ALGORITHMIC DISCRIMINATION AND INPUT ACCOUNTABILITY UNDER THE CIVIL RIGHTS ACTS

Robert Bartlett[*]
Adair Morse[‡]
Nancy Wallace[†]
Richard Stanton[°]

Abstract

The disproportionate burden of COVID-19 among communities of color, together with a necessary renewed attention to racial inequalities, have lent new urgency to concerns that algorithmic decision-making can lead to unintentional discrimination against members of historically marginalized groups. These concerns are being expressed through Congressional subpoenas, regulatory investigations, and an increasing number of algorithmic accountability bills pending in both state legislatures and Congress. To date, however, prominent efforts to define algorithmic accountability have tended to focus on output-oriented policies that may facilitate illegitimate discrimination or involve fairness corrections unlikely to be legally valid. Worse still, other approaches focus merely on a model's predictive accuracy—an approach at odds with long-standing U.S. antidiscrimination law.

We provide a workable definition of algorithmic accountability that is rooted in the caselaw addressing statistical discrimination in the context of Title VII of the Civil Rights Act of 1964. Using instruction from the burden-shifting framework, codified to implement Title VII, we formulate a simple statistical test to apply to the design and review of the inputs used in any algorithmic decision-making processes. Application of the test, which we label the *input accountability test*, constitutes a legally viable, deployable tool that can prevent an algorithmic model from systematically penalizing members of protected groups who are otherwise qualified in a legitimate target characteristic of interest.

[*] I. Michael Heyman Professor of Law & Faculty Co-Director of the Berkeley Center for Law and Business - UC Berkeley School of Law.

[‡] Soloman P. Lee Chair in Business Ethics and Associate Professor – UC Berkeley Haas School of Business.

[†] Professor & Lisle and Roslyn Payne Chair in Real Estate Capital Markets, Co-Chair, Fisher Center for Real Estate and Urban Economics – UC Berkeley Haas School of Business.

[°] Professor & Kingsford Capital Management Chair in Business Economics – UC Berkeley Haas School of Business.

ALGORITHMIC DISCRIMINATION AND INPUT ACCOUNTABILITY UNDER THE CIVIL RIGHTS ACTS

TABLE OF CONTENTS

I. INTRODUCTION

In fall 2019, the journal *Science* published research showing troubling evidence on inadvertent racial discrimination in the algorithm of health insurer UnitedHealth.[1] Hospitals were using the algorithm to allocate limited

[1] *See* Melanie Evans & Anna Wilde Mathews, *New York Regulator Probes UnitedHealth Algorithm for Racial Bias*, WSJ (Oct. 26, 2019), https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601.

hospital resources to the sickest patients. However, the researchers showed—because the algorithm used a patient's cost of care as the metric for gauging sickness and because African-American patients historically incurred lower costs for the same illnesses and level of illness—it caused African-Americans to receive substandard care as compared to white patients.[2] In this instance, not only did the seemingly race-blind algorithm produce bias, but it did so because of structural inequalities that caused African-Americans to exhibit a lower cost per illness, i.e., historically being unable to (or being advised to) spend less on healthcare relative to white patients.

A similar, but gender-focused, instance of algorithmic bias emerged at the same time, when Apple Inc. debuted its much-anticipated Apple Card.[3] Within weeks, Twitter was abuzz with headlines that the card's credit approval algorithm was systematically biased against women,[4] followed by the New York State Department of Financial Services announcing an investigation.[5]

Despite the potential for algorithmic decision-making to eliminate face-to-face biases, these episodes provide vivid illustrations of the widespread concern that algorithms may nevertheless engage in discrimination, even inadvertently.[6] The meteoric growth in algorithmic decision-making, spawned by the availability of unprecedented data on individuals and the accompanying rise in techniques in machine learning and artificial intelligence have greatly heighted this concern. Moreover, the laying bare of the inequalities and structural racism evident from the COVID-19 pandemic and the concurrent renewed attention on civil rights has heighten the necessity and urgency of addressing algorithmic bias. Indeed, acting on mounting anecdotes and evidence even before the pandemic, New York City,[7] Washington State,[8] and Congress[9] all introduced algorithm accountability bills to regulate governmental or corporate use of algorithms. Yet, a notable absence in these legislative efforts is a formal standard for courts or regulators

[2] *See* Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

[3] *See* Press Release, Apple Inc., *Introducing Apple Card, A New Kind of Credit Card Created by Apple* (March 25,2019), https://www.apple.com/newsroom/2019/03/introducing-apple-card-a-new-kind-of-credit-card-created-by-apple/.

[4] *See* Sridhar Natarajan & Shahien Nasiripour, *Viral Tweet About Apple Card Leads to Goldman Sachs Probe*, BLOOMBERG (Nov. 19, 2019), https://www.bloomberg.com/news/articles/2019-11-09/viral-tweet-about-apple-card-leads-to-probe-into-goldman-sachs.

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

[6] *See, e.g.*, Salon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL L. REV. 671, 673 (2016) ("If data miners are not careful, the process can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination.").

[7] *See* Zoë Bernard, *The First Bill to Examine 'Algorithmic Bias' in Government Agencies Has Just Passed in New York City*, BUSINESSINSIDER (Dec. 19, 2017), http://www.businessinsider.com/algorithmic-bias-accountability-bill-passes-in-new-york-city-2017-12?IR=T.

[8] H.B. 1655, 66th Leg., Reg. Sess. (Wash. 2019).

[9] H.R. 2231, 116th Cong. (2019).

to deploy in evaluating algorithmic decision-making, raising the fundamental question: *What exactly does it mean for an algorithm to be accountable?*

In this Article, we provide an answer. Central to our framework is the recognition that, despite the novelty of artificial intelligence and machine learning, existing U.S. antidiscrimination law has long provided a workable definition of decision-making accountability dating back to Title VII of the Civil Rights Act of 1964.[10] What has been missing is a translation of this definition into the context of statistical modelling at the heart of algorithmic decision-making. The first of our two primary contributions is thus to define algorithmic accountability following Title VII. Our second contribution emerges naturally from the first. The definition of what it means for an algorithm to be accountable under discrimination law lends itself to a formal test of accountability. We put forward a workable test that regulators, courts, and data scientists can apply in examining whether an algorithmic decision-making process complies with long-standing antidiscrimination statutes and caselaw.

Title VII and the caselaw interpreting it define what it means for any decision-making process—whether human or machine—to be accountable under U.S. antidiscrimination law. At the core of this caselaw is the burden-shifting framework initially articulated by the Supreme Court in *Griggs v. Duke Power Co.*[11] Under this framework, plaintiffs putting forth a claim of unintentional discrimination under Title VII must demonstrate that a particular decision-making practice (e.g., a hiring practice) lands disparately on members of a protected group.[12] If successful, the framework then demands that the burden shift to the defendant to show that the practice is "consistent with business necessity."[13] If the defendant satisfies this requirement, the burden returns to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use.[14] The focus of Title VII is on discrimination in the workplace, but the analytical framework for unintentional discrimination that emerged from the Title VII context now spans other domains and applies directly to the type of unintentional, statistical discrimination utilized in algorithmic decision-making.[15]

[10] 42 U.S.C. § 2000e (2012).

[11] Griggs v. Duke Power Co., 401 U.S. 424, 432 (1971).

[12] *See* Dothard v. Rawlinson, 433 U.S. 321, 329 (1977).

[13] 42 U.S.C. § 2000e–2(k) (2012); *see also Griggs*, 401 U.S. at 431 (noting that in justifying employment practice that produces disparate impact, "[t]he touchstone is business necessity").

[14] *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975).

[15] For example, this general burden-shifting framework has been extended to other domains where federal law acknowledges the possibility of claims of unintentional discrimination under a disparate impact theory. *See, e.g.,* Texas Dep't of Housing & Cmty. Affairs v. Inclusive Cmtys. Project, Inc., 135 S. Ct. 2507 (2015), (adopting the burden-shifting framework for disparate impact claims under the Fair Housing Act); Ferguson v. City of Charleston, 186 F.3d 469, 480 (4th Cir. 1999) (discussing cases adopting the Title VII burden-shifting framework in Title VI disparate impact cases), *rev'd on other grounds*, 532 U.S. 67 (2001).

The feature of the burden-shifting framework that is often overlooked in the recent legal and economic literature on algorithmic bias[16] is the second step of the analysis. This step requires a showing that a process and its inputs satisfy a legitimate business necessity. This implies that outcome-focused tests and fixes (i.e., actions to make sure a decision equalizes outcomes, conditional on observables, across race or some other protected category) are insufficient actions for an algorithmic decision-maker to ensure compliance. Output-oriented policies, as we will discuss at length, are critical for predictive accuracy and fairness arguments, but fail to comply with input accountability in Title VII.

To see why, consider the facts of the Supreme Court's 1977 decision in *Dothard v. Rawlinson.*[17] There, a prison system desired to hire job applicants who possessed a minimum level of strength to perform the job of a prison officer, but the prison could not directly observe which applicants satisfied this requirement.[18] Consequently, the prison imposed a minimum height and weight requirement on the assumption that these observable characteristics were correlated with the requisite strength required for the job.[19] This procedure resulted in adverse hiring outcomes for female applicants, resulting in a class of female applicants bringing suit under Title VII for gender discrimination.[20] Deploying the burden-shifting framework, the Supreme Court first concluded that the plaintiffs satisfied the disparate outcome step,[21] and it also concluded that the prison had effectively argued that hiring applicants with the requisite strength could constitute a business necessity.[22] However, the Court ultimately held that the practice used to discern strength—relying on the proxy variables of height and weight—did not meet the "consistent with business necessity" criterion.[23] Rather, absent evidence showing the precise relationship between the height and weight requirements to "the requisite amount of strength thought essential to good job performance,"[24] height and weight were noisy estimates of strength that risked penalizing females over-and-above these variables' relation to the prison's business necessity goal. In other words, height and weight were likely to be biased estimates of required strength whose use by the prison risked systematically penalizing female applicants who were, in fact, qualified.

The Court thus illustrated that in considering a case of statistical discrimination, the "consistent with business necessity" step requires the

[16] *See infra* Part II(B) and Part II(C).
[17] 433 U.S. 321 (1977).
[18] *Id*. at 331-32.
[19] *Id.*
[20] *Id*. at 323.
[21] *Id*. at 330-31.
[22] *Id*. at 332.
[23] *Id*.
[24] *Id*.

assessment of two distinct questions. First, is the use of proxies for an unobservable "target" characteristic (e.g., requisite strength) done in pursuit of a fundamental business necessity? Second, even with a legitimate target characteristic and predictive proxy input variables, are these input variables noisy at estimating the legitimate business necessity in a way that will systematically penalize members of a protected group who are otherwise qualified?

The first question involves defining a business necessity model that a court agrees can justify disparate outcomes across protected and unprotected groups. Often, the targets within the business necessity model are unobservable attributes or latent concepts an individual might possess. For example, a court might deem required strength, reliability, and intelligence to be valid targets within the business necessity model of prison officer employment. Likewise, in lending, courts have long held that, under the Fair Housing Act (FHA),[25] an individual's creditworthiness is an acceptable business necessity;[26] thus, variables capturing the expected cash flow of the individual enabling repayment are the targets for informing loan decisions.

The second question involves assessing a proxy input variable's relation with the target and protected categories. Interpreting *Dothard,* a proxy variable should only be related to the protected category through its relation to a valid target. In the case of *Dothard*, height can only be related to gender through its relationship to required strength. Likewise in the context of lending, redlining is prohibited because it violates this criterion.[27] A lender who engages in redlining refuses to lend to residents of a majority-minority neighborhood on the assumption that the average unobservable credit risk of its residents is higher than those of observably-similar but non-minority neighborhoods.[28] By assuming that all residents of minority neighborhoods have low credit, redlining systematically penalizes minority borrowers who actually have high credit worthiness.

[25] 42 U.S.C. §§ 3601-3619 (2012).

[26] *See infra* note 118.

[27] *See, e.g.,* Laufman v. Oakley Bldg. & Loan Co., 408 F. Supp. 489, 493 (S.D. Ohio 1976) (redlining on the basis of race violates the "otherwise make unavailable or deny" provision of § 3604(a) of the FHA); (interpreting identical language in § 3604(f)(2) of the FHA as prohibiting insurance redlining); Strange v. Nationwide Mut. Ins. Co., 867 F. Supp. 1209, 1213–14 (E.D. Pa. 1994) (insurance redlining); NAACP v. Am. Family Mut. Ins., 978 F.2d 287, 297 (6th Cir. 1995) (insurance redlining); *Laufman*, 408 F. Supp. at 496–97 (mortgage redlining); Nationwide Mut. Ins. Co. v. Cisneros, 52 F.3d 1351 (7th Cir. 1995) (insurance redlining); Wai v. Allstate Ins. Co., 75 F. Supp. 2d 1, 7 (D.D.C. 1999); Lindsey v. Allstate Ins. Co., 34 F. Supp. 2d 636, 641–43 (W.D. Tenn. 1999) (insurance redlining). Regulatory agencies charged with interpreting and enforcing the lending provisions of the FHA have defined redlining to include "the illegal practice of refusing to make residential loans or imposing more onerous terms on any loans made because of the predominant race, national origin, etc. of the residents of the neighborhood in which the property is located. Redlining violates both the FHA and ECOA." Joint Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18,266 (1994).

[28] The term red-lining derives from the practice of loan officers evaluating home mortgage applications based on a residential map where integrated and minority neighborhoods are marked off in red as poor risk areas. Robert G. Schwemm, *Housing Discrimination* 13–42 (Release # 5, 1995).

These two insights from *Dothard*—that statistical discrimination must be grounded in the search for a legitimate target variable and that the input proxy variables for the target cannot systematically discriminate against members of a protected group who are qualified in the target—remain as relevant in today's world of algorithmic decision-making as they were in 1977. The primary task for courts, regulators, and data scientists is to adhere to them in the use of big data implementations of algorithmic decisions (e.g., in employment, performance assessment, credit, sentencing, insurance, medical treatment, college admissions, advertising, etc.).

Fortunately, Title VII's burden-shifting framework, viewed through basic principles of statistics, provides a way forward. We recast the logic that informs *Dothard* and courts' attitude towards redlining into a formal statistical test that can be widely deployed in the context of algorithmic decision-making. We label it the *Input Accountability Test (IAT)*.

As we show, the IAT provides a simple and direct diagnostic to determine whether an algorithm is accountable under U.S. antidiscrimination principles. A user of an algorithm (e.g., a business or a regulator) seeking to satisfy the IAT would do so by turning to historical data called "training data" that was originally used to calibrate the algorithm. In settings such as employment or lending where courts have explicitly articulated a legitimate business target (e.g., a job required skill or creditworthiness),[29] the first step would be establishing that the "target" variables sought by the algorithm are indeed business necessity variables. Second, taking a proxy input variable (e.g., height) that the predictive model utilizes, the next step requires decomposing the proxy's variation across individuals into that which correlates with the target variable (or variables) and an error component. The final step requires testing whether that error component remains correlated with the protected category (e.g., gender). If the error is uncorrelated, this means the proxy input variable is unbiased with respect to a protected group; therefore, it will pass the IAT. In this fashion, the test provides a concrete method to harness the benefits of statistical discrimination with regard to predictive accuracy while avoiding the risk that it systematically penalizes members of a protected group who are, in fact, qualified in the target characteristics of interest.

We provide an illustration of the IAT in the *Dothard* setting, not only to provide a clear depiction of the power of the test, but also to introduce several challenges in implementing it and suggested solutions. These challenges include multiple incarnations of measurement error in the target, as well as understanding what "significantly correlated" means in our era of massive datasets. We offer an approach that may serve as a way forward. Beyond the illustration, we also provide a simulation of the test inspired by *Dothard* using a randomly constructed training dataset of 800 prison officers.

[29] *See infra* Part IV(A).

We also illustrate how the IAT can be deployed by courts, regulators, and data scientists. In addition to employment, we list a number of other sectors—including credit, parole determination, home insurance, school and scholarship selection, and tenant selection—where unintentional discrimination is also policed through the burden-shifting framework inspired by Title VII and where courts or statutes have provided explicit instructions regarding what can constitute a legitimate business necessity target.[30] In these settings, application of the IAT can provide a critical tool for ensuring algorithmic decision-making is lawful. We also discuss other domains such as automobile insurance and health care where claims of algorithmic discrimination have recently surfaced, but where existing discrimination laws are less clear whether liability can arise for unintentional discrimination. For those concerned about algorithmic discrimination in these domains, our discussion underscores the special need for algorithmic accountability legislation in these contexts. In the meantime, businesses in these domains are left to self-regulate—often through public pressure—and the IAT provides a tool to test their models for bias.

