Less Discriminatory Algorithms

EMILY BLACK*†, JOHN LOGAN KOEPKE**†, PAULINE T. KIM***, SOLON BAROCAS**** & MINGWEI HSU*****

In discussions about algorithms and discrimination, it is often assumed that machine learning techniques will identify a unique solution to any given prediction problem, such that any attempt to develop less discriminatory models will inevitably entail a tradeoff with accuracy. Contrary to this conventional wisdom, however, computer science has established that multiple models with equivalent performance exist for a given prediction problem. This phenomenon, termed model multiplicity, suggests that when an algorithmic system displays a disparate impact, there almost always exists a less discriminatory algorithm (LDA) that performs equally well. But without dedicated exploration, developers are unlikely to discover potential LDAs. These observations have profound ramifications for the legal and policy response to discriminatory algorithms. Because the overarching purpose of our civil rights laws is to remove arbitrary barriers to full participation by marginalized groups in the nation's economic life, the law should place a duty to search for LDAs on entities that develop and deploy predictive models in domains covered by civil rights laws, like housing, employment, and credit. The law should recognize this duty in at least two specific ways. First, under disparate impact doctrine, a defendant's burden of justifying a model with discriminatory effects should include showing that it made a reasonable search for LDAs before implementing the model. Second, new regulatory frameworks for the governance of algorithms should include a requirement that entities search for and implement LDAs as part of the model building process.

---

* Assistant Professor, Department of Computer Science and Engineering and the Center for Data Science, New York University.

** Project Director, Upturn.

*** Daniel Noyes Kirby Professor of Law, Washington University School of Law, St. Louis, Missouri.

**** Principal Researcher, Microsoft Research; Adjunct Associate Professor, Information Science, Cornell University.

***** Senior Quantitative Analyst, Upturn.

† Equal contribution.

†† The authors would like to thank the following individuals for their helpful feedback: Olga Akselrod, Elizabeth Edenberg, Talia Gillis, Stephen Hayes, Daniel Jellins, Cynthia Khoo, Michael McGovern, Paul Ohm, Catherine Powell, Manish Raghavan, Matthew Scherer, Andrew Selbst, Ridhi Shetty, Eric Sublett, Dan Svirsky, Suresh Venkatasubramanian, and the staff of Upturn. The authors are also grateful to the participants at the 2023 Privacy Law Scholars Conference, the participants at the Law & Technology Workshop, the Washington University School of Law faculty workshop for their comments, to Julia Monti and Kelly Miller for excellent research assistance, and to the Editors at *The Georgetown Law Journal* for their careful and thoughtful editorial work to bring this Article to print.

TABLE OF CONTENTS

INTRODUCTION

Companies now routinely deploy algorithmic systems as part of their basic business operations to determine who gets access to critical opportunities and resources. These developments have worried legal scholars and civil rights advocates, who are concerned that algorithms may reflect or reinforce existing societal biases.[1] From tenant screening systems to employment assessment and hiring technologies to credit underwriting models, reliance on these tools raises concerns that they will discriminate against or exclude historically marginalized groups.[2] These concerns have stimulated a vast scholarship on how the law should respond. One strand of the literature focuses on existing civil rights law and how it applies to new technologies, debating whether disparate impact doctrine is adequate to meet the challenges posed by algorithmic tools.[3] Another

1. *See, e.g.*, Pauline T. Kim, *Auditing Algorithms for Discrimination*, 166 U. PA. L. REV. ONLINE 189, 196 (2017); Haley Moss, *Screened Out Onscreen: Disability Discrimination, Hiring Bias, and Artificial Intelligence*, 98 DENV. L. REV. 775, 783 (2021); Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633, 689 (2020); William Magnuson, *Artificial Financial Intelligence*, 10 HARV. BUS. L. REV. 337, 354–55 (2020); Kristin Johnson, Frank Pasquale & Jennifer Chapman, *Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation*, 88 FORDHAM L. REV. 499, 502 (2019); Crystal S. Yang & Will Dobbie, *Equal Protection Under Algorithms: A New Statistical and Legal Framework*, 119 MICH. L. REV. 291, 294 (2020); Margaret Hu, *Algorithmic Jim Crow*, 86 FORDHAM L. REV. 633, 638 (2017); *Civil Rights Principles for the Era of Big Data*, LEADERSHIP CONF. ON CIV. & HUM. RTS. (Feb. 27, 2014), https://civilrights.org/2014/02/27/civil-rights-principles-era-big-data/ [https://perma.cc/N4E2-BBEP]; ACLU et al., *Principles*, CIV. RTS. PRIV. & TECH. TABLE (2020) [https://perma.cc/9REC-JDU7]; Letter from Upturn, ACLU & Leadership Conf. on Civ. & Hum. Rts. to Dr. Eric S. Lander, Dr. Lynne Parker & Dr. Alondra Nelson (July 13, 2021), https://www.upturn.org/work/proposals-for-the-biden-administration-to-address-technologys-role-in/ [https://perma.cc/L8JU-VMXY]; Letter from Leadership Conf. on Civ. & Hum. Rts to Ambassador Susan Rice (Oct. 27, 2021), https://civilrights.org/resource/letter-to-ambassador-rice-on-civil-rights-and-ai [https://perma.cc/5UE7-WXPZ].

