing the effect-based theory of disparate impact).

plied to the initial burden on the plaintiff and how narrowly to construe the "business justification."

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. In focusing chiefly on what goes into the algorithm ("inputs"), these approaches follow the logic of traditional discrimination law. But in doing so, they commit a fallacy for three reasons. First, they often fail on their own terms by not fulfilling their own loose definition of fairness. Second, they sometimes resist practical implementation and are unsuitable for the machine learning setting. Finally, they risk restricting access to credit for vulnerable populations and further entrench disadvantage.

I analyze three approaches, summarized in Table 1.[180] The first approach excludes protected characteristics as inputs, primarily as a method for negating a claim of intentional discrimination under the "disparate treatment" doctrine.[181] The second approach expands the exclusion of inputs to proxies for protected characteristics.[182] This approach recognizes that other inputs may act as "proxies" for protected characteristics and argues that proxies should be excluded too. The last approach I discuss restricts algorithm inputs to only preapproved inputs.[183] It thus differs from the first two approaches, which allow all inputs other than certain forbidden inputs.

The primary fallacy of these approaches is that they continue to scrutinize decision inputs, as traditional fair lending did, even though this strategy is no longer effective in the algorithmic context. At the heart of traditional fair lending lay a paradigm of causality that has become outdated in the algorithmic age: 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.[184] A defendant could then negate a claim of discrimination by showing

[180] 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, 1424 (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. *See id.* at 1394 (arguing that modification should encode or shape training data).

[181] *See infra* Part III.A.

[182] *See infra* Part III.B.

[183] *See infra* Part III.C.

[184] *See* MacCarthy, *supra* note 97, at 81.

that a policy had a causal relationship to a legitimate business interest.[185]

Machine learning, however, is a world of correlation and not causation. When we use a machine learning algorithm to predict an outcome, the focus is on the accuracy of the prediction, and that accuracy 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.

Table 1: Summary of approaches

Approach	What is the approach trying to achieve?	Can the approach be implemented?	Is the approach effective?	Is the approach otherwise undesirable?
Excluding protected characteristic (Section III.A)	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 III.B)	No consideration of race through proxies	Difficulty in defining and identifying proxies	Algorithm can recover protected characteristic better than classic proxies (like zip codes)	—
Restricting inputs to pre-approved characteristics (Section III.C)	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 selectio n of pre-approved variables could entren ch disadvantage. High cost to prediction accuracy.

[185] *See id.* ("However, an employer can defend itself by showing that this policy or practice was based on reasonable factors other than age; they do not have to prove that this reasonable business practice had less impact on older workers than other possible alternatives.").

A. 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 characteristic 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.[186]

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 protected characteristic.[187] Therefore, the requirement that an algorithm exclude a protected characteristic is seen as akin to avoiding the classification on the basis of a protected characteristic.[188]

[186] *See* Kleinberg et al., *supra* note 85, at 27 ("[T]he 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 101 at 507 ("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."). For discussion in the context of employment discrimination, see Sullivan, *supra* note 5, 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." *Id.* at 405. 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. *See id.*

[187] *See generally* Kleinberg et al., *supra* note 85, at 21–22 (describing examples of disparate treatment).

[188] This assumed translation between inclusions of a protected characteristic and "discriminatory intent" is not obvious. See Aziz Z. Huq, *What Is Discriminatory Intent?*, 103 CORNELL L. REV. 1211, 1242–63 (2018) for a discussion of the various interpretations of discriminatory intent in the context of the Equal Protection Doctrine. Discriminatory intent has been interpreted as "motivation" and "animus," which are human attributes and seem less relevant for algorithms. *See id*. at 1222, 1242 (describing motivation and animus as interpretations of discriminatory intent). The basis for attributing discriminatory intent to an algorithm is more appropriate under an "anticlassification" understanding of intent, like that articulated by Huq. *See id.* at 1251–57 (describing the "anticlassification" understanding of discriminatory intent). In the algorithmic setting the mainstream position seems to be that disparate treatment would require the exclusions of protected characteristics. *See generally* Kleinberg et al., *supra* note 85, at 21–22 (describing examples of disparate treatment which require exclusions of protected characteristics). For a related discussion of

What is particularly appealing about the exclusion approach is that in the automated setting, protected characteristics can formally be excluded. In the human decision-making context, by contrast, such formal exclusion is often not possible because the human has observed the protected characteristic, such as race. 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.[189] In the context of algorithmic decision-making, companies can guarantee the formal exclusion of protected 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.

But despite the intuitive appeal of this approach, as I will argue in this Section, 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.

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 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.

The approach of excluding protected characteristics implicitly assumes that an algorithm might want to use a protected characteristic in forming a prediction. It assumes, in other words, that a pro-

whether statistical discrimination violates the Equal Protection Doctrine, see Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 301–22 (2020).

[189] *See* Kleinberg et al., *supra* note 85, 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 from implicit bias and are unaware of how a protected characteristic shapes their decision. *See generally* Samuel R. Bagenstos, *Implicit Bias's Failure*, 39 BERKELEY J. EMP. & LAB. L. 37 (2018) (describing how implicit bias is unconscious bias individuals are unaware of).

tected characteristic could be empirically relevant, in that it could provide information on default probability. In this respect, the use of an algorithm alleviates at least the concern over the arbitrary use of protected characteristics, because an algorithm would not consider a protected characteristic unless it had informational value.

However, the empirical relevance of a protected characteristic for an accurate prediction is also precisely what gives rise to the concern that an algorithm can still discover a protected characteristic, even after its exclusion. This concern gets at the difficulty of using a statistical technology that is focused on empirical accuracy, while complying with legal restrictions that go beyond empirical relevance. For example, the ECOA prohibits pricing on the basis of gender regardless of whether gender is of empirical relevance to default prediction.[190]

One reason an algorithm would consider a protected characteristic is that the characteristic correlates with some other unobservable characteristic that is of true interest.[191] In such a case, the protected characteristic is not of interest in and of itself. 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.[192] Economists often describe this situation as "statistical discrimination" because race is used to infer other information.[193]

[190] ECOA, 15 U.S.C. § 1691(a)(1).

[191] It is not possible to perfectly establish whether a characteristic is what I call "causal" or not of an outcome. The point I wish to make is that protected characteristics may be predictive because of the underlying relationship to the target and not because they act as proxies.

[192] This example closely relates to the category of proxy discrimination that Prince and Schwarcz call "Indirect Proxy Discrimination." *See* Anya E. R. Prince & Daniel Schwarcz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data,* 105 IOWA L. REV. 1257, 1279–81 (2020) ("[P]roxy discrimination will tend to occur when a suspect variable is predictive of a desired outcome only because it proxies for another, quantifiable and potentially available, variable that causes the desired outcome but that is not included in the AI's training data.").

[193] Statistical discrimination is the use of protected characteristics to form accurate beliefs about unobservable characteristics. *See generally* Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972) (describing statistical discrimination); Kenneth Arrow, *The Theory of Discrimination* (Princeton Univ., Indus. Rels. Section, Working Paper No. 30A, 1973) (describing statistical discrimination in the labor context). There is more nuance to the types of correlations that an algorithm might want to discover than is presented here. *See generally* Prince & Schwarcz, *supra* note 192 (providing other examples of types of correlations an

Legally, "statistical discrimination" is likely to be prohibited by fair lending's disparate treatment doctrine.[194] The direct conditioning on a protected characteristic, even if it merely serves as a proxy for another characteristic, nonetheless leads to different pricing for protected groups. And the fact that there is empirical support for using the protected characteristic, in that its use increases prediction accuracy, would not serve as a defense.[195] Protected characteristics are also sometimes of direct interest. When a protected characteristic is causally or closely related to the outcome of interest, an algorithm has a direct interest in recovering the characteristic.[196] 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.[197] In the case of ECOA, the Act prohibits discrimination on grounds that are possibly causal of default.[198] ECOA prohibits discrimination based on age and based on whether a borrower receives his or her income from public assistance programs, as discussed above.[199] It is plausible that these two factors affect a borrower's predicted future income and therefore

algorithm may want to discover); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 820–28 (2020) (discussing other examples of correlations algorithms attempt to discover including sickness and recidivism).