Our approach differs from other approaches to "algorithmic fairness" that focus on "tuning" algorithms to ensure fair outcomes across protected and unprotected groups.[31] We differentiate ourselves from this outcome-based approach for several reasons. First, these approaches often pursue outcome-based adjustments due to a misperception that existing legal prohibitions on unintentional discrimination are ineffective when applied to an accurate algorithmic process. For instance, in their widely-cited article *Big Data's Disparate Impact*,[32] Salon Barocas and Andrew Selbst note that the business necessity defense merely requires that an employment algorithm is "predictive of future employment outcomes."[33] However, as we illustrate, faithful application of the burden-shifting framework reveals that predictive accuracy is not a necessary and sufficient condition to satisfy the business necessity requirement. Rather, cases such as *Dothard* underscore the importance of examining whether proxy input variables are systematically biased against protected groups even if they are predictive of a valid employment outcome. Focusing exclusively on calibrating outcomes thus skips a critical component of the accepted approach to policing unintentional discrimination.

Second, outcome-based approaches can themselves run afoul of U.S. antidiscrimination law, particularly given the Supreme Court's 2009 decision in *Ricci v. DeStefano*.[34] In *Ricci,* a city's efforts to calibrate a decision-making process to equalize hiring outcomes across members of protected and

[30] *See Id*.
[31] *See infra* Part II(B).
[32] Sarocas & Selbst, *supra* note 6.
[33] *Id*. at 672.
[34] 557 U.S. 557 (2009).

unprotected groups—regardless of whether individuals were qualified in a legitimate target of interest—were deemed impermissible intentional discrimination.[35] The decision thus calls into question the legality of explicit race-based adjustments of algorithmic outcomes, as would be required by "tuning" an algorithm to ensure outcomes meet a specified fairness criterion across protected and unprotected groups.

Given *Ricci*, addressing distributional concerns implicated by the rise of algorithmic decision-making requires a clear-eyed understanding of the channels through which algorithms can perpetuate structural inequalities and how these channels can be altered. Our focus on checking targets and testing inputs provides exactly this understanding. Specifically, when presented with an algorithm that produces disparities, our approach highlights the need to examine first whether the algorithm is pursuing a valid target, followed by assessing whether it utilizes an input that fails the IAT. An algorithm that raises neither concern but nevertheless produces disparities thus points toward the need for a broader conversation concerning the fundamental fairness of the target.

As an illustration, consider an example that has recently captured considerable attention in light of the disproportionate burden of the COVID-19 pandemic on communities of color. Given the scarcity of ventilators, many hospitals around the country have turned to algorithms to allocate this life-saving resource. A common approach is to rely on a patient's score from the Sequential Organ Failure Assessment (SOFA) that gauges the degree of dysfunction of six organ systems. As one state agency noted, allocating ventilators based on SOFA scores is "objective" and "equitable" insofar that these "tragically difficult decisions must be based on … biological factors related only to the likelihood and magnitude of benefit from the medical resources."[36] This approach also ensured that "[f]actors that have no bearing on the likelihood or magnitude of benefit, including race, gender, sexual orientation, gender identity, [or] ethnicity …, are irrelevant and not to be considered by providers making allocation decisions."[37]

Notwithstanding this stated desire for an equitable allocation of resources, however, SOFA-based triage algorithms have alarmed many clinicians given the adverse effect they are likely to have on communities of color,[38] informed by the fact that a legacy of structural racism and inequality

[35] We discuss this challenge in more detail in Part II(B).

[36] *See* EXEC. OFFICE OF HEALTH & HUMAN SERVS., MASS. DEP'T OF PUB. HEALTH, CRISIS STANDARDS OF CARE: PLANNING GUIDANCE FOR THE COVID-19 PANDEMIC (2020), https://d279m997dpfwgl.cloudfront.net/wp/2020/04/CSC_April-7_2020.pdf.

[37] *Id*. at 4.

[38] *See, e.g*., Emily Cleveland Manchanda, *Inequity in Crisis Standards of Care*, 383 NEW ENG. J. MED. e16 (2020) (arguing that SOFA-based triage algorithms "penalize people for having conditions rooted in historical and current inequities and sustained by identity-blind policies"); Panagis Galisatsatos et al., *Health Equity and Distributive Justice Considerations in Critical Care Resource Allocation*, 8 LANCET

has caused Black and Latinx Americans to suffer differential rates of chronic and life-shortening conditions, such as hypertension, diabetes, chronic kidney disease, and chronic obstructive pulmonary disease. Under a SOFA-based triage algorithm, people of color would, on average, have lower priority for a ventilator if they contract COVID-19.

Our approach provides a clear means to assess these concerns. Hospitals utilizing a SOFA-based algorithm appear to be doing so based on a desire to allocate resources to patients with the best chance of long-term survival—a target that would appear to be a plausible business necessity target. Application of the IAT would then focus on whether a condition such as diabetes or hypertension is correlated with race or ethnicity beyond its ability to predict survival. Assuming it passes the IAT, then concerns of fairness would require considering whether it is possible to re-define the business necessity target—ideally, with input from experts having diverse perspectives—in a manner that accounts for the structural inequities that contribute to the racial and ethnic disparities in outcomes. Thus, our approach is compatible with concerns about distributional outcomes, but in our view, it is essential to ask first whether these outcomes arise from an invalid algorithmic process.

This Article proceeds as follows. In Part II, we begin by articulating a definition for algorithmic accountability that is at the core of our input accountability test. As we demonstrate there, our definition of algorithmic accountability is effectively a test for "unbiasedness," which differs from various proposals for "algorithmic fairness" that are commonly found in the statistics and computer science literatures. Building on this definition of algorithmic accountability, Part III formally presents the IAT. We begin by situating the IAT within the context of employment, before discussing in Part IV how it can also be applied outside this context. Part V addresses several challenges in implementing the IAT, along with potential solutions. Part VI follows by presenting a simple simulation of how a court, regulator or firm might use the IAT in the setting of *Dothard*. Part VII concludes.

II. ACCOUNTABILITY UNDER U.S. ANTIDISCRIMINATION LAW

A. Accountability and the Burden-Shifting Framework of Title VII

We ground our definition of accountability in the antidiscrimination principles of Title VII of the Civil Rights Act of 1964.[39] Title VII, which focuses on the labor market, makes it "an unlawful employment practice for an employer (1) to ... discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of

RESPIRATORY MED. 758, 759 (2020) (arguing that "SOFA scores might be unfavourably higher in African Americans during this pandemic").

[39] 42 U.S.C. § 2000e (2012).

such individual's race, color, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities ... because of such individual's race, color, religion, sex, or national origin."[40] Similar conceptualizations of antidiscrimination law were later written to apply to other settings, such as the prohibition of discrimination in mortgage lending under the FHA.[41]

In practice, Title VII has been interpreted as covering two forms of impermissible discrimination. The first and "the most easily understood type of discrimination"[42] falls under the *disparate-treatment* theory of discrimination and requires that a plaintiff alleging discrimination prove "that an employer had a discriminatory motive for taking a job-related action."[43] Additionally, Title VII also covers practices which "in some cases, … are not intended to discriminate but in fact have a disproportionately adverse effect on minorities."[44] These cases are usually brought forth under the *disparate-impact* theory of discrimination and allow for an employer to be liable for "facially neutral practices that, in fact, are 'discriminatory in operation,'" even if unintentional.[45]

Critically, in cases where discrimination lacks an intentional motive, an employer can be liable only for disparate outcomes that are unjustified. The *burden-shifting framework*, initially formulated in *Griggs v. Duke Power Co.*[46] and subsequently codified by Congress in 1991,[47] provides the process for understanding when disparities across members of protected and unprotected groups are justified. This delineation is central to the definition of accountability in today's era of algorithms.

Under the burden-shifting framework, a plaintiff alleging unintentional discrimination bears the first burden. The plaintiff must identify a specific employment practice that causes "observed statistical disparities"[48] across members of protected and unprotected groups.[49] If the plaintiff succeeds in establishing this evidence, the burden shifts to the defendant,[50] who must then "demonstrate that the challenged practice is job related for the position in

[40] 42 U.S.C. § 2000e-2(a) (2012).

[41] 42 U.S.C. § 3605 (2012) ("It shall be unlawful for any person or other entity whose business includes engaging in residential real estate-related transactions to discriminate against any person in making available such a transaction, or in the terms or conditions of such a transaction, because of race, color, religion, sex, handicap, familial status, or national origin.").

[42] Int'l Bhd. of Teamsters v. United States, 431 U.S. 324, 335 n.15 (1977).

[43] Ernst v. City of Chi., 837 F.3d 788, 794 (7th Cir. 2016).

[44] Ricci v. DeStefano, 557 U.S. 557, 577 (2009).

[45] *Id*. at 577-78 (quoting Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971)).

[46] Griggs, 401 U.S. at 432.

[47] Civil Rights Act of 1991, Pub. L. No. 102-66, 105 Stat. 1071 (1991).

[48] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 979 (1988).

[49] *See also* Albemarle Paper Co. v. Moody, 422 U.S. 405, 425 (1975) (holding that the plaintiff has the burden of making out a prima facie case of discrimination).

[50] *See id*. at 425 (noting that the burden of defendant to justify an employment practice "arises, of course, only after the complaining party or class has made out a prima facie case of discrimination.")

question and consistent with business necessity."[51] If the defendant satisfies this requirement, then "the burden shifts back to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use."[52]

This overview highlights two core ideas that inform what it means for a decision-making process to be accountable under U.S. antidiscrimination law. First, in the case of unintentional discrimination, disparate outcomes must be justified by reference to a legitimate "business necessity."[53] In the context of employment hiring, for instance, this is typically understood to be a job-related skill that is required for the position.[54] Imagine, for instance, an employer who made all hiring decisions based on applicants' level of a direct measure of a job-related skill. Even if the outcome of these decision-making processes results in disparate outcomes across minority and non-minority applicants, these disparities would be justified as nondiscriminatory with respect to a protected characteristic.

Second, in invalidating a decision-making process, U.S. antidiscrimination law does so because of invalid "inputs" rather than invalid "outputs" or results. This feature of U.S. antidiscrimination law is most evident in the case of disparate treatment claims involving the use by a decision-maker of a protected category in making a job-related decision. For instance, Section (m) of the 1991 Civil Rights Act states that "an unlawful employment practice is established when the complaining party demonstrates that race, color, religion, sex, or national origin was a motivating factor for any employment practice, even though other factors also motivated the practice."[55] However, this focus on inputs is also evident in cases alleging disparate impact, notwithstanding the doctrine's initial requirement that a plaintiff allege disparate outcomes across members of protected and

[51] 42 U.S.C. § 2000e-2(k)(1)(A)(i) (2012); *see also Griggs*, 401 U.S. at 432 ("Congress has placed on the employer the burden of showing that any given requirement must have a manifest relationship to the employment in question.").

[52] Puffer v. Allstate Ins. Co., 675 F.3d 709, 717 (7th Cir. 2012); *see also* 42 U.S.C. § 2000e-2(k)(1)(A)(ii), (C).

[53] 42 U.S.C. § 2000e-2(k)(1)(A)(i). Likewise, even in the case of claims alleging disparate treatment, an employer may have an opportunity to justify the employment decision. In particular, absent direct evidence of discrimination, Title VII claims of intentional discrimination are subject to the burden-shifting framework established in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792 (1973). Under the *McDonnell Douglas* framework, a plaintiff must first "show, by a preponderance of the evidence, that she is a member of a protected class, she suffered an adverse employment action, and the challenged action occurred under circumstances giving rise to an inference of discrimination." *Bennett v. Windstream Commc'ns, Inc.*, 792 F.3d 1261, 1266 (10th Cir. 2015). If the plaintiff succeeds in establishing a prima facie case, the burden of production shifts to the defendant to rebut the presumption of discrimination by producing some evidence that it had legitimate, nondiscriminatory reasons for the decision. *Id*. at 1266.

[54] *See, e.g.*, *Griggs*, 401 U.S. at 432 (holding that the employer's practice or policy in question must have a "manifest relationship" to the employee's job duties); *see also Albermarle*, 422 U.S. at 425 ("If an employer does then meet the burden of proving that its tests are 'job related,' it remains open to the complaining party to show that other tests or selection devices, without a similarly undesirable racial effect, would also serve the employer's legitimate interest in 'efficient and trustworthy workmanship.'").

[55] 42 U.S.C. § 2000e-2(m).

unprotected groups. Recall that even with evidence of disparate outcomes, an employer that seeks to defend against a claim of disparate impact discrimination must demonstrate why these outcomes were the result of a decision-making process based on legitimate business necessity factors (i.e., the disparate outcomes were the result of legitimate decision-making inputs).[56] This focus on "inputs" underscores the broader policy objective of ensuring a decision-making process that is not discriminatory.

The practical challenge in implementing this antidiscrimination regime is that the critical decision-making input—an individual's possession of a job-related skill—cannot be perfectly observed at the moment of a decision, inducing the decision-maker to turn to proxies for it. However, the foregoing discussion highlights that the objective in evaluating these proxy variables should be the same: ensuring that qualified applicants from a protected class are not being systematically passed over for the job or promotion. As summarized by the Supreme Court in *Ricci v. DeStefano*, "[t]he purpose of Title VII 'is to promote hiring on the basis of job qualifications, rather than on the basis of race or color.'"[57]

This objective, of course, is the basis for prohibiting the direct form of statistical discrimination famously examined by economists Kenneth Arrow[58] and Edmund Phelps.[59] In their models, an employer uses a job applicant's race as a proxy for the applicant's expected productivity because the employer assumes that the applicant possesses the average productivity of his or her race. If the employer also assumes the average productivity of minority applicants is lower than non-minorities (e.g., because of long-standing social and racial inequalities), this proxy will ensure that above-average productive minorities will systematically be passed over for the job despite being qualified for it. Because this practice creates a direct and obvious bias against minorities, this practice is typically policed under the disparate treatment theory of discrimination.[60]

Beyond this clearly unlawful form of statistical discrimination, a decision-maker can use statistical discrimination to incorporate not just the protected-class variable but also other proxy variables for the business-necessity unobservable attributes. For instance, an employer might seek to predict a job applicant's productivity based on other observable characteristics that the employer believes are correlated with future productivity, such as an applicant's level of education or an applicant's

[56] *See, e.g.*, Dothard v. Rawlinson, 433 U.S. 321, 331 (1977) (holding that, to satisfy the business necessity defense, an employer must show that a pre-employment test measured a characteristic "essential to effective job performance" given that the test produced gender disparities in hiring).

[57] Ricci v. DeStefano, 557 U.S. 557, 582 (2009) (citing *Griggs* , 401 U.S. at 424).

[58] Kenneth J. Arrow, *The Theory of Discrimination*, *in* DISCRIMINATION AND LABOR MARKETS 3 (Orley Ashenfelter & Albert Rees eds., 1973).

[59] Edmund S. Phelps, *The Statistical Theory of Racism and Sexism*, 62 AM. ECON. REV. 659 (1972).

[60] *See* text accompanying note 55.

performance on a personality or cognitive ability test.[61] Indeed, it is the possibility of using data mining to discern new and unintuitive correlations between an individual's observable characteristics and a target variable of interest (e.g., productivity as a job skill or wealth as a credit risk variable) that has contributed to the dramatic growth in algorithmic decision-making.[62] The advent of data mining has meant that thousands of such proxy input variables are sometimes used.[63]

As the UnitedHealth algorithm example revealed, however, the use of these proxy variables can result in members of a protected class experiencing disparate outcomes. The problem arises from what researchers call "redundant encodings"—the fact that a proxy variable can be predictive of a legitimate target variable *and* membership in a protected group.[64] Relying on these proxy variables therefore risks penalizing members of the protected group who are otherwise qualified in the legitimate target variable.[65] In short, algorithmic accountability requires a method to limit the use proxy variables to those that are consistent with Title VII of the Civil Rights Act and to prohibit the use of those that are not.[66]

B. *Input Accountability Versus Outcome Fairness Approaches*

Our input-based approach differs significantly from that of other scholars who have advanced outcome-oriented approaches to algorithmic accountability. For instance, Talia Gillis and Jann Spiess have argued that the conventional focus in fair lending on restricting invalid inputs (such as a borrower's race or ethnicity) is infeasible in the machine-learning context.[67] Focusing on the context of algorithmic lending, Gillis and Spiess argue that a predictive model of default that excludes a borrower's race or ethnicity can

[61] *See, e.g.,* Neal Schmitt, *Personality and Cognitive Ability as Predictors of Effective Performance at Work*, 1 ANN. REV. ORGANIZATIONAL PSYCHOL. & ORGANIZATIONAL BEHAV. 45, 56 (2014) (describing web-based pre-employment tests of personality and cognitive ability).

[62] *See* Barocas & Selbst, supra note 6, at 677 ("By definition, data mining is always a form of statistical (and therefore seemingly rational) discrimination.").

[63] *See, e.g.*, Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE. J.L. TECH. 148, 164 (2020) (describing how ZestFinance uses an "all data is credit data" approach to predict an individual's creditworthiness based on "thousands of data points collected from consumers' offline and online activities").

[64] *See* Barocas & Selbst, *supra* note 6, at 691 (citing Cynthia Dwork et al., *Fairness Through Awareness*, 3 PROC. INNOVATIONS THEORETICAL COMPUTER SCI. CONF. 214 app. at 226 (2012)).

[65] As noted in the Introduction, redlining represents a classic example: An individual's zip code may be somewhat predictive of one's creditworthiness, but given racialized housing patterns, it is almost certainly far more accurate in predicting one's race. Assuming that all residents in a minority-majority zip code have low creditworthiness will therefore result in systematically underestimating the creditworthiness of minorities whose actual creditworthiness is higher than the zip code average.