2. *See, e.g.*, Khari Johnson, *Algorithms Allegedly Penalized Black Renters. The US Government Is Watching*, WIRED (Jan. 16, 2023, 7:00 AM), https://www.wired.com/story/algorithms-allegedly-penalized-black-renters-the-us-government-is-watching/ (discussing biases in tenant screening systems); Miranda Bogen, *All the Ways Hiring Algorithms Can Introduce Bias*, HARV. BUS. REV. (May 6, 2019), https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias [https://perma.cc/7MB9-2568] (discussing bias in employment assessment and hiring technologies); Will Douglas Heaven, *Bias Isn't the Only Problem with Credit Scores—and No, AI Can't Help*, MIT TECH. REV. (June 17, 2021), https://www.technologyreview.com/2021/06/17/1026519/racial-bias-noisy-data-credit-scores-mortgage-loans-fairness-machine-learning/ [https://perma.cc/47YW-X5JH] (discussing bias in credit underwriting models).

3. *See, e.g.,* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 701–12 (2016); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 903–16 (2017); Michael Selmi, *Algorithms, Discrimination and the Law*, 82 OHIO ST. L.J. 611,

strand of the literature eschews a litigation focus, arguing instead for a proactive approach that would regulate ex ante how algorithms are developed in an effort to prevent the deployment of biased tools.[4]

An often-unspoken premise in discussions about algorithmic fairness is that once a particular prediction problem has been defined, a unique solution exists. So, if, for example, a bank seeks to predict default by borrowers, it is assumed that one correct model exists that best meets that goal. It then follows that any deviations from that unique solution would necessarily entail a loss of performance. The implication is that pursuing goals like minimizing discrimination will unavoidably involve a tradeoff with accuracy—an idea perhaps inspired by the focus on the tensions between the two in the computer science community.[5] These assumptions—that a unique solution exists and that a fairness-accuracy tradeoff is inevitable—are descriptively inaccurate. Recent work in computer science has established that there are almost always multiple possible models with equivalent accuracy for a given prediction problem—a phenomenon termed model multiplicity.[6] Notably, this phenomenon is not limited to models with

634–43 (2021); Mikella Hurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 190–95 (2016).

4. *See, e.g.*, Ifeoma Ajunwa, *An Auditing Imperative for Automated Hiring Systems*, 34 HARV. J.L. & TECH. 621, 661 (2021); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 168–81 (2017); Andrew D. Selbst & Solon Barocas, *Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law*, 171 U. PA. L. REV. 1023, 1029–44 (2023); Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1537–52 (2019); Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 2032 (2021).

5. *See, e.g.*, Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez & Krishna P. Gummadi, Fairness Constraints: Mechanisms for Fair Classification, *in* 54 PROCEEDINGS OF THE 20TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS 962, 963 (Proc. Mach. Learning & Rsch. Eds., 2017), https://proceedings.mlr.press/v54/zafar17a.html; Christian Haas, The Price of Fairness–A Framework to Explore Trade-Offs in Algorithmic Fairness, 1, 3 (Int'l Conf. on Info. Sys., 2019 Proc., Munich, Germany); Michael Wick, Swetasudha Panda & Jean-Baptiste Tristan, Unlocking Fairness: A Trade-off Revisited, *in* 33RD CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 8751, 8751 (Curran Assocs. ed., 2019), https://papers.neurips.cc/paper_files/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html.