[194] *See generally* Kleinberg et al., *supra* note 85 (describing disparate treatment doctrine). This may not be 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 188, at 1249.

[195] *See supra* Part I.C.

[196] This is closely related to what is often referred to as "rational discrimination," which often comes up in the context of disability insurance. *See generally* Samuel R. Bagenstos, "*Rational Discrimination," Accomodation, and the Politics of (Disability) Civil Rights*, 89 VA. L. REV. 825 (2003).

[197] *See* Prince & Schwarcz, *supra* note 192, at 1281 (discussing the example of health 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.

[198] 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 42. For an economics perspective on this type of discrimination, see J. Aislinn Bohren, Kareem Haggag, Alex Imas & Devin G. Pope, *Inaccurate Statistical Discrimination: An Identification Problem* 10 (Nat'l Bureau of Econ. Rsch., Working Paper No. 25,935, 2020), proposing a new category of discrimination: "inaccurate statistical discrimination," which is a type of statistical discrimination that is based on inaccurate beliefs.

[199] *See supra* Part I.C.

closely relate to default risk. Yet, ECOA prohibits their consideration. Similarly, ECOA prohibits discrimination based on marital status, although this too could affect the likelihood of default when a mortgage is underwater.[200]

Other protected characteristics, such as gender and race, may also have a direct empirical relation to default risk, causing further schism between empirical accuracy and normative limitations. For example, it might be rational for a lender to consider race and gender when estimating future income assuming there is labor market discrimination against women and racial minorities.[201] If information about race and gender is available to an algorithm, an optimized prediction of default is likely to consider these protected characteristics. This would be true whether or not the protected characteristic is provided directly as an input.

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

[200] 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, Kyle D. Logue & Daniel Schwarcz, *Understanding Insurance Antidiscrimination Laws*, 87 S. CAL. L. REV. 195 (2014). 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, 2004 O.J. (L 373) (EC) (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.

[201] *See generally* David Neumark, *Experimental Research on Labor Market Discrimination*, 56 J. ECON. LITERATURE 799 (2018).

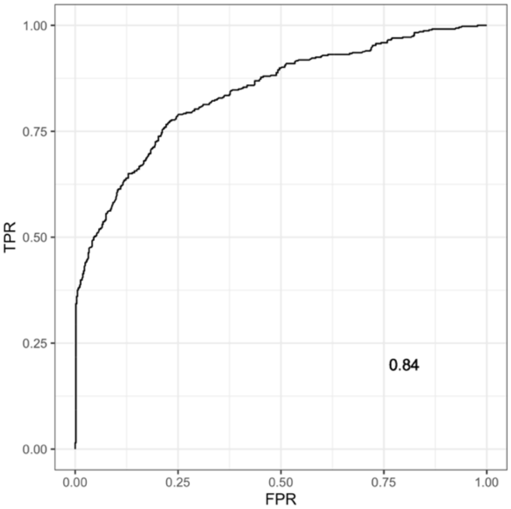

Figure 4: ROC curve for prediction of borrower "age." The "age" variable in the Boston Fed HMDA dataset is not a continuous variable of age but rather an indicator of whether the applicant's age is above or below the median in the Boston Metropolitan Statistical Area. The ROC curve plots the true-positive-rate 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 12 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 characteristic (ROC) curve. The number on the bottom 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.[202]

[202] In fact, these results are a lower bound of what is feasible with big data and machine learning. As discussed, the variables in the Boston Fed HMDA dataset are primarily more traditional pricing variables and the data are therefore not as rich as those likely available to algorithmic lenders. *See* Munnell et al., *supra* note 77. With nontraditional data, lenders might recover protected characteristics with even greater accuracy.

My 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. Even if an algorithm does not seek to recover the information—that is, even if it never tries to derive race or marital status—such characteristics are available to it because they are so embedded in 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 characteristics 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 of the underlying characteristics 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 a 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.

The gap between what is empirically relevant and what is normatively relevant suggests a blurring of the distinction between antidiscrimination law and affirmative action.[203] If two people with different default risks, because they are of a different age or gender, are forced to be treated equally, there is potentially a cross-subsidization from one group (mid-aged or male borrowers) to another group (older or female lenders).[204]

[203] *See* Strauss, *supra* note 28. In this context, I refer to affirmative action as forcing equal treatment of borrowers with different risk profiles. However, affirmative action is much broader than this example.

[204] This blurring of the lines is similar to arguments advanced with respect to the similar functioning of discrimination law and "reasonable accommodations" in the context of the Americans with Disabilities Act. *See* Christine Jolls, *Antidiscrimination and Accommodation*, 115 HARV. L. REV. 642, 697 (2001). Here, however, I argue that this form of affirmative action is a result of disparate treatment and the barring of direct conditioning on a protected characteristic and not necessarily disparate impacts.

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.[205]

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,[206] 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.[207]

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

[205] The extent to which disparate impacts are concerned in directly reducing outcome disparities is discussed above in Part I.B, and may depend on what type of reason is driving the disparities created by exclusion.

[206] In the 2000 HMDA dataset, for example, Black applicants are less than ten percent of all applications reported. *See Nationwide Summary Statistics for 2000 HMDA Data, Fact Sheet*, FED. FIN. INST. EXAMINATION COUNCIL (FFIEC) tbl.2 (2001), https://www.ffiec.gov/hmcrpr/hm00table2.pdf [https://perma.cc/HF4U-7HPH]. It's important to note that the Boston HMDA is skewed to overrepresent minorities relative to their share amongst mortgage applicants.

[207] *See also* Kim, *supra* note 5, 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."); Melissa Hamilton*, The Biased Algorithm: Evidence of Disparate Impact on Hispanics*, 56 AM. CRIM. L. REV. 1553 (2019) (considering the performance of the COMPAS risk assessment on Hispanics); Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAV. SCI. & L. 145 (2019) (discussing with respect to COMPAS and gender); Alice Xiang, *Reconciling Legal and Technical Approaches to Algorithmic Bias*, 88 TENN. L. REV. 3, 24 (2021) (discussing COMPAS and the case *State v. Loomis* where the court found it compelling to include gender to promote accuracy). *See generally* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Ashesh Rambachan, *Algorithmic Fairness*, 108 AEA PAPERS AND PROC. 22 (2018).

equally distributed across the population and higher "ability" people default less, perhaps because their earnings are higher.[208] 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.

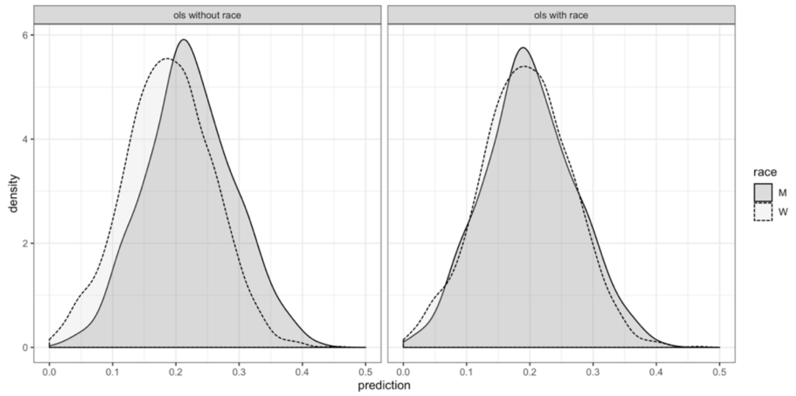

Figure 5: Simulated example of default risk using a "race blind" algorithm (on the left) and a "race aware" algorithm (on the right). The graphs plot the distribution risk for white (W) and minority (M) borrowers using an OLS regression.

Figure 5 shows that in my simulated example, predicting default risk only from college attendance results in non-white borrowers having a higher default probability.[209] 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.

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

[209] 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.

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.[210] 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 reliable means of reducing 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.[211] Recently, however, several scholars have suggested that discrimination law's position on the consideration of a protected characteristic may be more nuanced.[212]

B. Excluding Proxies for Protected Characteristics

A second approach to applying discrimination law to algorithmic pricing expands the prohibited inputs to also include "proxies" for protected characteristics. The discussion in the previous Section demonstrates that the exclusion of a protected characteristic may be meaningless if an algorithm can use proxies for that characteristic.[213] If there are proxies for a 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

[210] In Part II.B above, I present a case in which the lasso regression puts weight on the input "race," predicting that white borrowers are less likely to default.