[66] In theory, there are statistical methods that would estimate the precise degree to which a redundantly encoded proxy variable predicts a legitimate target variable that is independent of the degree to which it predicts membership in a protected classification. We discuss these methods and their shortcomings *infra* at notes 114 to 116 and in the Appendix.

[67] *See* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV 459 (2019).

still penalize minority borrowers if one of the included variables (e.g., borrower education) is correlated with both default and race.[68] Gillis and Spiess acknowledge the possibility that one could seek to exclude from the model some of these correlated variables on this basis, but they find this approach infeasible given that "a major challenge of this approach is the required articulation of the conditions under which exclusion of data inputs is necessary."[69] They therefore follow the burgeoning literature within computer science on "algorithmic fairness"[70] and advocate evaluating the outcomes from an algorithm against some baseline criteria to determine whether the outcomes are fair across protected and unprotected groups.[71] If they are not, the solution would be to "tune" the algorithm to ensure that they are.[72]

[68] *Id*. at 468-69.

[69] *Id*. at 469. Elsewhere in their article, Gillis and Spiess also suggest that input-based analysis may be infeasible because "in the context of machine-learning prediction algorithms, the contribution of individual variables is often hard to assess." *Id*. at 475. They illustrate this point by showing how in a simulation exercise, the variables selected by a logistic lasso regression in a predictive model of default differed each time the regression was run on a different randomly-drawn subsample of their data. However, this evidence does not speak to how an input-based approach to regulating algorithms would be deployed in practice. A lasso regression—like other models that seek to reduce model complexity and avoid over-fitting—seeks to reduce the number of predictors based on the underlying correlations among the full set of predictor variables. Thus, it can be used in training a model on a set of data with many proxy variables, and running a lasso regression multiple times on different subsamples of the data should be expected to select different variables with each run. However, once a model has been trained and the model's features are selected, the model must be deployed, allowing the features used in the final model to be evaluated and tested for bias. That is, regardless of the type of model fitting technique one uses in the training procedure (e.g., lasso regression, ridge regression, random forests, etc.), the model that is ultimately deployed will utilize a set of features that can be examined.

[70] For a summary, *see* Sam Corbett-Davies & Sharad Goel, The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning (Aug. 14, 2018) (unpublished manuscript) (available at https://arxiv.org/pdf/1808.00023.pdf). In particular, a common approach to algorithmic fairness within computer science is to evaluate the fairness of a predictive algorithm using a "confusion matrix." *Id*. at 4. A confusion matrix is a cross-tabulation of actual outcomes by the predicted outcome. For instance, the confusion matrix for an algorithm that classified individuals as likely to default on a loan would appear as follows:

	Default Predicted	**No Default Predicted**
Default Occurs	# Correctly Classified as Defaulting = N_{TP} (True Positives)	# Incorrectly Classified as Non-Defaulting = N_{FN} (False Negatives)
Default Does Not Occur	# Incorrectly Classified as Defaulting = N_{FP} (False Positives)	# Correctly Classified as Non-Defaulting = N_{TN} (True Negatives)

Using this table, one could then evaluate the fairness of the classifier by inquiring whether the classification error is equal across members of protected and unprotected groups. *Id*. at 5. For example, one could use as a baseline fairness criterion a requirement that the classifier have the same false positive rate (i.e., $N_{FP} / (N_{FP} + N_{TN})$) for minority borrowers as for non-minority borrowers. Alternatively, one could use as a baseline a requirement of treatment equality (e.g., the ratio of False Positives to False Negatives) across members of protected and unprotected groups. As noted in the text, given a stated fairness criterion, an algorithm can then be tuned to achieve it.

[71] *See* Gillis & Spiess, *supra* note 67, at 480 ("In the case of machine learning, we argue that outcome analysis becomes central to the application of antidiscrimination law.").

[72] A related line of research addresses disparities arising from redundant encodings by including a protected classification as an input variable when calibrating a predictive model. *See* Devin G. Pope &

We part ways with these approaches for three reasons. First, as noted above, our reading of the Civil Rights Act of 1964 and 1968 and the subsequent caselaw and codification informs us that an input-based approach is required under the burden shifting framework that has long-informed the policing of unintentional discrimination. Second, we do not view as insurmountable the challenge of articulating the conditions for excluding variables that are correlated with a protected classification, as we illustrate in Part III. Third, it is likely that "tuning" techniques are themselves problematic with respect to discrimination law.

Outcome-based approaches would almost certainly be deemed legally problematic following the Supreme Court's 2009 decision in *Ricci v. DeStefano*.[73] The facts giving rise to *Ricci* involved a decision by the city of New Haven to discard the results of an "objective examination" that sought to identify the most qualified city firefighters for promotion.[74] The city justified its decision to discard the results on the basis that they revealed a statistical racial disparity, raising the risk of disparate impact liability under Title VII.[75] A group of white and Hispanic firefighters sued, alleging that the city's discarding of the test results constituted race-based disparate treatment.[76] In upholding their claim, the Court emphasized the extensive efforts that the city took to ensure the test was job-related[77] and that there was "no genuine dispute that the examinations were job-related and consistent with business necessity."[78] Nor did the city offer "a strong basis in evidence of an equally valid, less-discriminatory testing alternative."[79] Prohibiting the city from discarding the test results was therefore required to prevent the city from discriminating against "qualified candidates on the basis of their race."[80]

The Court's assumption that the promotion test identified the most qualified firefighters makes it difficult to see a legal path forward for explicit race-based adjustments of algorithmic outcomes. Assuming the algorithm properly identifies qualified individuals in a specified target, such race-based adjustments would appear to be no different from what the city of New Haven attempted to do with the promotion test results. Rather, *Ricci* underscores the

Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 206 (2011); Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, John M. Olin Center For Law, Economics, and Business Discussion Paper No. 1019 (October 2019). The rationale for doing so is to "de-bias" the redundantly-encoded variable. We address problems with this approach in Part III(C)(ii) and in the Appendix.

[73] 557 U.S. 557 (2009).

[74] *Id*. at 562.

[75] *Id*.

[76] *Id*. at 562-63.

[77] *Id*. 586-588.

[78] *Id*. at 587; see also id at 589 ("The City, moreover, turned a blind eye to evidence that supported the exams' validity.")

[79] *Id*. at 589.

[80] *Id*. at 584 ("Restricting an employer's ability to discard test results (and thereby discriminate against qualified candidates on the basis of their race) also is in keeping with Title VII's express protection of bona fide promotional examinations.")

fundamental importance of ensuring that decision-making processes do not systematically discriminate against qualified individuals because of their race—the goal of the burden shifting framework of Title VII.

Yet our objective is not to dismiss output-focused considerations of fairness. Rather, our goal is instead to emphasize that the burden-shifting framework requires separating the question of whether the inputs of an algorithmic process are biased against members of a protected group from the question of whether the outcomes of an unbiased algorithm meet some criterion of fairness.

As an example, consider again the SOFA algorithm discussed in the Introduction.[81] As noted, many hospitals have used the SOFA algorithm to allocate scarce life-saving resources during the pandemic based on a patient's dysfunction of six organ systems. The stated rational for doing so is that these medical conditions are legitimate input variables in a patient's expected long-term survival. However, SOFA-based triage algorithms have alarmed many clinicians given the adverse effect they are likely to have on communities of color due to structural inequalities that cause Black and Latinx Americans to suffer differential rates of chronic and life-shortening conditions that contribute to a disqualifying SOFA score.

However, it is far from clear that a SOFA algorithm would be problematic under the burden-shifting framework. If a patient's expected long-term survival is the business necessity, then a personal medical history input variable of, say, diabetes may well pick up business necessity and not have a residual correlation with race beyond its correlation with survival. If that is true, the SOFA-based algorithm would therefore not be biased against members of a protected group *given the algorithm's objective*. But one has to wonder how society can view this outcome as equitable in light of the fact that the same structural inequalities that put people of color at a greater risk of contracting of diabetes would (under the SOFA-algorithm) put them at a greater risk of dying from COVID-19 given higher rates of infection among Black and Latinx Americans.

The fairness of the algorithm's outcomes would thus need to be considered separately from whether the algorithm uses invalid input variables in pursuing its objective. From an anti-subordination perspective, one way to address these concerns would focus on whether the business necessity target is in fact equitable in light of the structural inequalities that contribute to Black and Latinx patients having higher SOFA scores. For instance, as a matter of health policy, a state's department of public health could simply stipulate an alternative business necessity target following consultation with members of the medical community and other stakeholders. However, even with a more equitable target, our approach highlights the continuing need to

[81] *See* text accompanying notes 36-38.

monitor the inputs used in the decision-making model to ensure they are not biased against protected groups.

Likewise, separately considering the question of whether the inputs of a decision-making process are biased from the question of whether the outcomes of an unbiased algorithm are fair can highlight the need to address structural inequalities in a more systematic fashion. Lending provides a domain where this has been especially relevant. Under the FHA, courts have routinely held that creditworthiness is an approved legitimate business necessity target. Yet the determinants of creditworthiness (e.g., income, income growth, wealth) reflect long-standing racial and economic inequalities, and the process of creating credit scores is also subject to criticisms of racial bias. Thus, even an unbiased lending rule that targeted creditworthiness would result in lending outcomes that reflect these structural inequalities. In this context, absent a change in the business necessity target, rectifying inequitable lending outcomes requires an additional intervention, such as through subsidized loan programs and other policies designed to encourage lending to low and moderate-income families. Indeed, this approach is reflected in existing U.S. housing programs such as the Federal Housing Administration mortgage program (which seeks to provide mortgages to low and moderate-income borrowers)[82] and the Community Reinvestment Act (which seeks to encourage lenders to provide loans to residents of low and moderate-income neighborhoods).[83]

Finally, separately considering a model's inputs from the fairness of its outputs recognizes that the question of fair outcomes is fundamentally a policy question that requires engagement from a diverse community of stakeholders. As Richard Berk and others have noted, efforts to make algorithmic outcomes "fair" pose the challenge that there are multiple definitions of fairness, and many of these definitions are incompatible with one another.[84] The central challenge Berk raises is that an outcome fix will often result in *some* form of residual discrimination, raising the inevitable question: *how much* discrimination should be permissible in the outcomes?[85] For this reason, determination of distributional equity is accordingly best left to context-specific policy institutions that can evaluate the relevant trade-offs in a transparent fashion and with input from diverse perspectives.

[82] *See* James H. Carr & Katrin B. Anacker, *The Complex History of the Federal Housing Administration: Building Wealth, Promoting Segregation, and Rescuing the U.S. Housing Market and The Economy*, 34 BANKING & FIN. SERVS. POL'Y REP. 10 (2015) (describing the program).

[83] *See* Keith N. Hylton, *Banks and Inner Cities: Market and Regulatory Obstacles to Development Lending*, 17 YALE J. ON REG. 197 (2000) (describing the Act).

[84] *See* Richard Berk et al., Fairness in Criminal Justice Risk Assessments: The State of the Art 33 (May 30, 2017) (unpublished manuscript) (available at https://arxiv.org/pdf/1703.09207.pdf) (arguing that "[t]here are different kinds of fairness that in practice are incompatible").

[85] *See, e.g.,* Gillis & Spiess, *supra* note 67, at 486 (advocating an outcome test in which a regulator evaluates whether lending outcomes differ by race among "similarly situated" borrowers "should include a degree of tolerance set by the regulator").

C. Input Accountability Versus “Least Discriminatory” Predictive Accuracy

We differ also from scholars and practitioners who focus only on the final step in the disparate-impact burden-shifting framework. Recall that according to this framework, an employer who establishes that a business practice can be justified by a legitimate business necessity shifts the burden back to the plaintiff to show that an equally valid and less discriminatory practice was available that the employer refused to use.[86] Some commentators have mistakenly assumed that this test implies that the critical question to ask when evaluating an algorithm that produces a disparate impact is whether the algorithm uses the least discriminatory predictive model for a given level of predictive accuracy. Of course, for a data scientist with access to thousands of variables, it is easy to run many models and decide which creates the least disparate impact for a given level of accuracy in prediction. But this approach will not address whether any of the variables used in the model are systematically penalizing members of a protected group who are otherwise qualified in the skill or characteristic the model is seeking to predict.

Nonetheless, a number of commentators have, mistakenly we believe, argued that the central test for whether an algorithm poses any risk of illegitimate discrimination should be whether there are alternative models that can achieve the same level of predictive accuracy with lower levels of discrimination.[87] For instance, in an oft-cited discussion paper regarding fair lending risk of credit cards, David Skanderson and Dubravka Ritter advocate that lenders should focus on this step of the disparate-impact framework when evaluating the fair-lending risk of algorithmic credit-card models.[88] Specifically, Skanderson and Ritter note that “a model or a model’s predictive variable with a disproportionate adverse impact on a prohibited basis may still be legally permissible if it has a demonstrable business justification and there are no alternative variables that are equally predictive and have less of an adverse impact.”[89] For Skanderson and Ritter, the business necessity defense for an algorithmic decision-making process therefore boils down to whether it is the most accurate possible test in predicting a legitimate target

[86] *See* text accompanying note 52.

[87] *See, e.g.,* Nicholas Schmidt & Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination, (Nov. 8, 2019) (unpublished manuscript) (available at https://arxiv.org/pdf/1911.05755.pdf) (advocating for using “a ‘baseline model’ that has been built without consideration of protected class status, but which shows disparate impact, and then search[ing] for alternative models that are less discriminatory than that baseline model, yet similarly predictive.”).

[88] *See, e.g.,* David Skanderson & Dubravka Ritter, *Fair Lending Analysis of Credit Cards* (FED. RESERVE BANK OF PHILA., Discussion Paper No. 14-02, 2014), https://www.philadelphiafed.org/-/media/consumer-credit-and-payments/payment-cards-center/publications/discussion-papers/2014/d-2014-fair-lending.pdf?la=en.

[89] *Id*. at 38.

variable of interest. As they summarize in the context of lending, "If a scoring system is, in fact, designed to use the most predictive combination of available credit factors, then it should be unlikely that someone could demonstrate that there is an equally effective alternative available, which the lender has failed to adopt."[90]

However, validating an algorithm based entirely on the fact that it is the most predictive model available would validate algorithms that are clearly biased against members of a protected group who are qualified in the desired target. To illustrate, we offer a simple example. Consider an employer who needs employees that can regularly lift 40 pounds as part of their everyday jobs. Imagine this employer designs a one-time test of whether applicants can lift 70 pounds as a proxy for whether the applicant can repetitively lift 40 pounds. The employer can show that this test has 90% prediction accuracy. However, those applicants that fail the test who in fact could regularly lift 40 pounds are disproportionately female. Thus, the test, because it is not a perfect proxy, causes a disparate impact on female applicants.

Now assume that it can be shown that a one-time test of whether applicants can lift 50 pounds produces no disparate impact on females but has an accuracy rate of just 85%. Under Skanderson and Ritter's approach, the employer would have no obligation to consider the latter test, despite the fact that a 70-pound test will systematically penalize female applicants that can in fact satisfy the job requirement.

Not surprisingly, this approach to pre-screening employment tests has been routinely rejected by courts. In *Lanning v. Southeastern Pennsylvania Transportation Authority*,[91] for instance, the Third Circuit considered a physical fitness test for applicants applying to be transit police officers. The fitness test involved a 1.5 mile run that an applicant was required to complete within 12 minutes; however, the 12 minute cut-off was shown to have a disparate impact on female applicants.[92] The transit authority acknowledged that officers would not actually be required to run 1.5 miles within 12 minutes in the course of their duties, but it nevertheless adopted the 12 minute cut-off because the transit authority's expert believed it would be a more "accurate

[90] *Id*. at 43. This line of reasoning also informs Barocas and Selbst's conclusion that Title VII provides a largely ineffective means to police unintentional discrimination arising from algorithms. *See* Barocas & Selbst, *supra* note 62, at 701-714. According to Barocas and Selbst, the business necessity defense requires that an algorithm is "predictive of future employment outcomes." *Id.* If this is correct, it would logically follow that an employer will have no disparate impact liability from using the most predictive algorithmic model for a legitimate job-related skill since an equally predictive, less discriminatory alternative would not be available. However, this conclusion relies on an assumption that predictive accuracy is a necessary and sufficient condition to justify a decision-making process that produces a disparate impact. As we show, this is an incorrect assumption, as courts have been careful not to conflate the business necessity defense with predictive accuracy. A predictive model may be accurate in predicting whether an individual is likely to have a legitimate target characteristic but nevertheless be biased against members of a protected group who are otherwise qualified in the target characteristic.

[91] 181 F.3d 478 (3rd Cir. 1999), *cert. denied*, 528 U.S. 1131 (2000).