6. Several different terms have been used to describe related phenomena over the years in computer science and statistical scholarship. Leo Breiman first introduced the notion that various models could be equally effective at the same task. *See* Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199, 206 (2001) (using the term "the Rashomon effect"). Charles T. Marx, Flavio du Pin Calmon, and Berk Ustun resurfaced the idea that different models could have different predictions but similar performance, under the term "predictive multiplicity." *See* Charles T. Marx, Flavio du Pin Calmon & Berk Ustun, Predictive Multiplicity in Classification, *in* 119 PROCEEDINGS OF THE 37TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 6765, 6765 (Proc. Mach. Learning & Rsch. eds., 2020). Emily Black and Matt Fredrikson displayed similar behavior on different classes of models in concurrent work. *See* Emily Black & Matt Fredrikson, Leave-One-Out Unfairness, *in* FACCT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 285, 286 (Ass'n for Computing Mach. eds., 2021). Later, Emily Black, Manish Raghavan, and Solon Barocas introduced a term, "model multiplicity," to encompass not only how similarly performing models differ in their predictions, but also in their internals, which have impacts on the explanations of their predictions. *See* Emily Black, Manish Raghavan & Solon Barocas, Model Multiplicity: Opportunities, Concerns, and Solutions, *in* FACCT '22: PROCEEDINGS OF THE 2022 5TH ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 850, 850 (Ass'n for Computing Mach. eds., 2022). The terms

especially low accuracy; it also applies to models that achieve levels of accuracy that firms find perfectly acceptable for use in practice, even in high-stakes decisionmaking scenarios regulated by discrimination law.

Multiplicitous models perform the chosen prediction task equally well, but they may differ in many other ways, including which features they use to make their predictions, how they combine these features to make their predictions, and whether their predictions are robust to changing circumstances.[7] And, significantly for our purposes, these comparably performing models can have different levels of disparate impact across groups. As a result, when an algorithmic system displays a disparate impact, model multiplicity suggests that other models exist that perform equally well but have less discriminatory effects. In other words, in almost all cases, a less discriminatory algorithm exists.

In this Article we explore the phenomenon of model multiplicity and the resulting insight that a less discriminatory algorithm—what we refer to as an LDA—almost always exists whenever a model has a disparate impact. The availability of LDAs opens the possibility of reducing, or in some cases eliminating, the negative effects of algorithmic systems on marginalized groups without compromising business objectives. Unfortunately, there is no guarantee that the model development process will uncover less discriminatory models unless the developer makes an effort to search for them.[8] Models used to sort and select applicants are typically developed through a development pipeline entailing numerous choices that narrow the range of models explored.[9] Each of those choices eliminates consideration of some possible models, and unless developers are deliberately testing for models with less disparate impact along other branches of the pipeline, they are unlikely to happen upon them.[10]

These insights about model multiplicity have profound ramifications for the legal response to discriminatory algorithms. Civil rights statutes like Title VII, the Fair Housing Act, and the Equal Credit Opportunity Act already prohibit discrimination in employment, housing, and credit, including practices that have a disparate impact on disadvantaged groups.[11] Our core argument is that when it comes to algorithmic decision systems, the law should explicitly recognize that

introduced in these various works all have slightly different interpretations, and we follow that of model multiplicity in this work.

7. *See, e.g.*, Zhi Chen, Cynthia Rudin, Margo Seltzer, Takuya Takagi, Rui Xin & Chudi Zhong, Exploring the Whole Rashomon Set of Sparse Decision Trees, *in* 35TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 14071, 14071 (Curran Assocs. ed., 2022), https://arxiv.org/abs/2209.08040; Black et al., *supra* note 6, at 850; Marx et al., *supra* note 6, at 6765; Lucas Monteiro Paes, Rodrigo Cruz, Flavio P. Calmon & Mario Diaz, On the Inevitability of the Rashomon Effect, *in* 2023 IEEE INTERNATIONAL SYMPOSIUM ON INFORMATION THEORY (ISIT) 549, 549 (2023), https://ieeexplore.ieee.org/document/10206657.

8. *See* Black et al., *supra* note 6, at 855–56.

9. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 695–96 (2017).

10. *See* Black et al., *supra* note 6, at 853.

11. *See* Civil Rights Act of 1964, 42 U.S.C. § 2000e; Fair Housing Act, 42 U.S.C. § 3601; Equal Credit Opportunity Act, 15 U.S.C. § 1691.

entities relying on them have a legal duty to search for and implement LDAs before deploying a system with disparate effects. Without such a duty, developers are likely to be singularly focused on their chosen performance metric and will fail to identify ways to achieve the same goals with less discriminatory impact.