[211] *See* MacCarthy, *supra* note 97, at 73 ("These cases do suggest that the use of group variables in algorithms would be subject to strict scrutiny, even if their purpose is to reduce group disparities.").

[212] Deborah Hellman, for instance, argues that separately considering which inputs are predictive of future criminal activity may not in fact constitute disparate treatment. *See* Hellman, *supra* note 193, at 854.

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 and the requirement to assure that policies do not create a disparate impact, thereby requiring a consideration of people's forbidden characteristics, has recently been debated. *See* Ricci v. DeStefano, 557 U.S. 557, 579 (2009) (indicating that a promotion test was invalidated by an employer because of the concern that promotion based on the test would trigger disparate impact); *see also* Hellman, *supra* note 193, at 822; Kim, *supra* note 5, at 925; Jason R. Bent, *Is Algorithmic Affirmative Action Legal*, 108 GEO. L.J. 803, 809 (2020) ("Voluntary algorithmic affirmative action ought to survive a disparate treatment challenge under *Ricci* and under the anti-race-norming provision of Title VII."). *See generally* Primus, *supra* note 97.

[213] *See supra* Part III.A.

fair lending, this strategy is sometimes adopted by excluding salient examples of proxies, such as zip codes.[214]

Several scholars have proposed preventing algorithms from using variables that are highly correlated with a protected characteristic.[215] For example, Hurley and Adebayo propose a model bill—the Fairness and Transparency in Credit Scoring Act—that contains this type of provision.[216] The model bill requires that credit scores "not treat as significant any data points or combinations of data points that are highly correlated to immutable characteristics."[217] This approach was also articulated by HUD in its proposed rule on disparate impact, in a section related to algorithmic credit decisions. According to the proposed rule, a defendant can negate a claim's disparate impact by showing that a risk assessment algorithm excludes proxies for protected characteristics.[218]

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,[219]

[214] 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*, 88 U. COLO. L. REV. 1109, 1136 (2017) (discussing redlining).

[215] *See* Hurley & Adebayo, *supra* note 5.

[216] *See id.* at 196.

[217] *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. *See id.* at 190.

[218] Section 100.500 (c)(2) of HUD's proposed disparate impact rule relates to a case in which a plaintiff is challenging a defendant's use of a model with a discriminatory effect and lays out the defenses on which a defendant can rely. *See* HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42,854 (proposed Aug. 19, 2019) (to be codified at 24 C.F.R. pt. 100), *supra* note 11. According to § 100.500(c)(2)(iii), 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." *See id.* Ultimately, the Final Rule did not include the proposed provisions on algorithmic decisions, however HUD has not yet provided any alternative guidance on the topic.

[219] *See* Charles River Assocs., *Evaluating the Fair Lending Risk of Credit Scoring Models*, CRA INSIGHTS: FIN. ECON. 1, 3 (2014), https://media.crai.com/sites/default/files/publications/FE-Insights-Fair-lending-risk-credit-scoring-models-0214.pdf [https://perma.cc/6S56-LHAM] ("Ostensibly neutral variables that predict credit risk may nevertheless present disparate impact risk on a prohibited basis if they are so

and/or (2) does not contain informational value beyond its use as a proxy.[220] Hurley and Adebayo focus on variables that highly correlate with protected characteristics.[221] Other approaches require something beyond a correlation, such as requiring that the variable does 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[222] is difficult to implement in practice. The problem is that we do not have a good understanding of the "model" of default, nor of the variables that 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.[223] However, even zip codes, which have become the archetype of a proxy for race, are likely to contain informational value relevant to default risk.[224]

highly correlated with a legally protected demographic characteristic that they effectively act as a substitute for that characteristic.").

[220] *See Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003*, *supra* note 7. There are other approaches for the exclusion criteria. For example, Sunstein has argued that "[d]ifficult problems are presented if an algorithm uses a factor that is in some sense an outgrowth of discrimination." *See* Sunstein, *supra* note 101, at 509. In the context of credit pricing this would mean excluding many of the fundamental features used to price credit, even today, such as credit scores and wealth.

[221] *See* Hurley & Adebayo, *supra* note 5, 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.").

[222] *See* Prince & Schwarcz, *supra* note 192, at 1257 ("A practice producing a disparate impact only amounts to proxy discrimination when 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.").

[223] *See* Yu et al., *supra* note 84, at 28 (providing 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.* at 29.

[224] For example, the real estate fluctuations in a particular area. *See* Erik Hurst, Benjamin J. Keys, Amit Seru & Joseph Vavr, *Regional Redistribution Through the US Mortgage Market*, 106 AM. ECON. REV. 2982, 2982 (2016) (documenting large regional variation in default risk, despite the uniform pricing of Government Sponsored Enterprises (GSEs) across regions).

A further difficulty is that many variables can be an indicator of a protected characteristic and also independently contain information relevant to the outcome of interest. In most cases, we are not able to isolate the component of a variable that is merely a proxy for a protected characteristic and the component that contains independent information.

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 characteristics 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.[225] 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 both "race" and the ten variables that correlate most with "race." One way to consider the disparities between the groups is by considering 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,[226] 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.[227]

[225] This figure is similar to the figure produced in Gillis & Spiess, *supra* note 36, 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.

[226] This is partially because the distributions have altogether been condensed as a result of the use of fewer variables to distinguish between borrowers. *See supra* Part II.C.1.

[227] 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.

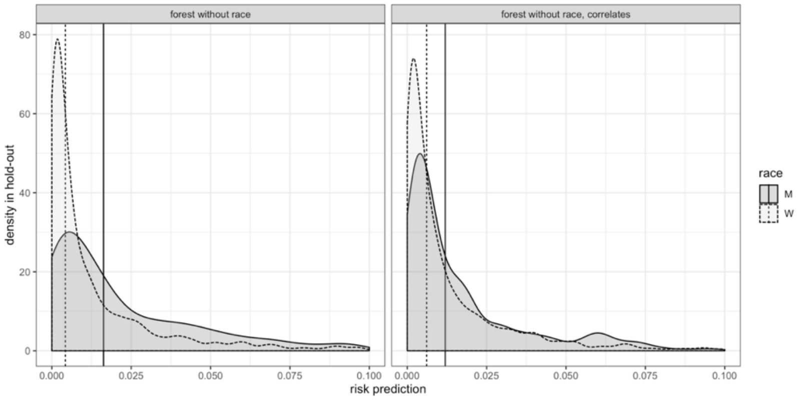

Figure 6: Distribution of risk predictions across groups for different inputs. The graph on the left shows the risk predictions when using all HMDA inputs other than race, plotted separately for the non-Hispanic white (W) and non-white (M) borrowers in the holdout group. The graph on the right shows the risk predictions when using HMDA inputs other than race and ten variables with the highest correlation to race. It, too, plots the predictions 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,[228] for the Boston Metropolitan Statistical Area.[229]

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. *See supra* Part II.B.

[228] 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.

[229] 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 Black populations 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 77, at 26.

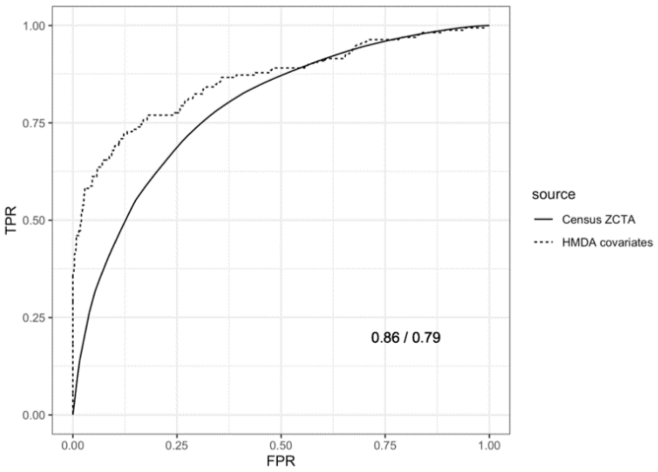

Figure 7: ROC curve for prediction of "Black" using HMDA covariates and using Census ZCTA

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 also be 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 features that do not feel like proxies.