[92] *Id*. at 482.

measure of the aerobic capacity necessary to perform the job of a transit police officer."[93]

In considering the transit authority's business-necessity defense, the court agreed that aerobic capacity was related to the job of a transit officer.[94] It also agreed that by imposing a 12 minute cut-off for the run, the transit authority would be increasing the probability that a job applicant would possess high aerobic capacity.[95] Nonetheless, the court rejected this "more is better" approach to setting the cutoff time:

> Under the District Court's understanding of business necessity, which requires only that a cutoff score be "readily justifiable," [the transit authority], as well as any other employer whose jobs entail any level of physical capability, could employ an unnecessarily high cutoff score on its physical abilities entrance exam in an effort to exclude virtually all women by justifying this facially neutral yet discriminatory practice on the theory that more is better.[96]

Accordingly, the court required "that a discriminatory cutoff score be shown to measure the minimum qualifications necessary for successful performance of the job in question in order to survive a disparate impact challenge."[97] In other words, in determining whether disparate outcomes are justified, the question to ask is not simply whether the model is accurate in predicting the target variable, but whether the choice of the process and inputs met the business necessity burden.[98]

The *Lanning* case focused on predictive accuracy in determining the minimum cutoff score for qualification, and its holding highlights why an employer cannot claim that it is a business necessity to use a high cutoff score simply because it is the most accurate score for finding qualified applicants.[99]

[93] *Id*.

[94] *Id*. at 492.

[95] *Id*. ("The general import of these studies is that the higher an officer's aerobic capacity, the better the officer is able to perform the job.").

[96] *Id*. at 493.

[97] *Id*.

[98] *See Lanning*, 181 F.3d. 478, 481 (3rd Cir. 1999) ("[U]nder the Civil Rights Act of 1991, a discriminatory cutoff score on an entry level employment examination must be shown to measure the minimum qualifications necessary for successful performance of the job in question in order to survive a disparate impact challenge."); *see also* Ass'n of Mex.-Am. Educators v. California, 195 F.3d 465 (9th Cir. 1999) (upholding, against a disparate-impact challenge under Title VII, a requirement that public school teachers "demonstrate basic reading, writing and mathematics skills in the English language as measured by a basic skills proficiency test" and holding as not clearly erroneous the district court's finding that the cutoff scores "reflect[ed] reasonable judgments about the minimum levels of basic skills competence that should be required of teachers.").

[99] The Third Circuit was clear that setting the cutoff was effectively about calibrating the predictive accuracy of the employment test. *See Lanning*, 308 F.3d 286, 292 (3rd Cir. 2002) ("It would clearly be unreasonable to require SEPTA applicants to score so highly on the run test that their predicted rate of

This same reasoning also applies to the selection of variables one uses in a predictive model.

For example, in the recent past, credit decisions were made primarily on application data plus credit history reports and any "soft information" a loan officer could glean from interacting with a borrower. For simplicity, imagine that all of these items only translated into 10 variables and that these variables predict default with predictive accuracy of 85%. With the advent of big data and machine learning, lenders now regularly use thousands of variables to assess an applicant's default probability. Imagine that multiple machine learning algorithms, after using thousands of variables, can now predict default with 90% accuracy. Assume further that all of these algorithms produce a greater disparate impact than the conventional 10-variable model, but one can nevertheless find the least discriminatory of these machine learning algorithms. Does the fact that the least discriminatory algorithm has the highest predictive accuracy justify the additional disparate impact caused by moving from 10 to 1,000 variables? Under the burden shifting framework, legitimate business necessity has not been established. Perhaps the machine learning model is indeed consistent with business necessity, but just because it is more accurate does not establish this fact.

Indeed, this latter example speaks directly to a controversial rule proposed in 2019 by the Department of Housing and Urban Development (HUD).[100] Given the increasing role of algorithmic credit scoring, the proposed rule-making expressly provides for a new defense for disparate impact claims under the FHA where "a plaintiff alleges that the cause of a discriminatory effect is a model used by the defendant, such as a risk-assessment algorithm…."[101] In particular, the proposed rule provides that in these cases, a lender may defeat the claim by "identifying the inputs used in the model and showing that these inputs are not substitutes for a protected characteristic and that the model is predictive of risk or other valid objective."[102] In other words, so long as a variable is not an undefined "substitute" for a protected characteristic, any model that predicts creditworthiness is sufficient to defeat a claim of disparate impact discrimination.

This approach to algorithmic accountability, however, suffers from the same defect noted previously with regard to those who have misapplied the "least discriminatory alternative" test. Specifically, by focusing solely on

success be 100%. It is perfectly reasonable, however, to demand a chance of success that is better than 5% to 20%.").

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

[101] *Id*. at 42,862. The rulemaking was intended to amend HUD's interpretation of the disparate impact standard "to better reflect" the Supreme Court's 2015 ruling in *Texas Department of Housing & Community Affairs v. Inclusive Communities Project, Inc*., 135 S. Ct. 2507 (2015), which upheld the ability of plaintiffs to bring disparate impact cases of discrimination under the FHA.

[102] *Id*.

whether a model is "predictive of risk or other valid objective," HUD's test leaves open the possibility that a lender can adopt a model that systematically discriminates against borrowers who are, in fact, creditworthy. Recall that in our hypothetical strength test, the ability to lift 70 pounds was, in fact, predictive of whether an applicant could regularly lift 40 pounds; however, it systematically discriminated against women who were qualified for the job. Worse still, by not even requiring that a model have any particular level of accuracy, HUD's test would seemingly permit the use of any proxy so long as it has *some* correlation with credit risk. Indeed, this approach would even appear to permit the use of explicit redlining in a predictive model so long as a lender could show that the average credit risk of a majority-minority neighborhood is marginally higher than that of non-majority-minority neighborhoods.

In contrast, the central goal of the burden shifting framework is to ensure that in evaluating a decision-making process, members of a protected class are not being systematically penalized despite being qualified in a target characteristic of interest.

III. THE INPUT ACCOUNTABILITY TEST

In this section, we move to the second aspect of our contribution: presenting our *input accountability test* (IAT) to test for discrimination under Title VII. In Part IV, we examine how the IAT can be extended to applications outside of the Title VII context.

We begin with some nomenclature. The design of a decision-making algorithm rests fundamentally on the relationships between a set of input variables, sometimes referred to as "features" in the context of machine learning, and an underlying latent skill or attribute of interest (strength, productivity, etc.), referred to as a "target." Valid target variables all fall under a legitimate business necessity fundamental model. This fundamental model can be a formal structural relationship, as is possible in life cycle modelling of credit risk, or, more likely, is a nonparametric combination of these target variables (i.e., the required job skills are a function of intelligence, reliability and strength). Today, the relationships between targets and features are increasingly analyzed and developed within artificial-intelligence and machine-learning processes, but an algorithmic decision-making process can also be based on human-selected data or even on personal intuition. The IAT applies to a decision-making algorithm regardless of whether the features (i.e., input variables) are determined through machine learning or human learning.

Our second contribution is a test that informs when an input variable's use produces statistical discrimination against a protected class that is unjustified according to the criteria developed in Part II. That is, the IAT detects if the use of an input results in systematically penalizing members of a protected group beyond the role of the input variable in extracting the business necessity goal.

A. The Test

We illustrate our test throughout with the facts giving rise to the 1977 Supreme Court decision in *Dothard v. Rawlinson*.[103] As noted previously, in *Dothard*, female applicants for prison officer positions challenged a prison's minimum height and weight requirements as inconsistent with Title VII.[104] Because the average height and weight of females were less than those for males, the female applicants argued that the requirement created an impermissible disparate impact for females under Title VII.[105] In response, the prison argued that a height and weight requirement was a justified job requirement given that an individual's height and weight are predictive of strength, and strength was required for prison officers to perform their jobs safely.[106] In short, the prison took the position that the general correlation between one's height/weight and strength was sufficient to justify the disparate outcomes this requirement caused for women. The Supreme Court, however, rejected this defense.[107] Rather, to justify gender differences in hiring outcomes, the prison would need to show that it had tested for the *specific type of strength* required for effective job performance; [108] in other words, the prison would have to be concerned with the aspects of strength that the proxy variables were and were not picking up that related to a prison officer's need for strength.

We use this setup and some hypothetical applicants to lay out the IAT. Imagine for example that twelve individuals apply for an open prison officer position, of which six applicants are male and six are female. In evaluating the applicants, the prison seeks to select those applicants who possess the actual strength required for successful job performance. For simplicity, assume that an individual's strength can be measured on a scale of zero to one hundred, and that a strength score of at least sixty is a true target for job effectiveness (i.e., a strength of sixty is a legitimate-business-necessity criterion). The challenge the prison faces in evaluating job applicants is that each applicant's actual strength is unobservable at the time of hiring, thus inducing the prison to rely on height as a proxy.

Assume that the use of a minimum height requirement results in the following distribution of applicants according to their actual but unobservable strength (Figure 1).

[103] 433 U.S. 321 (1977).

[104] *Id*. at 323-24.

[105] *Id*.

[106] *Id*. 331.

[107] *Id*. at 332.

[108] *Id*. at 332 ("If the job-related quality that the appellants identify is bona fide, their purpose could be achieved by adopting and validating a test for applicants that measures strength directly.").

Figure 1

	Results with Height Test		
Actual Strength	**Meets Height Requirement**	**Fails Height Requirement**	
100	x		
90	x		*Minimum*
80	x		*Required*
70	x	x	*Strength*
60		x	↓
50	x		
40	x	x	
30		x	
20		x	
10		x	
0			

x = applicant

Consistent with the prison's argument, there is a clear correlation between an applicant's height and actual strength. However, when we examine the gender of the applicants, we discover that only the six male applicants satisfy the minimum height requirement (Figure 2).

Figure 2

	Results with Height Test		
Applicant's Strength	**Meets Height Requirement**	**Fails Height Requirement**	
100	□		
90	□		*Minimum*
80	□		*Required*
70	□	●	*Strength*
60		●	↓
50	□		
40	□	●	
30		●	
20		●	
10		●	
0			

□ = male; ● = female

In this situation, a basic correlation test between height and strength has produced exactly the injury of concern noted in Part II(A): The imperfect relationship between height and strength results in penalizing otherwise qualified female applicants and benefiting unqualified male applicants. This can be seen from the fact that the only applicants who possessed sufficient strength but failed the height test were female. Likewise, the only applicants who met the height test but lacked sufficient strength were male. The screening test is thus systematically biased against female applicants for reasons unrelated to a legitimate business necessity.

This example points to the crux of the IAT. In general, the objective of the test is to ensure that a proxy variable is excluded from use if the imperfect relationship between the proxy variable and the target of interest results in systematically penalizing members of a protected group that are otherwise qualified in the target of interest. In other words, since the proxy variable (height) is not a perfect predictor of having the target strength, there is some residual or unexplained variation in height across applicants that is unrelated to whether one has the required strength. The question is whether that residual is correlated with gender. In Figure 2, it is.

To avoid this result in *Dothard*, the Supreme Court therefore required a better proxy for required strength. In particular, the prison would be required to "adopt[] and valida[te] a test for applicants that measures strength directly" in order to justify disparities in hiring outcomes.[109] For example, assume that the prison implemented as part of the job application a physical examination that accurately assessed required strength, which produced the following results (Figure 3).

[109] *Id*. at 332.

Figure 3

	Results with Perfect Strength Exam		
Actual Strength	**Passes Exam**	**Fails Exam**	
100	□		
90	□		*Minimum*
80	□		*Required*
70	□ ●		*Strength*
60	●		↓
50		□	
40		□ ●	
30		●	
20		●	
10		●	
0			

□ = male; ● = female

The examination was perfect in classifying all individuals – male and female – as qualified if they in fact were so. Note that, even under this perfect exam, more males than females would be deemed eligible for the position. This disparity, however, arises solely through differences in actual strength (a legitimate business necessity).

Figure 3 is an ideal outcome in the sense that the prison was perfect in measuring each applicant's actual strength, but perfect proxy variables are rarely available. Imagine instead that the prison asks the applicants to complete a simple muscle-mass index assessment (Figure 4).[110]

[110] For instance, imagine the prison assesses each applicant's mid-arm muscle circumference (MAMC) and requires a minimum measure which the prison believes is associated with having a minimum strength of 60. The MAMC is one of several techniques to measure muscle mass. *See generally* Julie Mareschal et al., *Clinical Value of Muscle Mass Assessment in Clinical Conditions Associated with Malnutrition*, 8 J. CLINICAL MED. 1040 (2019).

Figure 4

	Results with Muscle Mass Exam		
Actual Strength	**Meets Muscle Mass Requirement**	**Fails Muscle Mass Requirement**	
100	□		
90	□		*Minimum*
80	□		*Required*
70	●	□	*Strength*
60		●	↓
50	□		
40	●	□	
30		●	
20		●	
10		●	
0			

□ = male; ● = female

As can be seen, muscle mass proxies for required strength with a positive, significant correlation, but it does so with error. In particular, there are applicants who are sufficiently strong but fail the muscle mass requirement, and there are applicants who meet the muscle mass requirement but are not sufficiently strong. The difference from Figure 2, however, is that the proxy is unbiased: Neither male applicants nor female applicants are favored by the fact that the proxy does not perfectly measure required strength. This is illustrated by the fact that one male and one female fail the muscle mass requirement but possess sufficient strength for the job, and one male and one female meet the muscle mass requirement but lack sufficient strength. Because the proxy is unbiased with respect to gender, an employer should therefore be permitted to use muscle mass as a proxy for required strength.

B. The Test in Regression Form

Moving from concepts to practice, standard regression techniques provide a straightforward means to implement the IAT. In keeping with the foregoing example, we return to the modified facts of *Dothard,* in which a prison uses a job applicant's height as a proxy for whether they have the required strength to perform the job of a prison officer.[111] The prison does so based on the assumption that required strength is manifested in an

[111] Of course, there might be multiple proxies. For instance, imagine the job requirements were strength and IQ, in some combination. Such a specification could be handled by more complex formations on the right-hand side of the regression framework that we discuss here.

individual's height. However, height is also determined by a host of other causes that are unrelated to strength. If we represent this group of non-strength determinants of height for a particular individual i as ε_i, we can summarize the relationship between the height and strength as follows:

$$Height_i = \alpha \cdot Strength_i + \varepsilon_i,$$

where α is a transformation variable mapping the relationship of strength to expected height. If ε_i was zero for each individual i, the equation becomes $Height_i = \alpha \cdot Strength_i$. In such a setting, an individual's height would be precisely equal to the individual's strength, multiplied by the scalar α. Therefore, one could compare with perfect accuracy the relative strength of two individuals simply by comparing their heights.

Where ε is non-zero, using height as a proxy for strength will naturally be less accurate; however, using height in this fashion will pose no discrimination concerns if ε (the unexplained variation in height that is unrelated to strength) is uncorrelated with a protected classification. This was precisely the case in Figure 4: Strength was *somewhat* manifested through the muscle mass index. Thus, it would be a useful variable for predicting which job applicants had the required strength for the job. Moreover, while it was error-prone in measuring actual strength (i.e., $\varepsilon_i \neq 0$), using one's muscle mass index to infer strength would pass the IAT:

$$\varepsilon_i \perp gender;$$

the errors were not statistically correlated with gender, the protected category in our example.

To implement this test empirically, the prison would use the historical data it holds concerning its existing employees' measured height and strength and regress employee height on employee strength to decompose the variation of height into that which is correlated with strength and that which is unexplained. This process would estimate $\hat{\alpha} \cdot Strength_i$, where the $\hat{\alpha}$ is the estimated regression coefficient. Using this estimated relationship between strength and height, the difference between an employee's actual height and predicted height would constitute an estimated residual (ε_i) for each employee, or the portion of height unexplained by strength.[112] This decomposition thus takes out the linear correlation of the input variable with the target. One could equally do this decomposition on other transforms of the input variable (e.g., squared, natural logarithm, or non-parametric interval variables). Using these residuals, the prison would then examine whether they are correlated with employee gender.

[112] The regression will also estimate a constant term that is utilized in calculating the relationship between strength and height.

How would the IAT be used in a setting where the proxy is not a continuous measure (such as one's height or muscle mass) but rather a binary outcome of whether an individual possesses a specified level of the measure? Recall that this was the case in our hiring example where the prison first assessed an applicant's height and then applied a cut-off score to eliminate from consideration those applicants who fell below it. As reflected in *Dothard* and *Lanning*, applying a minimum cut-off score to a proxy variable is a common decision-making practice, including within machine learning.[113]

The application of the IAT would use the same framework as above, but it would use as the left-hand-side variable an indicator variable for whether an individual i was above or below the cutoff—for our example, $Height_i = 1$ for applicants above the cutoff and $Height_i = 0$ for applicants below it. To estimate a discrete 0-1 variable (*Height*) as a function of a target (e.g., *Strength*), the preferred model is a logistic estimation (or a comparable model for use with a dichotomous outcome variable). Logistic estimation is a transformation that takes a set of zeros and ones representing an indicator variable and specifies them in terms of the logarithm of the odds ratio of an outcome (in our example, the odds ratio is the probability of $Height_i$ being above the cut-off divided by the probability that it is below the cut-off). This transformation is then regressed on the target measure (*Strength*). To generate the residuals, one predicts the probability of a positive outcome and then generates the residual as the true outcome minus the predicted probability. As above, to pass the test, the residuals should not be significantly correlated with gender.