Our proposal for such a duty is novel because no court has yet addressed the question, yet it finds support in existing legal authorities under the civil rights laws. Because we build our argument on existing doctrine, we confine our argument to employment, housing, and credit—domains that have a clear connection to economic opportunity and justice—and to legally protected characteristics such as race, gender, and age. We do not address related issues, such as the use of algorithmic tools in the criminal legal system[12] or the legal protection of novel algorithmic groups.[13] Although important, they raise distinct concerns and warrant separate analysis.

Recognizing a duty to search for LDAs aligns with the purposes behind our civil rights laws, which were intended to remove arbitrary barriers to full participation by marginalized groups in our nation's economic life.[14] If landlords, employers, or lenders can rely on algorithms that systematically disfavor disadvantaged groups, even when less discriminatory alternatives are available, they will unnecessarily hamper our goal of achieving a society in which resources are more equitably distributed.

Placing the duty to search on entities that rely on algorithmic systems also makes practical sense because model developers are in the best position to undertake a fruitful search. The process of developing a model through the machine learning pipeline is inherently one of exploration,[15] involving iterative cycles of choosing, testing, and recalibrating.[16] Any responsible developer is constantly assessing candidate models to optimize performance and other desired metrics. Our argument is that minimizing disparate impacts should be one of the considerations in the search for the preferred model. In contrast to model developers, individuals who are subjected to these algorithms are especially poorly situated to

12. *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2218 (2019).

13. *See, e.g.*, Sandra Wachter, *The Theory of Artificial Immutability: Protecting Algorithmic Groups Under Anti-Discrimination Law*, 97 TUL. L. REV. 149, 149 (2022).

14. *See* Griggs v. Duke Power Co., 401 U.S. 424, 429–31 (1971) (explaining that Congress's objective in enacting Title VII "was to achieve equality of employment opportunities and remove barriers that have operated in the past to favor an identifiable group of white employees over other employees. . . . What is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification"); United Steelworkers of Am. v. Weber, 443 U.S. 193, 204 (1979) (holding that Title VII is "intended as a spur or catalyst" to eliminate discrimination (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 418 (1975))).

15. Our proposal applies to models learned from data. Though we use the term "machine learning pipeline," this process is not unique to the machine learning pipeline, as it is also the normal statistical modeling pipeline.

16. *See* Lehr & Ohm, *supra* note 9, at 696.

look for LDAs.[17] They will often lack access to basic data necessary to diagnose discriminatory effects, and even when they are aware of such disparities, they are highly unlikely to have the resources—technical, financial, and computational—to meaningfully search for LDAs on their own. Requiring them to identify alternative models would amount to a test of their resource capacity.

To incentivize entities to search for LDAs, the law should recognize this duty in at least two ways. First, under disparate impact doctrine, a defendant's burden of justifying a model with discriminatory effects should be recognized to include showing that it made a reasonable search for LDAs before implementing the model. Second, new regulatory frameworks governing algorithms should require entities to search for and implement less discriminatory models as part of the model building process.

The idea that defendants bear some burden regarding LDAs may seem counterintuitive at first because disparate impact doctrine traditionally associates less discriminatory alternatives with a plaintiff's burden.[18] In the usual account, a plaintiff first establishes a prima facie case of disparate impact, the defendant then must establish a business necessity defense, and finally, the plaintiff can still succeed by showing the existence of a less discriminatory alternative that the defendant refuses to adopt.[19] Under this traditional three-step framework, proving the existence of a less discriminatory alternative appears to be the plaintiff's burden.

However, a close reading of legal authorities over the decades reveals that the existence of a less discriminatory alternative is sometimes relevant to the defendant's burden of establishing a business necessity defense. More specifically, establishing that defense may entail demonstrating that an apparently available, less discriminatory alternative, was not in fact a viable option.[20] Some have argued this imposes an impossible task of proving a negative on the defendant. However, requiring a defendant to consider readily accessible alternative algorithms with less disparate effects does not entail a boundless inquiry. It only requires a showing that they made reasonable efforts to search for and implement identifiable LDAs. Though it may be difficult to identify less discriminatory alternatives in other contexts, such concerns have little force when it comes to alternative algorithms. Precisely because algorithmic systems are evaluated based on quantitative measures of accuracy and other properties, it is far easier to compare alternative algorithms. In other words, unlike in other contexts, it is

17. *See, e.g.*, Kim, *supra* note 3, at 920–21; Selbst, *supra* note 4, at 162–65; Barocas & Selbst, *supra* note 3, at 726; Virginia Foggo & John Villasenor, *Algorithms, Housing Discrimination, and the New Disparate Impact Rule*, 22 COLUM. SCI. & TECH. L. REV. 1, 51–52 (2020).