The final reason to be wary of this second exclusion approach is that most inputs used to price credit, even in the traditional context, correlate with a protected characteristic.[230] Restricting the use of variables that correlate with protected characteristics reduces lenders' ability to accurately predict default risk and personalize pricing accordingly.[231]

[230] *See supra* Part I.B.

[231] *See supra* Part I.A.

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.

C. Restricting the Algorithm to a Predetermined Set of Variables

A third approach restricts the inputs of an algorithm to inputs that are pre-approved. It was recently proposed by Prince and Schwarcz in the context of insurance: "[i]nstead of allowing use of any variable not barred, as in the traditional anti-discrimination model, this approach would only allow actors to use pre-approved variables."[232] This third 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.

A related recent proposal looks to restrict algorithmic inputs to a set of pre-vetted variables. According to Bartlett et al., algorithmic inputs should only include variables that do not penalize protected groups disproportionately, controlling for the variables' predictive relevance.[233] This test, which they call the "input accountability test," breaks down variables into a component that predicts the target, such as creditworthiness, and a component that does not relate to the target ("noise"). If the noise component of the variable is correlated with a protected characteristic, it must be excluded from an algorithm's inputs.[234]

Predetermining permissible variables could be implemented either by a regulator or by 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.[235] If lenders must show that a lending decision relies on inputs that are logically related to the outcome, they will need to ex-

[232] *See* Prince & Schwarcz, *supra* note 192, at 1306.

[233] *See* Bartlett et al., *supra* note 102, at 23.

[234] *Id.* at 31.

[235] Prince & Schwarcz, *supra* note 192, at 1316 ("[O]ne possible solution is to require those employing algorithms to convince regulators or others of causal connections between the variables utilized and the desired outcome.").

clude other variables. Predetermining which variables are related to the outcome will allow lenders to meet this burden.[236]

1. Entrenching Disadvantage

The main challenge for the third approach is to define which variables are permissible. That definition depends on what the restriction is meant to achieve. One version of the approach, for instance, might want to restrict the variables to only inputs that predict default. But if that is the goal, then there would be no reason to restrict inputs at all. After all, compared to a human, the algorithm is a better judge of whether an input predicts default.

Another version of the approach might limit the algorithm to characteristics that are used in traditional credit pricing, such as FICO scores or a borrower's income.[237] But this would undermine 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."[238] 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 version allows 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 ver-

[236] *See* Westreich & Grimmelmann, *supra* note 57, at 15 ("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 5, 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.").

Another approach which seeks to develop a pre-approval process for inputs was recently suggested by Yang and Dobbie. Their approach is based on a statistical method to prevent inputs that correlate with protected characteristics from serving as proxies. *See* Yang & Dobbie, *supra* note 188.

[237] *See supra*, Part I.A.

[238] *See* Ficklin & Watkins, *supra* note 177. Fintech can potentially be used "to address race-based financial inequality while also being attentive to the possibility that seemingly innocuous technologies can generate biased banking practices against minority communities." Julia F. Hollreiser, Note, *Closing the Racial Gap in Financial Services: Balancing Algorithmic Opportunity with Legal Limitations*, 105 CORNELL L. REV. 1233, 1235 (2020).

sion 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)."[239] 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.[240] Similar to the approach in Subsection III.B.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 any direct sense of how inputs relate to the structural model.[241] 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—among them that the public repayment of loans may lead to lower default

[239] *See* Bartlett et al., *supra* note 68, at 3.

[240] This articulation of the exclusion criteria is different to the one provided in Bartlett et al., *supra* note 102. According to this later article, an input would be required to be excluded even if it was included in the life-cycle model as long as the residual of the prediction of default correlated with race. *Id.* at 31, 34.

[241] 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 68, at 15.

rates due to reputation concerns.[242] 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.[243]

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. Without such limitations, machine learning pricing can increase accuracy for a few reasons. First, its replacement of human prediction with an automated system of prediction can add accuracy to the prediction.[244] Second, compared to other statistical methods, such as linear regressions, machine learning allows for greater flexibility in forming a prediction, which in turn increases accuracy.[245] Finally, the expansion of the type and number of inputs considered by an algorithm can further increase its accuracy.[246]

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.[247]

[242] *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) (discussing the various theories and empirical evidence on microlending).

[243] For example, suppose creditworthiness is affected by social attitudes to foreclosure, which are more prevalent among minority communities, but that these social norms were not part of the structural model of repayment. In such a case, the exclusion of this type of input may in fact increase bias.

[244] *See* Lkhagvadorj Munkhdalai, Tsendsuren Munkhdalai, Oyun-Erdene Namsrai, Jong Yun Lee & Keun Ho Ryu, *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 that the non-human expert prediction is superior in predicting default). For a similar discussion in the context of bail decisions, see Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig & Sendhil Mullainathan, *Human Decisions and Machine Predictions*, 133 Q. J. ECON. 237 (2018).

[245] *See supra* Part II.C.

[246] Some input proposals go beyond restricting non-traditional inputs and may also require restricting traditional credit inputs. It is unlikely, for example, that even the traditional credit inputs would pass Bartlett et al.'s "input accountability test," as it imposes very strict conditions on each individual input. This would mean that under this test, credit may not be personalized at all. *See* Bartlett et al., *supra* note 102, at 32.

[247] 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

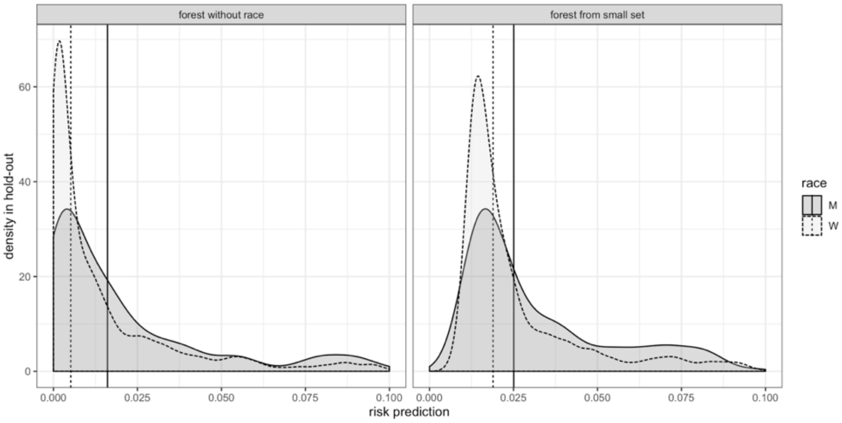

Figure 8: Distribution of risk predictions. The graph on the left shows the risk predictions using a random forest with the full set of inputs (other than race). The graph on the right shows the risk predictions using a random forest with a small set of more traditional credit inputs. Both graphs separate 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 fewer 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.[248]

loan.

[248] *See* Appendix B (ROC curves).

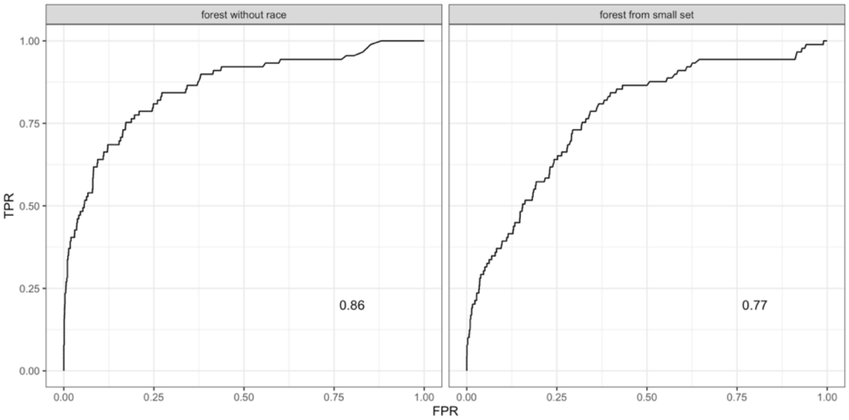

Figure 9: ROC curves corresponding to risk distributions in Figure 8. The ROC curve on the left shows the accuracy of the risk predictions using a random forest with the full set of inputs (other than race). The graph on the right shows the accuracy of the risk predictions using a random forest with 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.[249]

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.[250] 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

[249] Bartlett et al.'s "input accountability test" is likely to restrict inputs much further, possibly to the extent of prohibiting personalized credit pricing altogether. It is questionable whether even traditional inputs would pass this test. Bartlett et al., *supra* note 102, at 31.