We have thus far assumed a simple model of business necessity based on one target, strength. Yet, suppose the skills necessary for a prison officer include intelligence, reliability, and diligence. The fundamental business necessity model would then have multiple targets. In many contexts, the multiple targets are related. For example, muscle mass may not just pick up strength but also some aspects of diligence, as it takes grit to persevere at the gym regularly. A lender, for example, may choose a number of input variables (signals on family and social networks) that could be used to pick up missing aspects of wealth and expected income growth. When the targets are more than one, the application of the IAT would include all target variables on the right-hand side of the estimation, again using historical training datasets.

[113] *See, e.g.*, Elizabeth A.Freeman & Gretchen G.Moisen, *A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa*, 217 ECOLOGICAL MODELING 48 (2008) (reviewing criteria for establishing cutoffs in ecological forecasting).

C. Consequence of Failing the IAT

When an input variable fails the IAT, its use is inconsistent with Title VII. Thus, the variable should be excluded. This is a stark statement, and we do not take its assertion lightly.

i. Concerns with Excluding IAT-Failing Variables

The first concern with excluding an input variable that fails the IAT might be with the empirical reliability of the IAT in detecting input variables that are inconsistent with Title VII. We take this concern seriously, and we devote Part V to discussing empirical challenges in implementing the IAT. In particular, Part V discusses issues arising from (i) the unobservability of a target variable, (ii) measurement error in a target variable, and (iii) testing for a statistically uncorrelated relationship between the protected category and the residual (from decomposing the input variable into that which is correlated with the target(s) and that which is unexplained). We follow this discussion in Part VI with a simulation based on *Dothard* to show how the IAT can be implemented in general and in handling these challenges.

The second concern is that it might be possible to de-bias the input variables rather than dropping them. We address this concern in the following subsection and in the Appendix.

ii. The Possibility of De-biasing IAT-Failing Variables

If the residuals are correlated with a protected classification (e.g., gender), it may be possible to "de-bias" a model that predicts strength from height, most notably by adding an individual's membership (or lack of membership) in a protected class as an input in the predictive model. Indeed, this approach to de-biasing proxy input variables has been advanced by several scholars.[114]

However, as shown in the Appendix, the fact that de-biasing requires us to include one's membership in a protected group (e.g., an indicator variable for whether an applicant is female or not) in the predictive model impairs the utility of this approach. A predictive model that explicitly scores individuals differently according to gender constitutes disparate treatment, making it a legally impermissible means to evaluate individuals. To avoid this challenge, proponents of this approach have therefore advocated that, in making predictions, the model should assign all individuals to the mean of the protected classification;[115] in our example, one would do so by treating all

[114] *See* Devin G. Pope & Justin R. Sydnor, *Implementing Anti-Discrimination Policies in Statistical Profiling Models*, 3 AM. ECON. J. 206, 206 (2011); Crystal Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, John M. Olin Center For Law, Economics, and Business Discussion Paper No. 1019 (October 2019). We provide an example of this approach, as well as its limitations, in the Appendix.

[115] *See*, *e.g*., Pope & Sydnor, supra note 114, at 212.

individuals as if gender = 0.5 (i.e., (1 + 0) / 2) when estimating the effect of gender on predicted strength. Doing so introduces prediction error, however, and as demonstrated by Kristen Altenburger and Daniel Ho, this error can be especially problematic when the approach is deployed in common machine-learning models.[116] More troublesome, these prediction errors can themselves be systematically biased against members of a protected group who are otherwise qualified in the target. We illustrate this challenge in the Appendix, which presents a simple example showing that this "de-biasing" procedure may actually have almost no effect on the extent of bias in the final outcome.

These considerations reinforce our conclusion that a decision-making model should exclude any variable that fails our test. While this approach risks sacrificing some degree of predictive accuracy in favor of an unbiased decision-making process, our discussion in Part II(C) illustrates that U.S. antidiscrimination law has long made this trade-off. Additionally, a rule of exclusion also creates obvious incentives to seek out observable variables that can more accurately capture the target variable of interest, consistent with the holding of *Dothard* that the prison should adopt a test that more directly measured an applicant's strength.[117] Indeed, creating this incentive seems especially appropriate in the machine learning context given the capacity of machine learning processes to analyze an ever-increasing volume of data to identify proxies that can pass the IAT.

IV. APPLICATIONS

The fact that the IAT is rooted in general antidiscrimination principles makes it applicable to any setting where a decision-maker relies on statistical discrimination, regardless of whether conducted by humans or algorithms. Central to our argument is the idea of using a test to ascertain adherence to business necessity targets when designing a decision-making process.

In this section, we discuss additional implementations outside of the employment setting to illustrate the general applicability of the IAT whenever a court or other regulatory body has articulated business necessity targets that can justify disparate outcomes across members of protected and unprotected groups. In Part 4(A), we begin by examining settings where courts have expressly engaged in this process and defined legitimate business necessity targets under various U.S. antidiscrimination laws. In these domains, application of the IAT simply requires testing an algorithm's features against the specified business necessity target concept(s).

In Part 4(B) we turn to other domains where no legally imposed business necessity target currently exists for applying the IAT. These are domains

[116] *See* Kristen M. Altenburger & Daniel Ho, *When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions*, 175 J. INSTITUTIONAL & THEORETICAL ECON. 98, 109-18 (2018).

[117] *See supra* note 108.

where formal liability for claims of disparate impact or other claims of unintentional discrimination are currently less clear, absolving courts or policy-makers from having to define business necessity targets. Firms operating in these settings are, of course, free to self-regulate by applying the IAT to their own self-imposed business necessity targets. However, for those concerned about algorithmic discrimination in these domains, Part 4(B) underscores the special need for algorithmic accountability legislation in these contexts. To the extent legislation occurs, the IAT will provide a ready means to ensure algorithms are accountable so long as the legislation clearly specifies a business necessity target.

Finally, in Part 4(C) we lay out the case for the fundamental importance of properly defining a business necessity target for both policy-makers and firms.

A. Domains with Court-Defined Business Necessity Targets

Consider a regulator tasked with evaluating a decision-making algorithm in one of the following domains where legal claims of unintentional discrimination are recognized, and where courts have expressly defined a legitimate target attribute that can justify unintended disparities that vary across protected and unprotected groups:

Table 1

Domain:	**Legitimate Target Attribute:**
Credit determinations	Creditworthiness[118]
Home insurance pricing	Risk of loss[119]
Parole determinations	Threat to public safety[120]

[118] *See* A.B. & S. Auto Serv., Inc. v. S. Shore Bank of Chi., 962 F. Supp. 1056 (N.D. Ill. 1997) ("[In a disparate impact claim under the ECOA], once the plaintiff has made the prima facie case, the defendant-lender must demonstrate that any policy, procedure, or practice has a manifest relationship to the creditworthiness of the applicant…"); *see also* Lewis v. ACB Bus. Servs, Inc., 135 F.3d 389, 406 (6th Cir. 1998) ("The [ECOA] was only intended to prohibit credit determinations based on 'characteristics unrelated to creditworthiness.'"); Miller v. Countrywide Bank, NA, 571 F. Supp. 2d 251, 258 (D. Mass. 2008) (rejecting the argument that discrimination in loan terms among African-American and white borrowers was justified as the result of competitive "market forces," noting that prior courts had rejected the "market forces" argument insofar that it would allow the pricing of consumer loans to be "based on subjective criteria beyond creditworthiness").

[119] *See, e.g.,* Owens v. Nationwide Mut. Ins. Co., No. Civ. 3:03-CV-1184-H, 2005 WL 1837959, at *9 (N.D. Tex. Aug. 2, 2005) (holding that minimizing the "risk of loss in homeowner's insurance" was a legitimate business necessity under the FHA that justified the use of facially neutral policy of using credit to determine eligibility for homeowner's insurance).

[120] *See, e.g.,* Cal. Penal Code § 3041 (West 2018) ("The [Board of Parol Hearings] shall grant parole to an inmate unless it determines that the gravity of the current convicted offense or offenses, or the timing and gravity of current or past convicted offense or offenses, is such that consideration of the public safety requires a more lengthy period of incarceration for this individual."); *see also* Smith v. Sisto, No. CV-08-00104CBMHCX, 2009 WL 3294860, at *6 (E.D. Cal. Oct. 13, 2009) (denying a claim that denial of parole

Tenant selection	Ability to meet lease obligations,[121] pay rent,[122] and resident safety[123]
Post-secondary school admission	Predicted academic success[124]
Selection into special education	Educational ability[125]
State merit scholarship eligibility	Academic achievement in high school[126]

Just as employers are permitted to make hiring decisions based on the legitimate target variables capturing a job-required skill, courts in these settings have likewise determined that decision-making outcomes can lawfully vary across protected and unprotected groups only if decisions are based on the target variables noted in Table 1.

In applying the IAT in these settings, the regulator's task thus follows the same process noted in Part III. First, the regulator must evaluate whether the decision-making process does, in fact, seek to produce outcomes based on the legitimate target attributes. Second, using historical data for both the target variables and the model's full set of input features, the regulator would then apply the IAT to each input variable used in the model. Finally, any input variable that failed the test would be required to be excluded from the model.

B. Domains Without Court-Defined Business Necessity Targets

What about domains where there are no legally imposed business necessity targets? There are two reasons why this might be the case. First,

constituted discrimination and concluding that "[t]he need to ensure public safety provides the rational basis for section 3041").

[121] *See* 24 C.F.R. § 100.202(c)(1) (2020) (permitting under the FHA a landlord's "[i]nquiry into an applicant's ability to meet the requirements of ownership or tenancy").

[122] *See* Ryan v. Ramsey, 936 F. Supp. 417 (S.D.Texas 1996) (noting that under the FHA, "there is no requirement that welfare recipients, or any other individuals, secure apartments without regard to their ability to pay.").

[123] *See* Evans v. UDR, Inc., 644 F. Supp. 2d 675, 683 (2009) ("The policy against renting to individuals with criminal histories is . . . based [on] concerns for the safety of other residents of the apartment complex and their property").

[124] *See* Kamps v. Baylor Univ., 592 F. App'x 282 (5th Cir. 2014) (rejecting an age discrimination case based on law school admissions criteria that relied on an applicant's grade point average (GPA) because GPA is a quantitative predictor of academic success in law school and thus a "a reasonable factor other than age").

[125] *See* Ga. State Conf. of Branches of NAACP v. Georgia, 775 F.2d 1403, 1420 (11th Cir. 1985) (finding, in a Title VI case alleging that school district achievement grouping caused disparate impact on minority students, that the school district's effort to classify students based on assessment of ability was justified because it bore "a manifest demonstrable relationship to classroom education").

[126] *See* Sharif by Salahuddin v. N.Y. State Educ. Dept., 709 F. Supp. 345, 362 (S.D.N.Y. 1989) (finding that the state's use of SAT scores did not have a "manifest relationship … [to] recognition and award of academic achievement in high school" in a Title IX claim of disparate impact alleging that the state's use of SAT scores to determine student eligibility for merit scholarships had a discriminatory effect on women).

antidiscrimination laws may not formally regulate decision-making processes that result in unintended disparities across protected and unprotected groups. Second, the legal risk for unintentional discrimination may presently be unclear. We provide an example of each.

With respect to the first situation, insurance outside the context of home insurance provides one such example.[127] As Ronen Avraham, Kyle Logue, and Daniel Schwarcz show, a number of jurisdictions do not have any laws restricting providers of automobile or life insurance from discriminating on the basis of race, national origin, or religion.[128] Nor is there a federal antidiscrimination statute applicable to insurance outside of the context of home insurance.[129] Consequently, insurers likely have considerable discretion to rely on statistical discrimination to underwrite policies, which may produce unintended disparities across protected and unprotected groups.

With respect to the second situation, an example can be found in the provision of medical treatment, which is relevant given our examples in the Introduction of the triage algorithms in the context of COVID-19 and UnitedHealth's patient illness algorithm. Discrimination in healthcare provision is covered by Title VI of the Civil Rights Act of 1964, thus making it a more regulated setting than the insurance example. However, in *Alexander v. Sandoval*,[130] the U.S. Supreme Court held that Title VI does not provide for a private right of action to enforce disparate impact claims, greatly diminishing the risk that a provider of healthcare will face a claim of unintentional discrimination. This has also meant there has not been an occasion for courts to articulate a business necessity target.

Even in these domains, however, the IAT remains a relevant tool for policing discrimination for two reasons. First, despite the lack of a clear cause of action for unintentional discrimination, these two domains often constitute areas of vital importance to the health and welfare of individuals, creating strong incentives for members of the public to scrutinize whether the transition to algorithmic decision-making is adversely impacting members of protected groups. Indeed, it is precisely this concern that led to the public scrutiny of SOFA-based triage algorithms during the COVID-19 pandemic. Similar public scrutiny has been applied to racial disparities in the pricing of auto loans. For instance, a nationwide study by the Consumer Federation of America (CFA) in 2015 found that predominantly African-American

[127] As noted in Table 1, the FHA governs discrimination in home insurance.

[128] *See* Ronen Avraham et al., *Understanding Insurance Anti-Discrimination Laws*, 87 S. CAL. L. REV. 195, 239 (2014).

[129] *Id.* at 241. Additionally, the few cases alleging discrimination by insurance providers under 42 U.S.C. § 1981—a Reconstruction era statute that prohibits racial discrimination in private contracting—have required a showing of intentional discrimination. *See, e.g.,* Amos v. Geico Corp., No. 06-CV-1281(PJS/RLE), 2008 WL 4425370 (D. Minn. Sept. 24, 2008) ("To prevail under § 1981, plaintiffs must prove that GEICO intentionally discriminated against them on the basis of race.").

[130] 532 U.S. 275 (2001).

neighborhoods pay higher auto premiums,[131] calling into question the discriminatory impact of insurance pricing models.

To the extent public scrutiny induces firms and organizations to self-regulate, the IAT provides a means to examine whether their decision-making models are producing unintentional, illegitimate discrimination. For instance, in response to the CFA's study, the Property Casualty Insurers Association of America responded with a declaration that "Insurance rates are color-blind and solely based on risk."[132] To the extent insurers are sincere in this claim, the IAT provides them with a ready test to ensure compliance with this self-imposed business necessity target.

Additionally, we believe that the lack of a clear cause of action for unintentional discrimination makes these domains especially vulnerable to the concerns about discrimination that have motivated the emergence of algorithmic accountability bills. For legislatures seeking to impose antidiscrimination guardrails on algorithmic decision-making in these areas, the IAT provides a tool to do so provided they articulate what business necessity targets can justify disparities in outcomes.

C. Determining Legitimate Business Necessity Targets

Lastly, the centrality of a business necessity target in the IAT—as in the theory of disparate impact more generally—underscores the vital importance of how this target is set by policy-makers and applied by firms. Recall that neither the IAT nor the theory of disparate impact will prevent unintentional disparate outcomes from occurring if they reflect underlying disparities in the distribution of a business necessity target. For instance, as shown in our *Dothard example*, even proper application of the IAT can result in hiring predominantly male prison officers if the distribution of strength (the business necessity target) favors males.

For policy-makers seeking to control algorithmic discrimination, this fact highlights the important equity considerations that must inform the determination of the appropriate target. Recall again the UnitedHealth example. Utilization of this algorithm was based on the objective of ascertaining the sickest patients for allocating care, a self-imposed target which presumably has business necessity. While this approach was designed to ensure that every patient had equitable access to care, a considerable amount of the criticism erupted regarding the distributional consequences of this self-imposed target.

[131] TOM FELTNER & DOUGLAS HELLER, CONSUMER FED'N OF AM., HIGH PRICE OF MANDATORY AUTO INSURANCE IN PREDOMINANTLY AFRICAN AMERICAN COMMUNITIES (2015), https://consumerfed.org/wp-content/uploads/2015/11/151118_insuranceinpredominantlyafricanamericancommunities_CFA.pdf.

[132] Press Release, Am. Prop. Casualty Insurers Ass'n, Auto Insurance Rates Are Based on Cost Drivers, Not Race (Nov. 18, 2015) (available at https://www.pciaa.net/pciwebsite/cms/content/viewpage?sitePageId=43349).

The UnitedHealth episode provides just one example of the challenging equity considerations implicated in setting a business necessity target. While resolving this challenge is beyond the scope of this article, it is critically important that it is addressed. To mitigate the risk that the distributional implications of a proposed target go overlooked, policy-makers would thus be well-advised to develop business necessity targets with input from those with expertise in distributional equity and with engagement from diverse communities. As noted in Part II, the distributional implications of setting the business necessity target is also a primary reason why we believe it is necessary to separate the question of whether an algorithm uses invalid inputs from the question of whether the outcomes from a model are fair and equitable.

Additionally, for firms engaged in algorithmic decision-making, the centrality of a business necessity target also underscores the need for businesses to be vigilant that a purported target in a decision-making model is, in fact, a legitimate one to use. This is especially the case when working in a domain where courts have defined what can (and cannot) constitute a business necessity target.