18. *See Albemarle Paper Co.*, 422 U.S. at 425; Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 527 (2015).

19. *See* Civil Rights Act of 1964, 42 U.S.C. § 2000e-2(k); Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460, 11482 (Feb. 15, 2013).

20. *See, e.g.*, Robinson v. Lorillard Corp., 444 F.2d 791, 798 (4th Cir. 1971); United States v. St. Louis–S.F. Ry. Co., 464 F.2d 301, 308 (8th Cir. 1972); Head v. Timken Roller Bearing Co., 486 F.2d 870, 879 (6th Cir. 1973); Pettway v. Am. Cast Iron Pipe Co., 494 F.2d 211, 244–45 (5th Cir. 1974).

straightforward to determine whether a potential alternative is actually less discriminatory and comparably effective.[21]

Of course, defining the scope of the defendant's duty and what constitutes a reasonable search is not without difficulty, but the law often imposes general duties that are given more specific content by reference to available technologies, industry norms, and interpretive case law. Currently, there are some well-established methods for identifying potential models with equal performance.[22] By deliberately choosing to explore other nearby branches in the machine learning pipeline and testing the resulting models for reduced disparate impact, developers may significantly increase their potential to uncover less discriminatory models.

Drawing from real-world experience and the machine learning literature more generally, it is possible to sketch out some of the steps that a reasonable search would entail. These include collecting or inferring demographic data for disparate impact testing, testing models for disparate impact before deployment and on an ongoing basis, exploring how changes in the model development process might reveal less discriminatory alternatives of equivalent accuracy, and implementing these LDAs in practice. Companies should be expected to dedicate reasonable resources to each step in the process, where reasonableness is determined by the costs of interventions, evidence-based best practices in the relevant industry, and the severity of the disparate impact at issue. And firms should have to document this process and their justification for the point at which they have concluded their search.

A duty to search for LDAs will advance efforts to combat algorithmic discrimination by requiring businesses that rely on algorithmic-decision systems to avoid producing unnecessary disparate impacts. However, our proposal is just one small piece of a broader puzzle. It is not a singular solution, nor is it sufficient to combat algorithmic discrimination. Indeed, some fundamentally flawed algorithms should simply not be implemented, and the availability of LDAs will not remedy them. In many cases, the most effective intervention to reduce unlawful disparate impact may be for businesses to explore non-algorithmic alternatives. Nevertheless, a duty to search for LDAs is a critical part of fulfilling the promise of the civil rights laws.

This Article proceeds as follows. In Part I, we describe what model multiplicity is and why it occurs. We also develop some basic definitions necessary for identifying LDAs. In Part II, we review disparate impact doctrine, explaining the role of less discriminatory alternatives and discussing when the existence of such alternatives has affected the defendant's burden of showing business necessity. Part II also explores how legal authorities have defined what a less discriminatory

21. We limit our proposal to algorithmic systems precisely for this reason. The process of developing non-algorithmic policies and procedures for decisionmaking is comparatively much less structured, with fewer obvious forking paths to explore in parallel. And because some of these policies and procedures may not lend themselves to an obvious metric of evaluation—and to being evaluated according to the same metric—they may not be as readily compared against each other.

22. See *infra* Section V.C for a discussion of these methods.

alternative is. In Part III, we explore the implications of model multiplicity for discrimination law, arguing that entities that use algorithms in the housing, employment, and credit contexts should have a duty to search for and implement LDAs. In Part IV, we offer a real-world case study where a search for an LDA was successful. In Part V, we summarize the relevant technical literature, describing the steps that a firm could take in the model development pipeline to search for LDAs. Also, in Part V, we describe what the duty to search for less discriminatory algorithms should mean in practice. In Part VI, we address some caveats and potential objections. Finally, we conclude.

I. A MULTITUDE OF POSSIBLE MODELS

When discussing algorithms, it is often imagined that, for any given prediction problem, a single correct model exists. Particularly in legal scholarship, scholars often take a "naturalized" view of machine learning,[23] in which the solution to a prediction problem exists, and it is the job of computer scientists to apply objective, technical processes to discover it. When one assumes that there is a unique model that optimizes the goal of, for example, predicting when a borrower will default, it logically follows that pursuing goals like reducing disparate impact will inevitably involve a tradeoff with performance.[24]

But these assumptions about algorithms—that a unique solution exists and that reducing disparate impact involves tradeoffs with performance—are descriptively inaccurate. In fact, even for high-performing models, there are almost always multiple possible models that can reach the best achievable performance for a given prediction problem—a phenomenon computer scientists have termed *model multiplicity*.[25] These alternative models perform equally well for the given prediction task but can also produce different predictions for the same individual. In the aggregate, these equally accurate models can have different impacts across demographic groups. As a result, when an algorithmic system has a disparate impact, developers may be able to discover an alternative model that performs equally well but has less discriminatory impact. But without dedicated exploration, it is unlikely that developers will discover potential LDAs.