[250] *See* Corbett-Davies et al., *supra* note 68, at 802–03; *see also* Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg & Kilian Q. Weinberger, *On Fairness and Calibration*, *in* PROCEEDINGS OF THE 31ST INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 2 (Ass'n for Computing Machinery ed., 2017) https://dl.acm.org/doi/10.5555/3295222.3295319 [https://perma.cc/R2FV-99G4]; Prince & Schwarcz, *supra* note 192, at 1306 ("[If no solution is adopted] due to narrowly-defined notions of efficiency, then it must be acknowledged that this comes at the expense of these [anti-discrimination] laws' goals.").

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 I.A, reduced accuracy could hurt vulnerable borrowers who are excluded altogether from credit markets when the lender cannot accurately price risk.[251]

D. Required Shift from Causation to Correlation

Translating traditional discrimination law to the algorithmic context requires more than the small tweaks suggested by approaches that continue to focus on credit pricing inputs and on their causal relationship to the differential treatment of protected groups. Instead of falling prey to their input fallacy, we must recognize that in the machine learning context, we cannot identify causal relationships.[252]

As discussed in Section I.C, fair lending law has traditionally focused on causal questions. Many scholars continue to apply this causal framework to discrimination law in the algorithmic context,[253] and regulators, too, continue to contemplate causal relationships.[254]

[251] *See supra* Part I.A; Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L.J. 1043 (2019). 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 requires considering who bears the cost of algorithmic errors in determining how to apply notions of fairness. *Id.* at 1111–12.

[252] *See* Martin J. Katz, *The Fundamental Incoherence of Title VII: Making Sense of Causation in Disparate Treatment Law*, 94 Geo. L.J. 489 (2006) (discussing the causation requirement in anti-discrimination laws); Foster, *supra* note 49, at 1472 ("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).").

[253] 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. For example, the NCLC describes this intuitive type of argumentation: "[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." Yu et al., *supra* note 84 , at 29. Similar arguments have been made in the context of employment. *See* Kim, *supra* note 5, at 881 ("If, however, the variables are merely correlated and not causally related, there is no necessary connection between them, and the correlation may not hold in the future."); *see also* King & Mrkonich, *supra* note 57, at 555 (arguing law's causal inquiry should be distinguished from a social science understanding of causality); Bartlett et al., *supra* note 102, at 37, 39.

[254] 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

In particular, scholars continue to maintain that disparate impact's "business necessity" is not met when there is a mere correlation between the features and outcome variables,[255] even though correlation is in fact the only relationship identified by machine learning. Similarly, some have argued with respect to substantiating a disparate impact claim that "there must be a nexus or causal connection between some element of institutional practices and the disparate outcome."[256]

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 their 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.[257]

Because of the absence of identifiable causal relationships, input-based approaches are unsuitable for discrimination law in an algorithmic setting. This, we have seen, 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.

IV. THE FUTURE OF FAIR LENDING

Given the unsuitability of input-based approaches, we must rethink how to analyze discrimination in the new context of algorith-

impact." HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, 84 Fed. Reg. 42,854, 42,859 (proposed Aug. 19, 2019) (to be codified at 24 C.F.R. pt. 100). HUD's articulation of the defense relies on the ability to isolate inputs and separately evaluate their causal relationship to a disparate outcome. *Id.*

[255] *See* Westreich & Grimmelmann, *supra* note 57 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.").

[256] *See* MacCarthy, *supra* note 97, at 84.

[257] Another example is using data on the time it takes a person to fill out an online application as predictive of default risk. *See* Berg et al., *supra* note 133, at 2894. 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.

mic credit pricing. If we are to fight credit discrimination as fair lending requires us to—by preventing the substantively different treatment of protected groups and advancing distributional and fairness interests—we must shift our focus to outcome 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.[258] This Part proposes a new framework for conducting such outcome-based analysis.

A. OUTCOME TESTING

The framework that I propose is an outcome-based test that regulators should use to assess whether credit pricing discriminates against protected groups in violation of fair lending law. The test applies a lender's pricing rule to a dataset of hypothetical borrowers and then examines the properties of the outcome. The test 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.[259] At the second stage, the regulator then takes

[258] 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 algorithmic fairness requires an alternative framework to discrimination, such as affirmative action. *See* Dwork et al., *supra* note 58 (proposing "fair affirmative action"); *see also* Chander, *supra* note 50, at 1040 (proposing we deal with unfair outcomes as a result of biased inputs through affirmative action). Huq's recent proposal to evaluate algorithmic criminal justice measures 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. Huq, *supra* note 251, at 1128. 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 INFO., COMMC'N & SOC'Y 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.

Others have suggested that the concerns of algorithmic fairness be addressed through creating appropriate frameworks that allow further private or public scrutiny. *See*, *e.g.*, PASQUALE, *supra* note 150 (discussing greater transparency as one possible approach). 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*, 66 STAN. L. REV. ONLINE 35 (2013).

[259] 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. 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 is primarily focused on

that prediction or pricing rule and applies it to a dataset of people to see the distribution of prices the rule produces.[260] 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.[261] Finally, the regulator evaluates the outcome to determine whether the disparities created by the pricing rule amount to discriminatory conduct.[262] I use the example of race as a protected characteristic, but the analysis is generalizable to other protected characteristics.

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.[263] For the reasons discussed in Part III, the criteria used to de-

understanding whether this lender decision varies by race. *See* CFPB Home Mortgage Disclosure Regulation C, 12 C.F.R. § 1003.1(b)(iii).

[260] Elsewhere I have argued that it is difficult to analyze a prediction function in the abstract. Gillis & Spiess, *supra* note 36, at 473–74. Rather, the prediction function should be applied to a group of borrowers in order to examine its properties. *Id.* at 485. 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. *Id.* at 486. 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.

[261] I focus on the possibility of the regulator applying the pricing rule to a dataset, but it is also possible to require lenders themselves to create such a test internally, which is then reported to the regulator.

[262] 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, focusing 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* Corbett-Davies & Goel, *supra* note 58, at 6 (defining this category as any measure that can be calculated from a confusion matrix, which tabulates the joint distributions of a certain decision and outcomes by a group); *see also* Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, 50 SOCIO. METHODS & RSCH. 3 (2018). Two of these measures, the "true positive rate" (TPR) and "false positive rate" (FPR), discussed in Appendix B, provide a way to measure the prediction accuracy. Ancillary literature has focused on documenting how the various classification errors often cannot simultaneously be satisfied. *See* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, ARXIV, Sept. 19. 2016, at 2, https://arxiv.org/pdf/1609.05807 [https://perma.cc/X6DK-T2MM] ; Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5 BIG DATA 153 (2017).

Some recent legal literature has also focused on these types of outcomes. *See* Hellman, *supra* note 193; MacCarthy, *supra* note 97, at 91–94, 96.

[263] 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. *See, e.g.*, Tex.

termine 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 clear definition of what discrimination law, and disparate impact in particular, is meant to achieve.

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 Article. 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, which would be needed to analyze discrimination under "discrimination as anti-classification." Much of the definition of credit discrimination centers around the interpretation of discriminatory intent.[264] One possible interpretation of discriminatory intent may be animus towards a protected group or "taste-based discrimination," meaning discrimination that is based on a prejudicial preference for one group over the other.[265] While this interpretation tracks some understandings of discriminatory intent under the Equal Protection clause[266] it is unlikely to be a correct interpretation of current fair lending law.[267] A more likely interpre-

Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 576 U.S. 519, 520 (2015). A typical disparate impact claim begins with a demonstration of outcome disparities. This showing of disparities is rarely sufficient in and of itself, even for the first stage of a case, since a plaintiff is also required to isolate the particular policy or input that led to the disparity. *Id.* at 521 ("A disparate-impact claim relying on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity."). Despite the role outcome analysis plays in a disparate impact case, there is little guidance on how exactly to conduct this analysis.