A case in point comes from the credit markets, whereby lenders may have incentives to deploy predictive algorithms to estimate demand elasticities across different borrowers to engage in price discrimination. Price discrimination is made possible by the fact that certain borrowers are more prone to accept higher-priced loans rather than engage in price shopping. (Technically, their demand is less "elastic"—that is, sensitive—to changes in price). These borrowers may not shop around for a host of reasons: They might live in financial desert locations of low competition, lack the knowledge to shop for the best rate, need to transact in a hurry, have a discomfort with financial institutions due to prior discrimination, and/or have a history of being rejected for loans. Empirical studies document that loan officers and mortgage brokers are aware of variation in borrowers' interest rate sensitivity and engage in price discrimination.[133]

A loan applicant's "price sensitivity" or "willingness to shop" may therefore be an additional unobserved characteristic that is of interest to a lender. In other words, a lender's profit margin depends on both creditworthiness (the court-determined legitimate business necessity from Table 1) and shopping profiles. A lender might therefore design an algorithm that seeks to maximize profits by uncovering credit risk and shopping profiles. Furthermore, the lender (if lending were not in a formally-regulated domain) would argue that profits are legitimate business necessity. Yet, as

[133] *See, e.g.*, SUSAN E. WOODWARD, URBAN INST., A STUDY OF CLOSING COSTS FOR FHA MORTGAGES xi (2008), https://www.huduser.org/Publications/pdf/FHA_closing_cost.pdf ("In neighborhoods where borrowers may not be so familiar with prevailing competitive terms, or may be willing to accept worse terms to avoid another application, lenders make higher-priced offers …").

noted in Table 1, lending is a domain where courts have expressly held that if a lending practice creates a disparate impact, "the defendant-lender must demonstrate that any policy, procedure, or practice has a manifest relationship to the creditworthiness of the applicant."[134] That is, while differences in creditworthiness can justify disparate outcomes in lending, differences in shopping behavior cannot.

The concern of algorithmic profiling for shopping behavior is of general concern because empirical evidence, again in lending, finds that profiling on lack-of-shopping almost certainly leads to higher loan prices for minority borrowers. For instance, Susan Woodward and Robert Hall[135] as well as Mark Cohen[136] find that adverse pricing for minority borrowers has generally been the rule when it comes to lenders who engage in price discrimination. In separate work,[137] we likewise find empirical evidence that, even after controlling for borrower credit risk, "FinTech" lenders charge minority homeowners higher interest rates. We interpret these pieces of evidence as consistent with loan originators using a form of algorithmic price discrimination. Were these algorithms subject to an internal or external "accountability audit," it is likely that the proxy variables used would fail the IAT because, no matter how well the algorithm performed in detecting the profitability of a loan, the target for the test would, by law, be creditworthiness—not an outcome that included price sensitivity. In this fashion, simply asking what target variable an algorithm seeks to detect can illuminate illegitimate algorithmic discrimination.

Finally, we want to end this applications section on a positive note. In many discussions with lenders, it has become evident that, at least in the finance realm, firms want to be able to validate what they are doing or what they intend to do before they invest and commit to a predictive algorithm. In this regard, the IAT can provide these firms with a useful tool for validating the use of proxy variables.

V. CHALLENGES IN IMPLEMENTING THE INPUT ACCOUNTABILITY TEST

Implementing the IAT faces several challenges, which we list below and then discuss in the context of the hiring test used in Part III ($Height_i = \alpha \cdot Strength_i + \varepsilon_i$), where the target variable is $Strength$.

[134] *A.B. & S. Auto Service, Inc.*, 962 F. Supp. 1056, 1056 (N.D. Ill. 1997).

[135] Susan Woodward & Robert E. Hall, *Consumer Confusion in the Mortgage Market: Evidence of Less Than a Perfectly Transparent and Competitive Market*, 100 AM. ECON. REV. 511 (2010).

[136] Mark Cohen, *Imperfect Competition in Auto Lending: Subjective Markup, Racial Disparity, and Class Action Litigation*, 8 REV. L. ECON. 21 (2012).

[137] Robert Bartlett et al., *Consumer Lending Discrimination in the FinTech Era* (Nat'l Bureau of Econ. Research Working Paper No. 25943, 2019), https://www.nber.org/papers/w25943.

A. Unobservability of the Target Variable

A first challenge in applying the IAT is the unobservability of the target variable of interest. The problem of an unobservable target is the key reason for constructing an algorithm to screen an applicant (or make some other decision), since the motivation for using statistical inference in the first place is the challenge of measuring business necessity target attributes (which are often latent) such as creditworthiness, productivity, longevity, or threat to public safety.[138]

In designing a machine-learning algorithm, this problem thus also arises in the training procedure, where a model estimates the relationship between various proxy input variables and an outcome of interest. In practice, the solution is to turn to historical data, which can be used to train the predictive model,[139] at times skipping any effort at directly measuring the target of interest. In our example, for instance, the prison may have taken muscle mass measures of strength for its prison officers at some point in the past, and it can use these data, along with other performance data (e.g., job performance assessments), and the application-reported height variable to calibrate its height cut-off model. These same data can be used for running the IAT. To be sure, the data may suffer from selection bias given that the employer will not observe performance of the applicants who were not hired. Accordingly, in both training a model and in running the IAT, one must be attendant to measurement error—a point we discuss in subsection (ii).

Nonetheless, this first challenge for the IAT—that the target is unobservable—is in many ways one of transparency. That is, data concerning the target attributes exist (after all, these data were required to train the model), but they may not necessarily be available to a regulator or researcher applying the IAT. As Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein emphasize, transparency in the training data is therefore an important step in ensuring the ability to evaluate whether algorithmic decision-making facilitates discrimination.[140] We agree. The ability to examine the training data used in designing a model would allow a regulator, litigant or researcher to conduct the IAT.

[138] *See* Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGIS. ANALYSIS 113, 132 (2019) ("One way to think about the goal of prediction is to overcome a missing information problem.").

[139] *Id.*

[140] *Id.* at 114 (arguing that harnessing the benefits of algorithmic decision-making while avoiding the risk of discrimination "will only be realized if policy changes are adopted, such as the requirement that all the components of an algorithm (including the training data) must be stored and made available for examination and experimentation").

B. Measurement Error

In addressing the unobservability problem of latent targets, one can inadvertently mis-measure it. This challenge of measurement error—or what is alternatively referred to as "label bias"[141]—has been studied in the computer science and economics literatures, providing useful guidance for addressing it when applying our test.[142]

Consider, for instance, judicial bail decisions where data scientists have used past judicial bail decisions to train algorithms to decide whether a defendant should be released on bail pending trial.[143] In many states, judges are required to consider the risk that a defendant poses for public safety, and in training the model, the business necessity target is often defined to be whether a released defendant was later arrested prior to the trial.[144] However, heavier policing in minority neighborhoods might lead to minority defendants being arrested more often than non-minorities who commit the same offense.[145] Or, said another way, perhaps the released minority defendant, who was re-arrested, was doing nothing illegal when re-arrested. Consequently, Sam Corbett-Davies and Sharad Goel have warned that this form of label bias risks causing a model to estimate a positive relationship between a defendant's race (and correlates of race) and whether the defendant poses a risk to public safety, simply due to the correlation of race with measurement error.[146]

Likewise, as Jon Kleinberg and others have noted, an employer who seeks to measure employee productivity through the number of hours that an employee spends at work will likely be using a biased measure of productivity if there are gender differences in how efficiently an employee works at the office (for example, to attend to childcare obligations before or after work).[147] Similar to the bail example, this form of label bias is problematic because the measurement error may be correlated with a protected characteristic, in this case, gender.[148]

[141] Corbett-Davies & Goel, *supra* note 70, at 3.

[142] See *id*. at 17-18.

[143] *See, e.g*., Berk et al., *supra* note 84, at 31-33.

[144] *Id*. at 31.

[145] Corbett-Davies & Goel, *supra* note 70, at 18.

[146] *Id*.

[147] *See* Kleinberg et al., *supra* note 138, at 139.

[148] Note that in both of these examples, the disparate outcomes for members of protected groups arose from the use of an estimate of a business necessity target that was mis-measured; that is, arrests are a noisy (and biased) measure for dangerousness and hours-worked is a noisy (and biased) measure for productivity. The concerns about the resulting disparities are thus different from those where the disparities arise from disparities in the target of interest. For instance, in the case of the SOFA algorithm for sorting patients for ventilator access during the COVID-19 pandemic, a health-condition such as diabetes may legitimately imply a greater mortality risk, inducing a hospital to prioritize patients that do not have diabetes if the business necessity is to prioritize patients with a greater expected long-term survival. However, the prevalence of higher rates of diabetes among African-Americans implies that they would

These examples illustrate the general point that measurement error in a target variable will create discriminatory bias when the measurement error is correlated with membership in a protected group. This result occurs because a statistical model that seeks to estimate the predictors of a true target y that is mis-measured as $y + \mu$ will inevitably discover that the protected classification (and any correlate of it) predicts the level of the mis-measured target.

For similar reasons, when measurement error in a target variable is correlated with a protected classification, application of our test may fail to detect this bias. Returning to the *Dothard* example, imagine that we applied the IAT to *Height* as before, but we use a measure for strength, *Strength**, that has measurement error μ that is correlated with gender. Formally, the test would be:

$$Height_i = \alpha \cdot Strength_i^* + \varepsilon_i$$

which is equivalent to:

$$Height_i = \alpha \cdot (Strength_i + \mu_i) + \varepsilon_i$$

In such a setting, the IAT may fail to reveal that the unexplained portion of height is correlated with the protected classification of gender. The reason is because the unexplained variation between "true" *Strength* and *Height* is $(\mu_i + \varepsilon_i)$, but the IAT will not be able to detect how gender is correlated with μ_i because it is part of *Strength**, the mis-measured target. In short, measurement error in a target variable is a critical issue to consider regardless of whether one is calibrating a model or running our test.

Recognition of this latter point is implicit in Kleinberg, Ludwig, Mullainathan, and Sunstein's argument for making training datasets transparent. Often, the data for a target will reveal fairly obvious risks that the measurement error is biased with respect to a protected classification (such as the example cited earlier when an employer uses hours-worked as a measure for productivity). At the same time, other instances when this problem arises may be less obvious. In these situations, transparency about the target proxy can nevertheless allow regulators and third-party researchers to scrutinize whether measurement error is correlated with a protected classification.

As an example, we revisit the controversy surrounding the widely-used health care algorithm deployed by UnitedHealth that we highlighted in the

get less access to ventilators under a SOFA algorithm. In this case, the target may be legitimate and measured correctly, but there may be a need for a fairness correction, as discussed in Part II(B).

Introduction.[149] UnitedHealth was transparent that it used a patient's cost of care as its proxy for the unobservable target (sickness). Using this information, researchers were subsequently able to show that this proxy for sickness had measurement error that was correlated with being an African-American patient, causing these patients to receive substandard care as compared to white patients. In particular, using actual morbidity data, these researchers showed that African-American patients historically incurred lower costs for the same illnesses and level of illness.[150] In short, transparency about the target's proxy allowed these researchers to examine how the measurement error was correlated with a protected classification, calling into question the use of this proxy for the target.

Beyond the benefits of transparency about a target, this last example also underscores the need to run the IAT with alternative measures of the target to identify mis-measured targets. This is particularly important given potential selection bias in the measure used for a target. Consider, for instance, a model that seeks to predict creditworthiness based solely on whether a borrower defaults in the training data. By construction, the training dataset consists only of those borrowers who received a loan; borrowers who do not get a loan provide no information. Thus, it is infeasible to estimate actual creditworthiness within the broader group of all applicants. This issue is often referred to as a "selective labels" problem within the computer science and economics literatures.[151] The literature on selective labels in training a model has suggested a process of interventions to correct the misestimations.[152] Another approach would be to implement the IAT through a structural estimation of theoretic representations of the target business necessity.[153]

[149] Melanie Evans & Anna Wilde Mathews, *New York Regulator Probes UnitedHealth Algorithm for Racial Bias*, WALL ST. J. (Oct. 26, 2019, 7:00 AM), https://www.wsj.com/articles/new-york-regulator-probes-unitedhealth-algorithm-for-racial-bias-11572087601.

[150] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447 (2019).

[151] *See* Himabindu Lakkaraju et al., *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*, 2017 KDD '17: PROC. 23RD ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 275; Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 256 (2018).

[152] *See, e.g.,* Maria De-Arteaga et al., *Learning Under Selective Labels in the Presence of Expert Consistency* (July 4, 2018) (unpublished manuscript) (available at https://arxiv.org/pdf/1807.00905v1.pdf) (proposing a data augmentation approach that can be used to leverage expert consistency to mitigate the partial blindness that results from selective labels).

[153] For instance, consider a credit scoring algorithm that predicts credit risk based on default rates for loans that were previously extended to a group of borrowers. A model built using these target data (i.e., whether or not a borrower defaults) suffers from bias insofar as it only includes default data for loans that were approved by a lender. This selective labels problem can result in bias if the human decision-maker who approved the loans based the approval decision on borrower characteristics that were observable to the loan officer but are unobservable to the data scientist because they do not appear in the dataset. Imagine, for instance, that a loan officer records data on a loan applicant's occupation and, for low-paying occupations, the loan officer also evaluates informally an applicant's attire, which the officer believes is associated with creditworthiness. Assume the loan officer approves loans to well-dressed applicants in occupations that would otherwise make them ineligible for a loan and that these applicants are, in fact, more creditworthy than their occupation would suggest. Training a predictive model using only default

Another version of the problem of measurement error comes in the context of threshold analysis. In our example, the prison asserted that it needed a minimum required level of strength. As a result, the target was not the continuous variable of strength, but the applicant possessing a strength level of at least 60, which we assumed was a legitimate business necessity threshold for a prison officer job. But what if the level of strength needed is not obvious? What if the prison erroneously thought the true level of required strength was 80? We previously referred to this setting as a mis-asserted target threshold. Cases such as *Lanning v. Southeastern Pennsylvania Transporation Authority* underscore the potential for these target thresholds to be mis-asserted in a way that results in intentional discrimination, such as when they are purposefully set at a level that will adversely affect members of a protected group.

In Figure 5, we assume that, as in Figure 3, the prison implements a physical exam that perfectly measures actual strength. If the prison mistakenly sets the minimum required strength threshold at 80 (the dashed line), the resulting problem is that more women cluster in the just-failed space (between the dashed and straight line), which is the region between the mis-asserted target threshold relative to the true required strength level. As the figures shows, if an employer did not want to hire women, it could intentionally implement a mis-asserted target, knowing that more women would be excluded.

data and occupation at the time of application would therefore suggest to the model that "high risk" occupations are actually more creditworthy than they are because they default infrequently. Moreover, given racial, ethnic and gender differences in the composition of certain occupations, this model would likely be biased in addition to being inaccurate. However, evidence of this bias would become apparent in applying the IAT if one were to run the test using an estimate for creditworthiness that was based on borrowers' cash flow data as opposed to default data.

Figure 5

	Results with Valid Strength Exam		Minimum Required Strength	
Actual Strength	**Passes Exam**	**Fails Exam**	*True*	*Model*
100	□			
90	□			
80	□			⇩
70		□ ●	⇩	
60		●		
50		□		
40		□ ●		
30		●		
20		●		
10		●		
0				

□ = male; ● = female

In this setting, the exam would pass the IAT insofar that it was unbiased with respect to gender in predicting whether an applicant had strength of at least 80. However, the employer's use of the exam would nevertheless fail our definition of accountability set forth in Part II because the employer has set the cut-off at a level where qualified females are systematically exluded from the position. As emphasized in *Lanning,* this example underscores the importance of supplementing the IAT with the ability to scrutinze whether a classification threshold has been set at a level that is justified by actual business necessity.

C. Testing for "Not Statistically Correlated"

The third challenge in applying the IAT concerns how to reject the null hypothesis that no correlation exists between a set of proxy variable residuals and a protected category. In our *Dothard* illustration, the use of *Height* as a proxy for *Strength* would pass the IAT if the unexplained variation between *Strength* and *Height* (denoted as ε_i) is uncorrelated with *Gender*, as given by the test:

$$\begin{array}{ll} \text{Regression:} & \varepsilon_i = \beta_0 + \beta_1 Gender_i \\ \text{Null Hypothesis:} & \beta_1 = 0. \end{array}$$

The tradition in courts and elsewhere is to use a statistical significance level of 0.05;[154] i.e., we are willing to allow for a 5% probability of making the "Type I" error of rejecting the null hypothesis ($\beta_1 = 0$) by chance, when it is actually true. A related concept is the p-value of an estimate: the probability of obtaining an estimate for β_1 at least as far from zero as the value estimated, assuming the null hypothesis is true. If the p-value is smaller than the statistical significance level, one rejects the null hypothesis.

However, a problem with focusing on p-values is that as the sample size grows increasingly large, realized p-values converge to zero if the sample estimate for β_1 is even trivially different from the null. This is because as the sample size grows larger, the uncertainty of our estimates (usually measured by their "standard error") gets closer and closer to zero, causing any coefficient (even magnitude-irrelevant ones) to look different from an exact null of $\beta_1 = 0$ in a p-value test. In particular, a company that brings a large dataset to bear on an IAT test might be disadvantaged relative to firms with less data.