Model multiplicity has significant implications for laws and policies that address algorithmic discrimination. We discuss those in Parts II and III. In this Part, we first explore the phenomenon theoretically, explaining why equivalent models with less disparate impact almost always exist for a given prediction problem. Then, we offer a more precise definition of *less discriminatory* algorithms. Finally, we briefly discuss how multiplicitous models are discovered in practice and provide a few concrete examples where model multiplicity led to the discovery of less discriminatory algorithms.[26]

---

23. *See* Lehr & Ohm, *supra* note 9, at 661.

24. *See, e.g.*, Zafar et al., *supra* note 5, at 962; Haas, *supra* note 5, at 7; Wick et al, *supra* note 5, at 8751.

25. Several different terms have been used to describe related phenomena over years of computer science and statistical scholarship. *See supra* note 6.

26. We offer a more detailed study of a particularly illustrative case in Part IV.

A. MODEL MULTIPLICITY

Given the common assumption that a unique algorithmic solution exists for a prediction problem, the statistical literature offers a surprising insight: there are almost always many equally accurate models for a given prediction problem.[27] Recent work has used the term model multiplicity to describe the phenomenon of multiple equally performing models existing for the same prediction task.[28] Despite exhibiting the same accuracy, these models can differ from each other in many other ways. Most importantly for this Article is the possibility that equally accurate models may differ in their individual predictions. When viewed in the aggregate, the difference in these individual predictions may mean that some interchangeable models will have less discriminatory effect.

To make the point concrete, consider a bank that uses an algorithmic system to make loan decisions.[29] The prediction task can be defined simply: classify individual applicants as creditworthy or not. Because creditworthiness cannot be directly measured, the bank uses some other measure—for example, the likelihood of nonpayment after 6 months—as a proxy.[30] Model multiplicity means that a machine learning process could develop multiple models that exhibit the same overall accuracy in predicting nonpayment, but which make different individual predictions. For example, while one model might predict that a Person X will not default, an alternative model, which exhibits the same accuracy as the first, might predict that Person X *will* default.

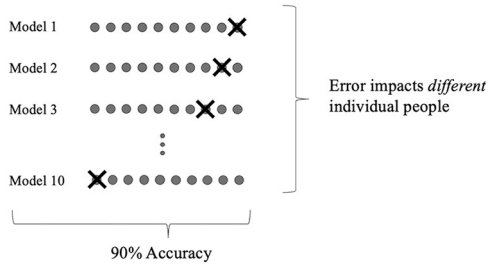
27. *See, e.g.*, Chen et al., *supra* note 7, at 14071; Black et al., *supra* note 6, at 850; Marx et al., *supra* note 6, at 6765; Hsiang Hsu & Flavio P. Calmon, Rashomon Capacity: A Metric for Predictive Multiplicity in Classification, *in* 36TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 28988, 28988 (Curran Assocs. eds., 2022).

28. *See supra* note 6.

29. For ease of explanation, we use the example of a binary classification problem, although the phenomenon of model multiplicity applies to other kinds of predictions as well.

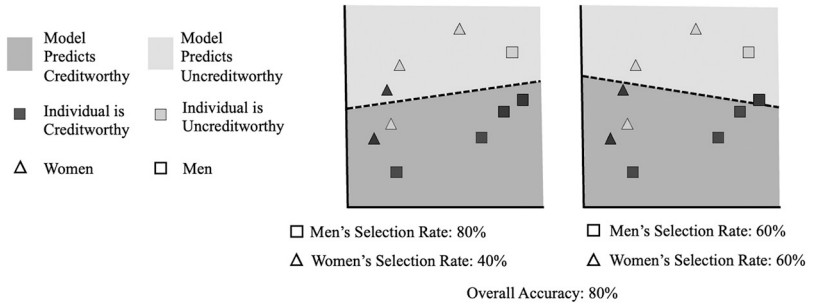
30. *See* Samir Passi & Solon Barocas, Problem Formulation and Fairness, *in* FAT* '19: PROCEEDINGS OF THE 2019 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 39, 40 (Ass'n for Computing Mach. eds., 2019) (illustrating the difficulties of defining a "good" employee in sales, and forcing algorithms to use concrete data such as applicants' sale figures instead of personality).