[264] *See supra* Part I.C.

[265] This is how economists typically refer to discriminatory intent as animus. Huq, *supra* note 188, at 1242.

[266] *See, e.g.*, William D. Araiza, *Animus and Its Discontents*, 71 FLA. L. REV. 155, 158–59 (2019) (finding application of the animus concept to the Equal Protection Clause and religion clauses, "suggests the flexibility and portability of the animus concept").

[267] *See, e.g.*, *Consumer Financial Protection Bureau Issues Two Final Rules to Promote Access to Responsible, Affordable Mortgage Credit*, CFPB (Dec. 10, 2020), https://www.consumerfinance.gov/about-us/newsroom/consumer-financial-protection-bureau-issues-two-final-rules-promote-access-responsible-affordable-mortgage-credit [https://perma.cc/47VM-9NZW].

If credit discrimination is limited to cases of animus, by definition a human sentiment, the use of algorithms for lending decisions is unlikely to raise a concern and will not require scrutinizing algorithmic inputs. Huq, *supra* note 251, at 1088–90. As long as the algorithm was set up to predict the correct object, such a credit risk, any tation of discriminatory intent is "anti-classification," focusing on whether a protected characteristic was a criterion for a lending decision.[268] Considering whether "similarly situated" borrowers were treated differently provides the second-best way of analyzing whether a protected characteristic was used as a criterion when this question cannot otherwise be answered directly.[269]

The second question that can be analyzed using outcome analysis is whether the pricing rule increases or decreases disparities relative to some baseline, which may be a way to analyze "discrimination as discriminatory effect."[270] The type of incremental analysis I propose is appropriate under the position that disparate impact plays a role beyond identifying discriminatory intent and is meant to address policies that have a discriminatory effect even when lacking intent.[271] This approach recognizes the role that disparate impact plays in balancing the benefits of accurate predictions and the business interest of lenders, with the need to prevent further disparities in credit markets, highlighted by inputs reflecting a biased world and biased measurement. The traditional way in which disparate impact played that role is the burden shifting framework.[272] However, fair lending's

way in which the data is used cannot be motivated by animus but by an attempt to provide an accurate prediction. *See id.* at 1086–87. Thus, even the direct use of a protected characteristic in pricing would not reflect animus and therefore would not be discriminatory. It is important to keep in mind that if the humans who design the algorithm are motivated by animus, this would trigger discrimination laws. *See id.* at 1089 (explaining intent could be found when "an algorithm's designer [is] motivated by either an animosity toward a racial group, or else a prior belief that race correlates with criminality, and then deliberately design the algorithm on that basis").

[268] *See* Huq, *supra* note 188, at 1251; *see also* Reva B. Siegel, *Equality Talk: Antisubordination and Anticlassification Values in Constitutional Struggles Over* Brown, 117 HARV. L. REV. 1470 (2004) (discussing racial anti-classification in the context of *Brown v. Board of Education*).

[269] *See supra* Part III; *see also* Xiang, *supra* note 207, at 26 ("[A]nti-classification masks a history of distinguishing between benign and malicious uses of protected class attributes."); Bent, *supra* note 212, at 852 ("[A]nticlassificatory ideal of colorblindness is impossible in the machine-learning context, because we cannot create truly colorblind algorithms. Machines are too effective in identifying proxies.").

[270] *See supra* Part I.C.

[271] *See supra* Part I.C. (discussing the "effects-based" theory of disparate impact).

[272] As long as lenders were able to demonstrate that some disparity was the result of a legitimate business interest, the disparities were tolerable. Susan S. Grover, *The Business Necessity Defense in Disparate Impact Discrimination Cases,* 30 GA. L. REV. 387, 387 (1996); *see* Gillis & Spiess, *supra* note 36, at 470–71 n.35. How broadly or narrowly the business justification was defined determines the weight given to those business interests over the goal of reducing credit market disparities. If the "business justification" of the disparate impact doctrine is to be taken to mean literally any business justification, then any algorithm that increases prediction accu-

existing burden shifting framework is unsuitable in the algorithmic setting.[273] Outcome testing provides a new set of tools to functionally perform the same kind of balancing of policy goals.

1. Comparing Borrowers Who Are Similarly Situated

An important question for discrimination law is whether borrowers who are similarly situated are treated the same.[274] In traditional disparate impact cases, this is required as part of the prima facie case. Discrimination law has long recognized that there are differences that are a legitimate basis on which to distinguish between borrowers.[275] Despite the significance of the definition of who is "similarly situated" under traditional fair lending, there is little guidance on this question.[276]

racy arguably meets the threshold of the justification.

[273] *See* Gillis & Spiess, *supra* note 36, at 462.

[274] This requirement originates in the seminal Title VII case, *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973). Some courts were willing to extend the "McDonnell Douglas standard" to the credit context. *See* Robert G. Schwemm, *Introduction to Mortgage Lending Discrimination Law*, 28 J. MARSHALL L. REV. 317, 329 (1995) (summarizing fair lending cases and the requirement for the plaintiff 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). For a more skeptical view of the application of the "similarly situated" requirement to the credit context, see Judge Posner in *Latimore v. Citibank Federal Savings Bank*, 151 F.3d 712, 713 (7th Cir. 1998). 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*, 120 YALE L.J. 728 (2011) (discussing problems that arise from the judiciary's dependence on "comparators" in evaluating discrimination claims); Ernest F. Lidge III, *The Courts' Misuse of the Similarly Situated Concept in Employment Discrimination Law*, 67 MO. L. REV. 831 (2002).

[275] According to the Supreme Court in *Inclusive Communities*, even the prima facie case of the plaintiff cannot rely only on a showing of disparities. Tex. Dep't of Hous. & Cmty. Affairs v. Inclusive Cmty. Project, Inc., 576 U.S. 519, 541 (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.").

[276] 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 (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 a disparate impact case, see the expert opinion discussed in Ayres et al., *supra* note 102, at 235–39. 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 include. *Id.* at 236–39.

There is some ambiguity over whether the requirement to demonstrate that a

A new interpretation to this old question can be used as an outcome-based test in the algorithmic setting. In a 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 protected groups, controlling for the legitimate grounds for differentiation? In a sense, this question reverse engineers the basic classification question of whether borrowers are distinguished based on the protected characteristic.[277] 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.[278] This set can be intuitively understood as adding control variables into a regression in that they explain differences between people.[279] 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.[280] As this set expands, more of the

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 274, at 746–47 (discussing how circuits differ on this issue); *see also* Schwemm, *supra* note 274, at 328–31 (describing the utility of finding a similarly situated individual for proving intentional discrimination in mortgage discrimination cases using the burden-shifting framework). Furthermore, it is unclear whether the requirement is part of a disparate treatment case as well as a disparate impact case. *See* Goldberg, *supra* note 274, at 731–33 (outlining the confusion of the comparator analysis).

[277] *See* Goldberg, *supra* note 274, at 731 ("[E]valuating allegations of discrimination requires courts and others to see something that is not observable directly: whether an accused discriminator has acted because of a protected characteristic.").

[278] There are some similarities between my framework and the framework proposed by Dwork et al. *See supra* note 58. Their approach is based on a similarity metric between individuals who are treated fairly if the classifier ensures similar outcomes for similar individuals. *Id.* at 214.

[279] This is similar to the analysis discussed in Ayres et al. *See supra* note 102. 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. *Id.* at 235–39. A recent paper suggests that controlling for covariates may produce skewed results and proposes a method to correct for omitted variable bias. *See* Jongbin Jung, Sam Corbett-Davis, Ravi Shroff & Sharad Goel, *Omitted and Included Variable Bias in Tests for Disparate Impact*, ARXIV, Aug. 30, 2019, at 2 https://arxiv.org/pdf/1809.05651 [https://perma.cc/57U7-LD3J], (introducing the authors' "risk-adjusted regression" method).

[280] *See* Goldberg, *supra* note 274, at 756–57 (noting courts' concerns with small sample sizes when evaluating comparators in discrimination cases).