The source of the problem is the fact that in any statistical test we are actually trading off the probabilities of making two different errors: Type I errors (when we wrongly reject the null when it is, in fact, true) and Type II errors (when we wrongly fail to reject the null when it is, in fact, false). The "significance level" of a test is the probability of making a Type I error. Keeping this fixed (e.g., at 5%) as the sample size increases means that we are keeping the probability of a Type I error fixed. But at the same time, again because the standard error of our estimates is going to zero as the sample size gets large, the probability of a Type II error is actually converging to zero. If we care about both types of error, it makes sense to reduce the probability of *both* as the sample size increases, rather than fixing the probability of Type I errors and letting that of Type II errors go to zero. This point has been made forcefully by many authors, especially Edward Leamer, and a number of solutions have been proposed for adjusting the significance level as the sample size increases.[155] A full consideration of these different approaches is

[154] *See, e.g.,* Karen A. Gottlieb, *What Are Statistical Significance and Probability Values?* 1 TOXIC TORTS PRAC. GUIDE § 4:10 (2019)("Through a half century of custom, the value of 0.05 or 1 in 20 has come to be accepted as the de facto boundary between those situations for which chance is a reasonable explanation (probabilities > 0.05) and those situations for which some alternative is a reasonable explanation (probabilities < 0.05)."); *see also* Eastland v. Tennessee Valley Authority, 704 F.2d 613, 622 n. 12 (1983) (noting, in an employment discrimination lawsuit, that "a probability level of .05 is accepted as statistically significant" in determining whether racial disparities in pay were statistically significant).

[155] *See, e.g.,* Edward Leamer, SPECIFICATION SEARCHES: AD HOC INFERENCE WITH NONEXPERIMENTAL DATA (1978) (proposing p-value adjustment to minimize error losses associated with Type I and Type II error); I.J. Good*, Standardized Tail-Area Probabilities*, 16 J. STATISTICAL COMPUTATION AND SIMULATION 65 (1982) (proposing p-value adjustment based on a "Bayes/non-Bayes compromise"); Mingfeng Lin et al., *Too Big to Fail: Large Samples and the p-Value Problem,* 24 INFO. SYS. RES. 906, 908-15 (2013) (surveying approaches to adjusting p-values in large samples, recommending the reporting of effect sizes and confidence intervals, and using coefficient/p-value/sample-size plots for interpreting the data along with Monte Carlo simulations); Eugene Demidenko, *The p-value You Can't Buy*, 70 AM. STATISTICIAN 33, 34-37 (2016) (proposing the use of d-values for assessing statistical inference in large datasets).

beyond the scope of this Article. Our basic point is to note when the issue will be relevant when applying the IAT and that there are a number of solutions to it. We provide below an example of one such approach to illustrate how it can be utilized to discern when a seemingly significant result when applying the IAT is actually a function of the large sample size and not evidence of a discriminatory proxy variable.

D. Nonlinearities or Interactions Among Proxies

Machine learning models are often focused on forming predictions based on nonlinear functions of multiple variables. In introducing the IAT, our specification focused on linear settings, but the IAT could in principle be amended to handle nonlinear models as well. For example, rather than just running the test regression once, we could run it repeatedly, with each of a set of basis functions of the explanatory variables on the left-hand side. Our goal in this Article has been to translate the legal notion of accountability under Title VII into the context of statistical modelling at the heart of algorithmic decision-making; therefore, we leave a more thorough consideration of this topic to future work. However, in general, implementation of the IAT could be made part of the type of feature selection and feature analysis protocols that are used in practice with both linear and non-linear machine-learning processes.[156]

VI. SIMULATION

We conclude with a simulation illustrating how to use the IAT to examine disparities arising from a hiring algorithm. In addition to demonstrating the application of the IAT, the simulation also illustrates how to address many of the concerns noted in Part V. As in Part III, we base the simulation on the facts of *Dothard*.

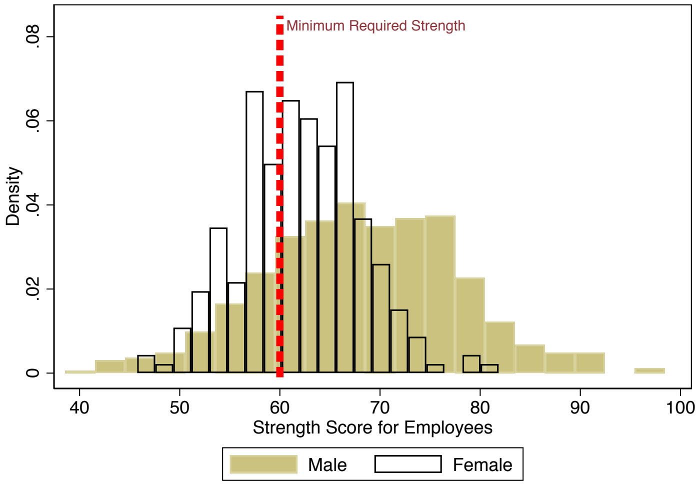
[156] In particular, a related literature in computer science focuses on feature selection to enhance model interpretability. *See* Datta et al. *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, *Proceedings of IEEE Symposium on Security & Privacy 2016*, 598–617, 2016 (proposing a quantitative-input-influence (QII) protocol based upon Shapley values to determine the importance of features and clustering metrics to summarize feature influence); *see also* Phillipe Bracke et al., *Machine learning explainability in default risk analysis*, Bank of England Staff Working Paper No. 816 (June 5, 2019) (implementing QII method in predicting mortgage defaults). More formally, Lundberg, et al., *Consistent Individualized Feature Attribution for Tree Ensembles*, arXiv:1802.03888v3 [cs.LG], March 7, 2019 and Merrill et al., *Generalized Integrated Gradients: A practical method for explaining diverse ensembles*," ArXiv 2019, build upon game-theoretic SHAP (Shapley Additive explanation) values and propose new feature credit-assignment algorithms that can handle a broad class of predictive functions with both piecewise-constant (tree-based), continuous (neural-network or radial-basis-function based), and mixed models.

A. Set-Up

The simulation assumes that the prison has historical records for 800 employees, of which roughly one-third are female (n=256) and two-thirds are males (n=544). We further assume that the prison uses these historical records to develop a sorting algorithm for considering a pool of 1,200 applicants. The 800 employees are endowed with an *unobservable* strength level, which we model as a random variable distributed normally with (i) a mean of 68 and a standard deviation of 10 for male employees and (ii) a mean of 62 and a standard deviation of 6 for female employees. With these modeling assumptions, females have lower mean strength but a smaller standard deviation, as plotted below in Figure 6. To be an effective prison officer requires a minimum strength of 60, the business necessity. The prison's past hiring is not perfectly effective at sorting which officers will meet this threshold; therefore, even among the employees, there are officers who fall below the required strength for the job. For now, we assume that the prison can implement a costly physical exam to measure true strength for these employees. (We abstract from other aspects of effectiveness such as psychological and managerial skills needed for prison-officer work.)

We assume the strength of applicants is likewise distributed randomly. However, for obvious reasons, the applicant pool has not been previously selected for strength as employees have. Therefore, we model strength across applicants as a random variable distributed normally with a mean of 50 and a standard deviation of 10 for male employees and a mean of 44 and a standard deviation of 6 for female employees.

Figure 6
Distribution of Strength Across Prison Officers



The prison managers cannot directly observe applicants' strength, and implementing a full physical exam across applicants is costly. Therefore, the prison decides to use height as a proxy variable for an applicant's strength, since it is easily measured on applications. We model height as a sum of a baseline 50 inches (with a normally-distributed error of 4 inches) plus a concave (quadratic) function increasing in strength. Female height has the same relation to strength but a ten percent lower baseline. The resulting mean height in the employee training dataset is 5'10" with a standard deviation of 5".

Finally, as in *Dothard*, the prison seeks to filter applicants by imposing a minimum height requirement. To determine the height cut-off, the prison runs a classification analysis. In doing so, the prison determines that they want to ascertain that an individual will be above the strength threshold with an 80% certainty, i.e., they want only a 20% risk of incorrectly classifying an applicant as eligible for hiring (above the strength threshold of 60) when the person in fact has a strength of less than 60. Based on the height and strength of the prison officers, this results in a 5'10" cut-off. The prison applies this cut-off to all 1,200 applicants.

Among the 370 female applicants, 344 (93%) fail the height test. In contrast, among the 830 male applicants, 504 (61%) fail the height test. These disparities suggest that the height cut-off may discriminate against female applicants, but we cannot definitively conclude this from the high rejection rates because, as we saw in Figure 6, females in our samples have lower strength than males on average.

B. Applying the Input Accountability Test

Assume that in advance of deploying the height test, the prison instead decides to conduct the IAT to ensure that any disparities in hiring would be based on differences in predicted applicant strength. Table 2 presents the results from the test. To run the IAT, the prison would return to the training data it possesses regarding its employees' actual strength and height that it used to determine the 5'10" cut-off. In panel A, we present the first step of regressing the proxy variable of height on employee strength, the target of interest. Because the prison is focused on using a cutoff for height, we estimate a logistic regression of whether an employee passes the height cut-off as a function of the employee's strength. (To do so, we use as our dependent variable an indicator variable that equals 1 for employees that are at least 5'10" and 0 for all others.) Note that this indicator variable is on the left-hand side of the regression (and not strength) because we want to decompose whether an employee meets the height cut-off into two components – the part that can be predicted from an employee's strength and the part that cannot be predicted from an employee's strength (the "residual").

Stated differently, logistic regression effectively estimates the probability that an employee is 5’10” based on employee strength. Therefore, the residual, which is equal to one minus this predicted probability for each employee, can be viewed as the variation in whether an employee meets the height threshold of 5’10” that is unrelated to an employee’s strength. In panel B, we present the results from regressing the residual from panel A onto the indicator variable for female.

Table 1

	(1)	(2)	(3)	(4)	(5)
Panel A: First Step of IAT (Dependent Variable =Column Heading)					
	Cut-Off Height	Cut-Off Muscle Mass	Muscle Mass	Strength Assessment	Cut-Off Muscle Mass
Strength	0.0206***	0.0377***	0.9965***		0.0387***
	[0.00155]	[0.000747]	[0.0191]		[0.0000138]
Muscle Mass				0.675***	
				[0.0307]	
Observations	800	800	800	800	2,000,000
[Pseudo] R-squared	0.111	0.466	0.772	0.376	0.496
Panel B: Second Step of IAT (Dependent Variable=Residuals from Step 1)					
Female	-0.354***	-0.013265	-0.3552	-8.858***	-0.0013***
	[0.0327]	[0.02625]	[0.379]	[0.542]	[0.000505]
Observations	800	800	800	800	2,000,000
R-squared	0.128	0	0	0.25	0
d-value					50%

Standard errors in brackets
*** p<0.01, ** p<0.05, * p<0.1

Panel A of Column (1) reports that strength only accounts for a small part of the variation (R-squared = 0.111) for whether an employee is (or is not) taller than 5’10”. In Panel B, our column (1) results show that the residual of the first step regression has a negative, significant correlation with gender, thus failing the IAT. Females incur a penalty because the proxy variable for the business necessity of required strength has residual correlation with gender.

Imagine that the prison realizes this flaw in using a height cut-off and decides instead to consider incurring an extra cost for doing a muscle-mass index evaluation of applicants. Because the evaluation is imperfect in assessing true strength, we assume that the results of a muscle-mass index evaluation is equal to an individual's strength plus random noise.[157] To implement this screening procedure, the prison first applies the muscle-mass index evaluation to existing employees so that it can estimate the minimum muscle mass an individual should have to be above the minimum strength threshold with an 80% certainty. The classification analysis produces a muscle-mass cut-off score of 64. As above, the prison then conducts the IAT.

In column (2) of panel A we present the results of the IAT for the muscle-mass index evaluation based on the employee training data. To implement the IAT, we run the same regressions that we used for testing the height cut-off, but we substitute an indicator variable for whether an employee has a muscle mass of at least 64 for the indicator variable for whether an employee is at least 5'10". In panel A, column (2) shows that the probability that an employee has a muscle mass of at least 64 is (unsurprisingly) related to an employee's strength, resulting in a much larger R-squared. Importantly, the residual should not fail the IAT, because (by construction) it has no bias against females. In column (2) of panel B, we see that this is indeed the case; the coefficient on female is statistically insignificant and small in magnitude.

In column 3, we instead consider a continuous variable version of muscle mass as a scoring variable rather than a cut-off version of the indicator variable. Perhaps the underlying job-required strength is not a threshold but a strength score that will feed into wage-setting or other profiling of individuals that focus on continuous rather than discrete measures. To implement the IAT in this context, we use the same training data that was used for column (2) of Table 2; however, the regression specification for the first step takes the form of a linear regression of employees' muscle mass scores on their measured strength. As in column (2), column (3) shows that muscle mass is a legitimate business necessity variable. In panel A, we find that muscle mass and strength are very correlated, with strength accounting for almost 80% of the variation in muscle mass. Column (3) of panel B shows that muscle mass again passes the IAT: the residual is uncorrelated with the female indicator variable.

In the final two columns of Table 2, we demonstrate the importance of two challenges we introduced in Part 5.

First, we use column (4) to illustrate the concern about measurement error in the target (strength). Thus far, we have been working under the assumption that the prison can take an accurate measurement via a physical exam of the training dataset employees. However, what if instead the prison cannot measure actual strength but uses a strength score made by a manager. (We label this assessment measure an employee's "Strength Assessment"). As noted above, a central challenge in real world settings is that target

[157] We model the random noise as a random variable drawn from a normal distribution having a mean of zero and a standard deviation of 5.

variables used to train predictive models are typically estimated in this fashion and may contain measurement error that is correlated with a protected characteristic. We therefore simulate an employee's Strength Assessment as biased against females.[158] In this regard, the simulation replicates the same problem illustrated with the UnitedHealth algorithm discussed previously (where the illness severity measure was inadvertently biased against African Americans).

In addition to employees' Strength Assessment, assume that the prison also has at its disposal data from the muscle measure index evaluation used in columns (2) and (3). Even without perfect data regarding employee strength, the prison can still use these data with the IAT to evaluate whether its preferred estimate of the target (an employee's Strength Assessment) suffers from bias. To implement this test, we treat muscle mass as an alternative measure of the target of interest (strength), and we treat the Strength Assessment as a proxy for strength, as we did for height in columns (1) – (3). Accordingly, the first step of the IAT is conducted by regressing employees' Strength Assessment on the muscle mass evaluation data. The results are shown in column (4) of panel A. Not surprisingly, an employee's muscle mass is closely related to an employee's Strength Assessment. In column (4) of panel B, we show the results of regressing the residuals from this regression on the gender variable. As shown in the table, Strength Assessment fails the IAT. In this fashion, the IAT can be used to test whether an estimate for a target suffers from biased measurement error, so long as one has an alternative estimate for the target (even a noisy one) that is believed to be unbiased.

The final column in Table 2 illustrates the concern of large data samples. For this column, we implement the same muscle mass test as in column (2), except that we randomly draw 2 million employees for the training dataset rather than 800 employees. (For all 2 million employees, we model their strength using the same assumptions used for the original 800 employees). For each employee, we likewise calculate muscle mass as employee strength plus a random variable distributed normally with a mean of 0 and a standard deviation of 5. Thus, in our simulated setting, muscle mass is a noisy estimate of employee strength but it has zero bias with respect to gender. Even so, however, the possibility remains that in drawing random measurement error for our sample, very slight differences may exist by chance between the average measurement error of females and males. (This is equivalent to observing that even if a coin is unbiased, it may still return more than 50% heads in a trial of 100 flips). Moreover, as we described in Part III, the p-

[158] In particular, for males, we model the Strength Assessment measure as strength plus random noise; however, for females, we model Strength Assessment as concave in strength (like the height variable)—a quadratic concave function of strength plus random noise. This modeling assumption implies that the managers evaluating females do not fairly evaluate them, especially for the stronger females.

value may converge to 0 for any small deviation, as sample sizes approach infinity. Thus, even a small (economically non-meaningful) correlation may look significant. This would create a setting of a large-dataset proxy variable failing the IAT, not because of a fundamental problem, but just because of the use of a fixed p-value. This is what we find in column (5). The coefficient on female in column (5) of Panel B is very small (-0.0013) but statistically significant, notwithstanding the fact that we modeled measurement error from a distribution that had exactly zero gender bias.

As noted in Part 5(C), when the IAT is applied to a large dataset, it is therefore critical to check whether a proxy that fails the IAT might have failed the test simply because of the large number of observations in the sample. That the seemingly statistical finding in column (5) may be an artifact of a trivial difference within a large dataset can initially be seen by the fact that the R-squared in column (5) of Panel B is 0%; if effectively no variation in the residuals can be explained by gender, how can it be that this proxy is penalizing females in a systematic fashion? Additionally, as noted previously, a number of formal solutions exist to examine this issue more fully. Here, we illustrate one such approach using the concept of the "d-value" proposed by Eugene Demidenko.[159] Rather than focus on a comparison of group means, the d-value is designed to examine how a randomly chosen female fared under this proxy variable relative to a randomly chosen male. Specifically, in the context of the IAT, the d-value answers the question "what is the probability that members of a protected group are being penalized by the proxy?" As shown in the last row of column (5) of Panel B, the d-value is approximately 50%, indicating that the probability that females are penalized by the use of a muscle-mass proxy is effectively a coin-toss; that is, there is no evidence that female applicants are being systematically penalized by the use of this proxy.