Figure 1. An illustration of how ten different models can exhibit the same accuracy while giving different individual predictions on a hypothetical group of ten people.



Model multiplicity exists because, for any given error rate, there are different ways to distribute accurate predictions over a population. Consider a model, such as in Figure 1, built on a dataset that contains ten different people. If the minimum achievable error rate is 10 percent, then the model will misclassify one person. As Figure 1 demonstrates, this means there can be ten different models that exhibit a 10% error rate, but make different predictions, each misclassifying a *different* person in the dataset. Over larger datasets, the number of potential combinations multiplies. By distributing errors differently, equivalent models can have differing effects across different demographic groups. Consider, for example, two models predicting outcomes for a population containing equal numbers of men and women, as Figure 2 illustrates. Both models exhibit 80 percent accuracy, but one selects more men than women, whereas the other selects men and women at the same rate. The two models display the same overall performance but differ in the disparity between the selection rates for the two groups. Overall, model multiplicity means that the distribution of outcomes across different groups can vary between equivalently accurate models, such that certain distributions of error lead to less disparate impact than others.

Figure 2. An example of two multiplicitous models displaying equal error rates overall yet differing in their disparities in selection rates across groups. The graph to the left has a steep difference in selection rate between men and women, whereas the graph to the right does not. The darker region of the graph refers to places where the model predicts an individual to be creditworthy, and darker points correspond to individuals who are indeed creditworthy. The lighter region of the graph refers to areas where the model predicts an individual to be uncreditworthy, and lighter points correspond to individuals who are indeed uncreditworthy.[31]



Thus, so long as there is imperfect accuracy in a given prediction algorithm,[32] model multiplicity guarantees that another model with similar overall accuracy exists that would generate predictions differently, and may reduce disparities across groups. Models used in domains covered by discrimination law commonly achieve levels of accuracy that are far from perfect, leaving a good deal of

31. Some formulations of linear models, such as ordinary least squares regression (OLS), can be expressed as a unique solution to an equation, and this fact may seem to be directly at odds with the concept of model multiplicity, especially as presented in our figure. However, even in the case of model types with a unique solution, differences in the setup of how the model is developed (i.e., the model development pipeline, which we discuss in the next subsection) can still lead to a variety of solutions. For example, variations in how the available data are partitioned into training data (which will be inputs to the equation with a unique solution) and test data (which will not), or, as Marx et al. focus on, differences in the choice of weight of regularization parameters, can lead to a sizeable amount of variance in individual predictions across models with a unique solution and access to the same model development resources (i.e., overall data available). *See* Marx et al., *supra* note 6, at 6771–72.

32. As a theoretical matter, when developing a model with an infinite amount of training data, there exists only one optimal model—the so-called Bayes optimal model. In practice, though, models are always trained on a finite amount of data and so never achieve the accuracy of the Bayes optimal model. This gap leaves open the possibility that there is more than one model with the best achievable accuracy in practice. *See* Black et al., *supra* note 6, at 850, 853. In fact, models used in domains covered by discrimination law can be very far from the Bayes optimal model. *See* Christo Wilson, Alan Mislove, Avijit Ghosh & Shan Jiang, *Auditing the Pymetrics Model Generation Process* 1, 17–18 (2020), https://cbw.sh/static/audit/pymetrics/pymetrics_audit_result_whitepaper.pdf.

error to shuffle around in the service of reducing disparities in selection rates.[33] Unless a company has fortuitously discovered the model with minimum possible disparity among all equivalent models,[34] the phenomenon of model multiplicity almost always[35] means that there exists a model with indistinguishable accuracy but less disparate impact—i.e., a less discriminatory algorithm.[36]

B. DEFINING LESS DISCRIMINATORY ALGORITHMS

Drawing on the insights of model multiplicity, we define here the concept of a *less discriminatory algorithm*, or LDA. An LDA has two critical features. First, the alternative algorithm must be less discriminatory than a given baseline model. And second, the proposed algorithm should be comparable or equivalent to the baseline model in performance.