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.[281] The difference between the empirical question of who is the same versus who should be treated the same becomes particularly apparent when we consider that fair lending law prohibits discrimination based on protected characteristics, even if they are directly related to default.[282] As discussed above, age and marital status may change a borrower's default risk, yet these characteristics cannot be used to distinguish between people.[283]

Testing for disparities among the "similarly situated" may seem like a return to input-based approaches, as it relies on the selection of the legitimate bases for differentiation.[284] If the test requires selecting normatively relevant criteria for distinction, then it may look similar to restricting an algorithm to pre-approved inputs.[285] However, this test differs from the input-based approaches that I criticized in Part III.[286] That is because restricting an algorithm's inputs to the similarly situated set would, while sufficient, not be necessary for this test. After all, there may be many inputs that increase prediction accuracy while not creating significant disparities.[287] This is especially important in the case of characteristics that would help increase access to credit for protected groups but are unlikely to be included in the similarly situated set, such as timely rental payments.[288] Moreover, a regulator may set the tolerance level such that some disparity is permissible when using inputs beyond the "similarly situated" set.

[281] Note that the similarly situated set is separate from the set of characteristics that is predictive of the outcome. *See generally* Ayres et al., *supra* note 102, at 235–39 (outlining the methodology of an expert witness in disparate impact lending litigation). 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. *See* discussion *supra* Part III.

[282] *See supra* notes 197–200 and accompanying text.

[283] *See supra* notes 197–200 and accompanying text.

[284] *See* discussion *supra* Part I.C.

[285] *See supra* notes 222–23 and accompanying text.

[286] *See* discussion *supra* Part III.

[287] *See* Xiang, *supra* note 207, at 23 ("[I]ncluding [protected class variables] can actually improve both fairness and accuracy, depending on how they are used.").

[288] *See* discussion *supra* Part II.C.2.

In general, creating a test that relies on similarly situated characteristics makes the tradeoff between accuracy and other policy goals explicit, rather than rendering it opaque as input-based approaches do when they restrict inputs to those that seem intuitively relevant to default.[289] It also means that this set can be adjusted and tested, whereas the restrictions of input-based approaches are not legible and adaptable.[290] The one disadvantage of this approach, however, 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.[291]

2. Considering Incremental Change

Another meaningful way to consider the disparities created by algorithmic pricing is to do so relative to a baseline, such as traditional credit pricing. Rather than considering the absolute levels of disparities created by a pricing rule, as in Figure 12,[292] the focus is on how these disparities compare to traditional credit pricing rules.[293] Similarly, a regulator could compare the prices produced under the use of traditional lending variables with new data available to a lender, such as consumer and payment behavior. In fact, the type of analysis conducted throughout this paper, and particularly in Parts II.C[294] and III.C,[295] would be the starting point to analyze the incremental effects of new technologies.

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.[296] When personalized pricing relies on biased inputs, it is unlikely to ever produce

[289] See discussion *supra* Part I.A for an overview of how credit decisions are made.

[290] *But see supra* note 66 and accompanying text (discussing some discretion enjoyed by loan officers).

[291] *See* discussion *supra* Part I.B.2.

[292] For an in-depth explanation of Figure 12, see discussion *infra* Appendix B.

[293] See discussion *supra* Part I for an overview of traditional credit pricing.

[294] The analysis in this Part looked at the way a new statistical technology can increase disparities when relying on "biased world" inputs, and how the use of big data can reduce disparities caused by "biased measurement."

[295] The analysis in this Part considered the tradeoff between accuracy and disparity.

[296] *See* Berk et al., *supra* note 262, at 31 ("At the same time, the benchmark is current practice. By that standard, even small steps, imperfect as they may be, can in principle lead to meaningful improvements in criminal justice decisions. They just need to be accurately characterized.").

pricing that is not disparate for protected groups.[297] This type of test is therefore more appropriate for the effect-based interpretation of disparate impact,[298] as it seeks to balance both the concern for further entrenching disadvantage with the interests of lenders and importance of functioning credit markets.

Furthermore, as discussed in Section II.C.2, 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 in turn benefit vulnerable groups.[299] 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 CFPB sent to an algorithmic lender, Upstart, in 2017 and later extended in 2020.[300] In its update on the No-Action Letter on 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)."[301] The focus of Upstart's analysis was therefore the

[297] *See* discussion *supra* Part I.B.

[298] *See supra* notes 101–02 and accompanying text.

[299] *See supra* note 26 and accompanying text.

[300] Letter from Christopher M. D'Angelo, Assoc. Dir. For Supervision, Enf't & Fair Lending, CFPB, to Thomas P. Brown, Paul Hastings, LLP (Sept. 14, 2017), https://files.consumerfinance.gov/f/documents/201709_cfpb_upstart-no-action-letter.pdf [https://perma.cc/DC79-TYMY] [hereinafter 2017 No-Action Letter]; Letter from Edward Blatnik, Acting Assistant Dir., Off. of Innovation, CFPB, to Alison Nicoll, Gen. Couns., Upstart Network, Inc. (Nov. 30, 2020), https://files.consumerfinance.gov/f/documents/cfpb_upstart-network-inc_no-action-letter_2020-11.pdf [https://perma.cc/4NSP-4T6V] [hereinafter CFPB 2020 No-Action Letter]. The 2017 No-Action Letter was the first and only No-Action letter that the CFPB had provided. *See* Ficklin & Watkins, *supra* note 177. For the general policy, see Policy on No-Action Letters; Information Collection, 81 Fed. Reg. 8,686 (Feb. 22, 2016).

[301] Ficklin & Watkins, *supra* note 177; *see also* Patrice Ficklin, Tom Pahl & Paul Watkins, *Innovation Spotlight: Providing Adverse Action Notices When Using AI/ML Models*, CFPB (July 7, 2020), https://www.consumerfinance.gov/about-us/blog/innovation-spotlight-providing-adverse-action-notices-when-using-ai-ml-models

incremental change in moving from traditional credit pricing to algorithmic credit pricing. It is this type of analysis that should form the core of fair lending analysis.

In its most recent No-Action letter to Upstart, the CFPB provided further details on the type of testing Upstart is required to perform.[302] The letter creates a framework for the periodic reporting, which it calls the Model Risk Assessment Plan (MRAP). Under MRAP, Upstart is required to report on the predictive accuracy by group and to test how its model compares to other credit models in enabling access to credit.[303]

An incremental approach could use an evolving baseline, allowing gradual progress towards reducing disparities. Using the status quo as the baseline is useful when considering whether to allow the initial use of a new technology but runs the risk of setting a low bar for fintech lenders in the long run. Therefore, as consumer lending moves beyond traditional credit pricing, there is a need to set a new bar of what is expected from lenders that use advance prediction technologies and new data.

In summary, outcome-based testing could provide important information about two questions that are meaningful to discrimination analysis: whether similarly situated borrowers are treated the same and whether a change in pricing increases or decreases disparity, which are highly relevant both under "credit discrimination as classification" and "credit discrimination as discriminatory effects."[304]

B. 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, with respect to both the decision inputs and how those inputs are used to produce predictions and pricing rules.[305] 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.[306]

[https://perma.cc/2BJE-8E7D] (commenting on AI's ability to comport with existing regulations).

[302] CFPB 2020 No-Action Letter, *supra* note 300, at 3.

[303] *Id.* at 3.

[304] *See* discussion *supra* Part I.C.

[305] *See* discussion *supra* Part II.A.

[306] *See* discussion *supra* Part I.A.

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. Machine learning pricing introduces new transparency and can therefore lead to 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.[307] In the traditional credit pricing context, regulators respond to materialized prices, meaning actual prices charged to actual borrowers.[308]

It is notable that the EU proposed regulation of AI from April 2021, which provides a comprehensive framework for the regulation of AI across many domains,[309] adopts a model of ongoing testing and reporting. The proposal requires system developers to engage in ex-ante and ongoing testing of algorithms, the results of which are reported to the relevant national supervisory authorities, depending on the domain in which the AI is deployed.[310] The proposal, which designates creditworthiness assessments as a use of AI that is "high-risk"[311] and therefore subject to heightened regulation, requires developers of AI systems to lay down risk management systems that test the system prior to placement of the product on the market,[312] as well as post-market monitoring.[313] The proposal also requires the monitoring of bias and its detection and correction, explicitly allowing the use of protected characteristics in this process.[314] Importantly, the proposal requires developers to create a conformity assessment, reported to the regulator, in order to certify that they have met the requirements of the proposal.[315]

In the machine learning pricing context, regulators themselves can analyze pricing rules before they are applied to real borrowers, similar to the simulations throughout this paper, creating the poten-

[307] *See* Hurley & Adebayo, *supra* note 5, at 160–61 (noting machine learning's ability to detect patterns to predict future data).