This finding, of course, is hardly a surprise given that we designed the simulation for column (5) to ensure that it was an unbiased proxy. In this fashion, the use of a d-value can highlight when a seemingly significant finding is a function of the large sample size and not evidence of a discriminatory proxy variable.[160]

[159] *See* Demidenko, *supra* note 155.

[160] To the extent one utilizes the d-value in this fashion, a natural question is what level of a d-value would constitute evidence of a discriminatory proxy. Given that the d-value answers the question "what is the probability that members of a protected group are being penalized by the proxy?", any result that yields a d-value deviating from 50% would presumably be evidence of a discriminatory proxy, allowing for a percentage difference to incorporate a far tail sampling draw. This conclusion follows from the conventional judicial reliance to on p-values, which likewise assumes that any finding with a p-value of less than 0.05 is evidence of discrimination. That said, in adopting such an approach, it would be important to utilize a d-value analysis only upon a finding that a proxy fails the IAT using a conventional statistical test. The reason stems from the fact that in smaller samples, even an unbiased proxy could result in a d-value that is slightly different from 50% due sample variance. For example, the d-value for column (3) is just slightly less than 51%, despite the fact that muscle mass is modeled as an unbiased proxy. However, running the same simulation with 50,000 observations produces a d-value of 50%.

VII. CONCLUSION

The era of Big Data places the antidiscrimination mandate at the heart of the Civil Rights Acts of 1964 and 1968 at a critical cross-roads. By relying on data-driven, statistical models, machine learning provides a promising alternative to the type of subjective, face-to-face decision-making that has traditionally been fraught with the risk of bias or outright animus against members of protected groups. Yet left unchecked, algorithmic decision-making can also undermine a central goal of U.S. antidiscrimination law. As we have shown throughout this Article, any decision-making rule that simply maximizes predictive accuracy can result in members of historically marginalized groups being systematically excluded from opportunities for which they are qualified to participate.

Ensuring that algorithmic decision-making promotes rather than inhibits equality thus demands a workable antidiscrimination framework. To date, however, prevailing approaches to this issue have focused on solutions that fail to grapple with the unique challenge of regulating statistical discrimination. Prominent regulatory approaches (such as reflected in HUD's recent proposed rule-making) have frequently prioritized predictive accuracy despite the fact that such an approach ignores the central risk posed by statistical discrimination demonstrated in our simulation. Conversely, interventions emanating from the field of computer science have largely focused on outcome-based interventions that could themselves lead to claims of intentional discrimination.

Because we derive our input accountability test from caselaw addressing statistical discrimination—in particular, the burden-shifting framework—the IAT advances a vision of algorithmic accountability that is consistent with the careful balance courts have struck in considering the decision-making benefits of statistical discrimination while seeking to minimize their discriminatory risks. By enhancing the predictive accuracy of decision-making, statistical discrimination can greatly enhance the ability of an employer, lender or other decision-maker to identify those individuals who possess a legitimate target characteristic of interest. However, cases such as *Griggs* and *Dothard* underscore the danger of simply focusing on predictive accuracy because a proxy that predicts a target variable can nonetheless result in systematically penalizing members of a protected group who are qualified in the target characteristic. That such discriminatory proxies have been consistently declared to be off limits underscores the conclusion that predictive accuracy alone is an insufficient criterion for evaluating statistical discrimination under U.S. antidiscrimination law.

At the same time, our approach is also consistent with the focus in *Griggs* and *Dothard* that differences in a legitimate target can justify disparities that differ across members of protected and unprotected groups. As we show, so long as a proxy used to predict a legitimate target variable is unbiased with respect to a protected group, it will pass the IAT, even if it results in disparate outcomes. The IAT can therefore provide greater transparency into whether disparate outcomes are the result of a biased model or more systemic disparities in the underlying target variable of interest, such as credit risk. In so doing, it can provide vital information about whether the proper way to address observed disparities from an algorithmic model is through de-biasing the model or through re-defining the target in a more equitable fashion or addressing disparities in the underlying target variable of interest (such as through targeted subsidies or other transfers). More generally, because the goal of the IAT is to avoid penalizing members of a protected group who are otherwise qualified in a target characteristic of interest, our approach will also be immune to the concern informing cases such as *Ricci v. DeStefano* that our test is biased against qualified individuals.

Finally, our approach provides clear "rules of the road" for how to exploit the power of algorithmic decision-making while also adhering to the antidiscrimination principles at the heart of the Civil Rights Acts of 1964 and 1968. In particular, the IAT offers data scientists a simple test to use in evaluating the risk that an algorithm is producing biased outcomes, mitigating a key source of the regulatory uncertainty surrounding the growing use of algorithmic decision-making. Additionally, our exploration of the early caselaw considering statistical discrimination also reveals that these rules of the road encompass more general concepts to guide both data scientists and regulators when evaluating algorithmic discrimination. These include the notion that, fundamentally, algorithmic decision-making is an effort to assess an unobservable attribute, such as productivity, criminality, longevity, or creditworthiness, through the use of one or more proxy variables. Consequently, evaluating an algorithm must begin with transparency about this target characteristic. And they likewise include the fact that correlation between the unobservable characteristic and the proxy is not, by itself, sufficient to justify the use of the proxy under antidiscrimination principles.

APPENDIX
DE-BIASING PROXY VARIABLES VERSUS DE-BIASING PREDICTIVE MODELS

In this Appendix, we conduct a simulation exercise to illustrate how attempting to de-bias a proxy variable used in a predictive algorithm may do little to de-bias the ultimate predictions. The example we use assumes that a college admissions director wishes to use applicants' standardized test scores (STS) to predict college success. For this purpose, we assume that a student's performance on the STS is a function of just two equally-important factors: *aptitude* and *family wealth*. In our simulation, wealth contributes to test performance because children from wealthier households often purchase expensive test preparation classes. To keep the simulation tractable, we assume that wealth does not affect college performance; its only effect is on a student's STS.

Our simulation uses a hypothetical training dataset of 1,000 college graduates where the admissions director has data on each student's STS at the time of application, student race, and the student's ultimate college performance (e.g., a weighted grade point average or other measure of performance). We divide the race of students, X_i^R, equally so that 500 students are non-White ($X_i^R = 0$) and 500 are White ($X_i^R = 1$). We assume that wealth and aptitude are distributed as follows:

$$X_i^{Wealth} \sim \begin{cases} N(0,1)\ if\ X_i^R = 0 \\ N(5,1)\ otherwise \end{cases}$$

$$X_i^{Aptitude} \sim N(0,1)$$

That is, a student's wealth is defined to be a random variable drawn from a normal distribution for all students. However, the mean and standard deviation for White students are 5 and 1, respectively, while it is 0 and 1 for non-White Students. In contrast, a student's aptitude is modeled as a random variable drawn from a normal distribution having a mean of 0 and a standard deviation of 1 for all students regardless of race.

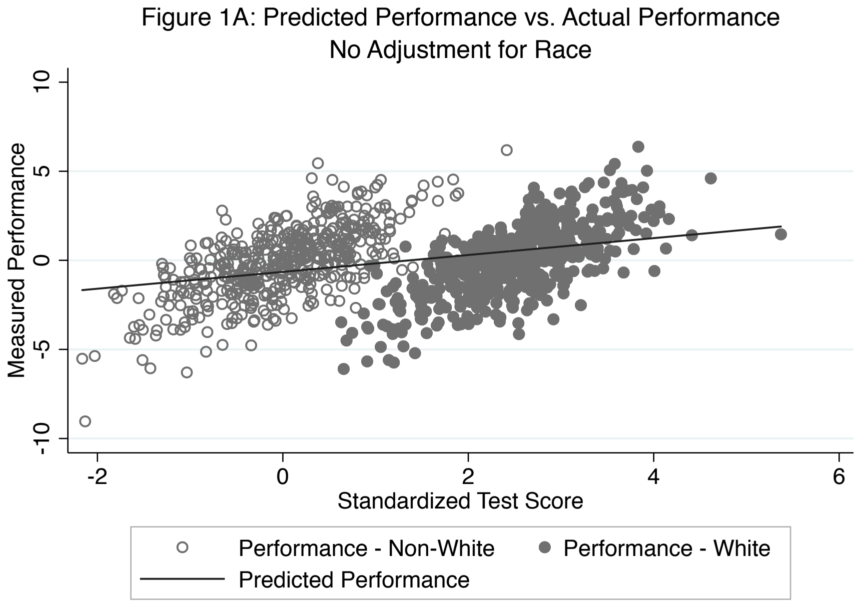
Note that under these distributional assumptions, there is very little common support in wealth across race categories. This is by design to illustrate the point noted by Kristen Altenburger and Daniel Ho that in these settings, the effort to de-bias proxy variables can produce the largest estimation errors.[161] As noted, a student's STS (X_i^{STS}) is a function of X_i^{Wealth} and $X_i^{Aptitude}$, with each variable given equal weight:

[161] *See* Altenburger & Ho, *supra* note 116, at 111. These settings arise "where sharp preexisting demographic differences may exist across groups." *Id*.

$$X_i^{STS} = 0.5\left(X_i^{Wealth}\right) + 0.5\left(X_i^{Aptitude}\right)$$

Finally, we simulate college performance (*Performance*$_i$) to be entirely determined by aptitude multiplied by a scalar (which we assume here to be 2).

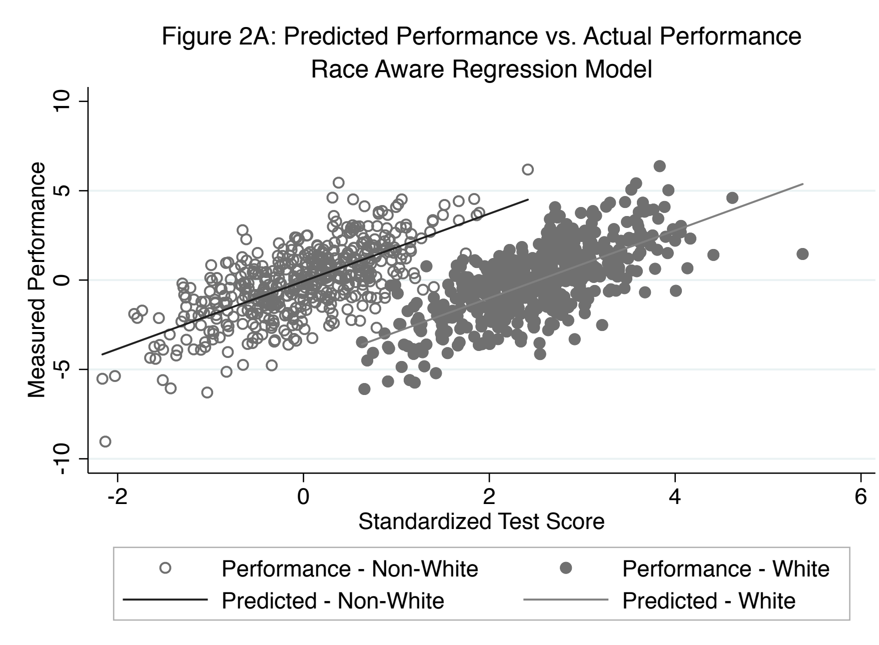
Aptitude is unobservable to the admissions director, inducing her to estimate whether she can use STS to predict college performance. In Figure 1A, we plot the relationship between college performance and STS for White and non-White graduates separately based on data simulated using the foregoing assumptions. We also include a line that provides the predicted college performance from a simple regression of college performance on STS. As shown in the Figure, White graduates had much higher STS on average, as would be expected from their higher family wealth.



The admissions director would like to admit students that are likely to have a positive measure of college performance (i.e., *Performance>0*). She therefore runs a simple regression of *Performance* on *STS*, which produces a regression coefficient ($\hat{\beta}^{STS}$) of 0.47. This estimate indicates that a one-point change in *STS* is associated with a 0.47 change in *Performance*. Using this regression estimate, she generates the fitted line shown in Figure 1A, which provides a predicted measure of *Performance* based solely on *STS*. The fitted line predicts that *Performance* is zero at roughly 1.3, suggesting that using a minimum *STS* of 1.3 would admit students with an expected college performance of at least 0. However, had the admissions director applied this cutoff to these individuals, the bias in *STS* would result in significant bias against non-White students owing to their lack of access to test preparation classes:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	13	465

Now assume that the admissions director seeks to control for the greater wealth (and therefore, the greater test preparation bias) among White applicants. Using the same data, she expressly adds X_i^R as a control variable in the regression of *Performance* on *STS.* Doing so allows her to predict *Performance* as a function of both *STS* and *Race.* The results are presented in Figure 2A.



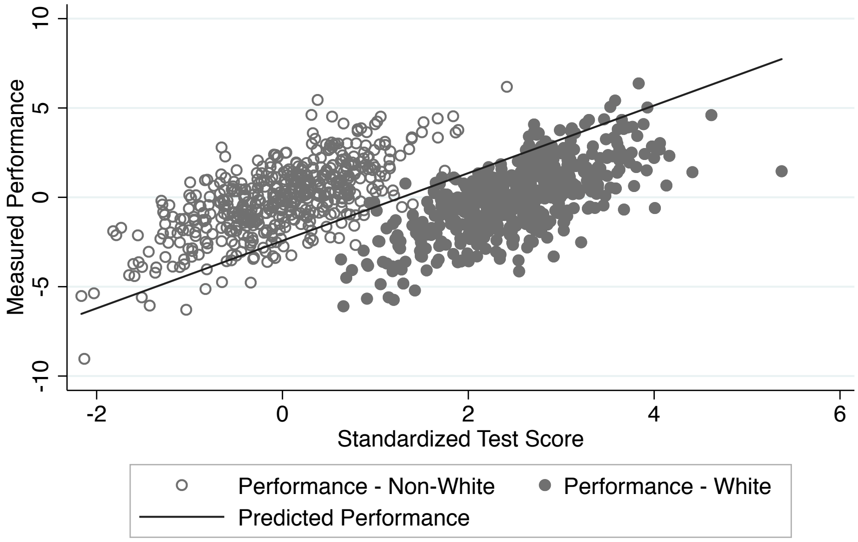
This procedure corrects for the racial bias that arises from using only *STS* to predict *Performance.* This can be seen by the two fitted regression lines, which do a much better job of predicting measured performance across the two racial groups than in Figure 1A. The reason stems from the fact that this regression specification estimates a different y-intercept for each racial group in estimating the relationship between *STS* and *Performance*. Specifically, the regression yields a y-intercept for X_i^R of -4.72, which indicates that in using *STS* to predict *Performance*, it is necessary to deduct 4.72 from the expected performance of White students. (Recall that the difference in average wealth across White and non-White students is 5.0, so this adjustment eliminates the bias that wealth creates when using *STS* as a measure of aptitude). With that adjustment, the regression coefficient for *STS* increases from 0.47 to 1.89 because the regression has effectively removed the confounding effect of wealth on *STS* so that it more accurately reflects aptitude. As above, the admissions director evaluates each fitted line and determines that the fitted line for non-White students predicts that *Performance* is zero where *STS* is also zero, and that the fitted line for White students predicts that *Performance* is zero where *STS* is 2.53. Applying a minimum test cut-off of 0 for non-White students and 2.53 for White students would result in the following students being deemed qualified:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	250	248

This procedure solves the racial bias created by using only *STS* to estimate *Performance*, but it is clearly problematic insofar that it requires a different minimum cut-off for White and Non-White students. This is disparate treatment. To avoid this problem, the admissions director therefore turns to the approach advanced by Devin Pope and Justin Sydnor as well as by Crystal Yang and Will Dobbie.[162] This procedure involves using the regression estimates generated for Figure 2A but treating all students as if they had the average value of race, which is 0.5 in this example. Making this adjustment means that every student receives a deduction of -2.36 (i.e., 0.5 * -4.72) after multiplying their score by the slope coefficient for *STS* of 1.89, which remains purged of the confounding influence of wealth. This enables the admissions director to estimate a single fitted regression line as shown in Figure 3A:

[162] *See supra* note 66.

Figure 3A: Predicted Performance vs. Actual Performance
Race Blinded Regression Model



The fitted line predicts that *Performance* is zero at approximately 1.28, which the director uses as the minimum cut-off. Had the director applied this cut-off to this group of individuals, the following results would have occurred:

	Non-White	White
# of Qualified Candidates Predicted by Test Score	15	468

In effect, the results are largely identical to those obtained by using only *STS* to predict performance. The reason stems from the lack of common support in wealth across White and non-White students, resulting in the need for a significant negative adjustment to every White student when estimating performance from *STS*. Applying half of this negative adjustment to *every* student thus works against the de-biasing of the slope coefficient for *STS*. In short, the slope coefficient for *STS* in Figure 3A is unbiased with respect to non-White students, but the predictive model is not. This problem was significant in this example because there is so little common support in wealth across White and non-White students—a problem that will exist whenever there are significant demographic differences across protected and unprotected groups.