[308] *See supra* note 259 and accompanying text.

[309] European Union Proposal*, supra* note 17, at 21.

[310] *Id* at 49*.* ("[National supervisory authorities must comply with the requirements in Chapter 2 and] provide national competent authorities and notified bodies with all the necessary information to assess the compliance of the AI system with those requirements.").

[311] *Id.* at 27 and Annex III (containing the full list of high-risk uses).

[312] *Id.* at 46–48.

[313] *Id.* at 74–75.

[314] *Id.* at 48.

[315] *Id.* at 58–59.

tial for ex ante testing. As argued throughout this Article, the effects of changes in credit markets on disparities between groups is unclear and cannot be adequately studied from a theoretical perspective. This means that only testing 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. They are unsure about how to comply with discrimination law in a machine learning setting. The outcomes-based testing approach that I propose would provide lenders with valuable legal certainty.

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 has 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 the 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.[316] 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.[317] 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

[316] *Committee Passes Bills to Promote Innovation, Strengthen the Financial System and Protect Consumers, Small Businesses and Investors*, U.S. HOUSE COMM. ON FIN. SERVS. (May 9, 2019), https://financialservices.house.gov/news/documentsingle.aspx?DocumentID=403739 [https://perma.cc/6QLM-PAF3].

[317] Fair Lending Report of the Bureau of Consumer Financial Protection, June 2019, 84 Fed. Reg. 32,420, 32,422–23 (July 8, 2019).

practices do not increase disparities.[318] Additionally, the recent Request for Information on the use of AI in finance—emphasizing the use of AI in credit—reflects the coordinated and increased interest in creating a regulatory regime for AI lending.[319] These regulatory efforts indicate that 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 approaches to resolving the tension between old law and new realities are not promising. Current approaches are inadequate because they continue to commit the input fallacy, even though scrutinizing decision inputs as in traditional fair lending is no longer feasible or effective in the algorithmic context. This input fallacy is committed by both proponents and opponents of a broad disparate impact 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 should 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 impossible 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 allows regulators to respond to the unique discrimination challenges posed by the fintech industry and to leverage technological advancement for the sake of greater fairness in lending.

But the conclusions go beyond credit pricing. They apply to all 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 leave behind the input fallacy, recognize the new challenges of algorithmic decision-making, and embrace its opportunities.

[318] Ficklin & Watkins, *supra* note 177 (noting that innovations like machine learning may expand access to credit).

[319] Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning, 86 Fed. Reg. 16,837 (March 31, 2021).

APPENDICES

APPENDIX A: SIMULATION DATA[320]

As discussed in the main Article, I demonstrate my main points in a stylized simulation exercise that is calibrated to real data.[321] 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 approaches to the application of discrimination law to algorithmic decision-making. I also use the simulation exercise to present my proposed regulatory framework.

The simulation demonstration is based on real mortgage application data from the Boston Fed HMDA dataset. In general, the HMDA requires mortgage lenders to disclose loan-level information on mortgage applications and whether they were granted or denied.[322] 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.[323] I specifically use the Boston Fed HMDA dataset,[324] 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.[325]

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.[326]

[320] The following is adapted from a previous publication that I co-authored. *See* Gillis & Spiess, *supra* note 36.

[321] *See* discussion *supra* Part II.B.

[322] *See* Munnell, et al., *supra* note 77, at 25 (mentioning data points collected because of the HMDA).

[323] *See supra* note 206 and accompanying text.

[324] *See* Munnell et al., *supra* note 77, at 25–28 (describing how the Boston Fed created this unique dataset and a discussion of their findings).

[325] HMDA data do not contain information about credit histories, debt burdens, or loan-to-value ratios among other factors. *Id.* at 25. The Boston Fed also used census data on neighborhood characteristics. *Id.* at 26.

[326] *Id.* at 13–14, 32.

The most significant advantage of using HMDA data is that they contain demographic characteristics, such as borrower race, gender, age, and marital status along with various neighborhood characteristics.[327] 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 variables discussed in Subsection II.A.1; however, the types of variables are broader than what is typically used by mortgage originators in setting the "par-rate" in traditional lending.[328]

The Boston Fed HMDA dataset 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 needs 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.[329]

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[330] 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

[327] Such as the appreciation of housing properties in the neighborhood. *Id.*

[328] For a full description of the variables in the Boston Fed HMDA dataset, see *id.*

[329] The methodology is similar to that discussed in the Online Appendix in Gillis & Spiess, *supra* note 36. Further details are provided in Appendix A.

[330] TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE AND PREDICTION 119–29 (2d ed. 2009) (discussing logistic regression).

matched HMDA-McDash dataset.[331] As a result, for every individual in the Boston Fed 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 Fed HMDA dataset. Second, for every individual in the bootstrap sample, that individual's default is simulated based on the default probability 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 these 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:

$$\text{predicted probability of default } = \text{logistic}(\alpha + \beta_1 \text{characteristic}_1 + \beta_2 \text{characteristic}_2 + \ldots)$$

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 10. 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% will have an 80% probability of default.

[331] *See* Fuster et al., *supra* note 65, (manuscript at 12) (explaining the use of both the HMDA and McDash datasets).

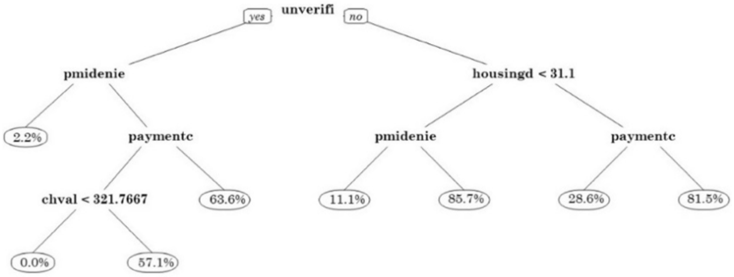

Figure 10: A decision tree that predicts default probability on simulated data

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 Article, I use a rule obtained 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 be a group of borrowers who were classified as "defaulters" but end up repaying the loan and not defaulting (type I error).[332] 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 in-

[332] 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.

crease the type II error. The more accurate a prediction, the smaller the tradeoff between these two types of 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 were correctly classified as "positive" relative to the total number of people classified as "positive":

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

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 that did indeed default. 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":

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

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 11 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. Figure 11 shows that the prediction for white borrowers is more accurate for nearly every classification cutoff.

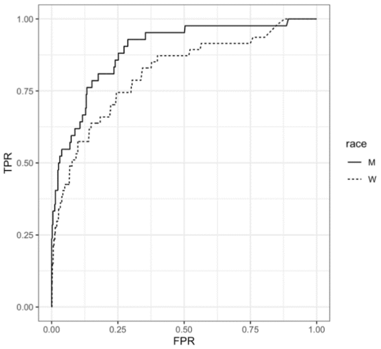

Figure 11: ROC curve for risk prediction using all inputs (other than race), plotted separately for whites (W) and non-whites (M) borrowers.

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 a useful metric to describe overall relative accuracy.

Another example is Figure 12, which plots the ROC curve for the prediction of "marital status" from other HMDA dataset variables. Figure 12 shows that a borrower's marital status can be predicted fairly accurately using the other HMDA variables.

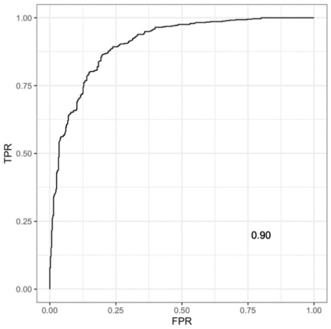

Figure 12: ROC curve for prediction of "marital status." The "marital status" variable is a dummy variable equal to 1 if the applicant is married and 0 if the applicant is unmarried or separated in the Boston Fed HMDA dataset. The ROC curve plots the true-positive-rate and false-negative-rate for different cut-off rules. The number in the lower right corner is the Area Under Curve (AUC).