We define "discriminatory" by reference to existing civil rights laws. Specifically, we focus on discrimination based on legally protected characteristics like race and sex,[37] leaving aside debates about whether other forms of systemic disadvantage should also be forbidden. Apart from who is protected, the concept of discrimination also requires a definition of what discrimination *is*. Although computer scientists have advanced a variety of formal definitions,[38] we rely on

33. *See* Wilson et al., *supra* note 32, at 17–18.

34. Of course, if a model is perfectly accurate, there are no errors that can be distributed differently, and so there are no less discriminatory alternatives. Such a model is extremely unlikely to exist.

35. In addition to the increasing work stream drawing attention to, and providing evidence for, model multiplicity, recent work has begun to provide methods for estimating the size of the set of equally accurate models for a given prediction problem. *See* Marx et al., *supra* note 6, at 6771–72; Lesia Semenova, Cynthia Rudin & Ronald Parr, On the Existence of Simpler Machine Learning Models, *in* FACCT '22: PROCEEDINGS OF THE 2022 5TH ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, *supra* note 6; Hsu & Calmon, *supra* note 27, at 28994. The experimental results of these papers consistently provide evidence for the fact that many prediction problems do admit a large set of equally effective models. *See, e.g.*, Semenova et al., *supra*, at 1827–58. Given this evidence, it is unlikely that a model practitioner would happen upon the least discriminatory model in a large set of equally effective models if they were not expressly searching for such a model—an argument that has been made and experimentally verified in the case of robustness and explainability. *See, e.g.*, *id.*; Alexander D'Amour, Katherine Heller & Dan Moldovan et al., *Underspecification Presents Challenges for Credibility in Modern Machine Learning*, J. MACH. LEARNING. RSCH., 2022 at 1, 3, https://arxiv.org/abs/2011.03395. Thus, we use the term "almost always" to connote the idea that, given evidence that the set of equivalently accurate models is *often* large and that the chance of finding the least discriminatory model within them is vanishingly small, there is *almost* always a less discriminatory model available.

36. As we discuss in Sections II.C and VI.B, "equally effective" need not be the standard against which less discriminatory alternatives are assessed. Nevertheless, as we hope to show, even under this more restrictive standard, developers have substantial room to discover LDAs.

37. Federal law prohibits discrimination in employment based on race, color, national origin, religion, sex, sexual orientation, and gender identity (Title VII), as well as age (ADEA) and disability (ADA). In the housing context, the Fair Housing Act (FHA) prohibits discrimination based on race, color, national origin, religion, sex, familial status, and disability. In the credit context, the Equal Credit Opportunity Act (ECOA) prohibits discrimination based on race, color, national origin, sex, sexual orientation, gender identity, marital status, age, and receipt of public assistance.

38. *See, e.g.*, Pratyush Garg, John Villasenor & Virginia Foggo, Fairness Metrics: A Comparative Analysis, *in* 2020 IEEE INTERNATIONAL CONFERENCE ON BIG DATA 3662 (2020), https://arxiv.org/abs/2001.07864; Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour & Kristan Lum,

existing legal theories and, in particular, disparate impact doctrine, which scrutinizes disparities in selection rates that systematically disadvantage marginalized groups. A disparity in selection rates occurs when the rate of positive outcomes differs between groups—for example, when a model used for lending decisions approves a lower proportion of Black than white applicants or a model used to screen resumes recommends greater proportions of male than female candidates. An algorithm is less discriminatory compared with another if it results in a meaningful reduction in disparity in selection rates between groups.

Second, a less discriminatory alternative algorithm must have performance equivalent to a baseline model. Comparing model performance requires defining what metric of performance is relevant. We generalize here and define model performance as whatever metric the model is trained to optimize that is appropriate for the given context. For example, the performance of a lending model might be measured by its accuracy in predicting the likelihood that a borrower will default. Depending on the context, developers might choose to optimize a variety of different metrics when training a model, but for the sake of simplicity, we will use accuracy as a stand-in for any more specific performance metric.[39]

Identifying models that perform equally well also requires determining to what degree models can differ from one another along the relevant performance metric and still be considered equivalent. In other words, it is necessary to decide on a threshold level of difference in performance beyond which models are considered meaningfully different. In the model multiplicity literature, equivalent accuracy refers to levels of accuracy that are "functionally indistinguishable" (e.g., accuracy rates of 97.8989 and 97.8990).[40] What differences are considered meaningful in practical settings will depend upon what is being measured and how model outputs are used to make real-world decisions. For example, when models appear interchangeable as an operational matter from the perspective of the organization deploying them, they should be considered equivalent in performance. We refer to this context-specific threshold, or bound, as ε ("epsilon"), such that models whose performance is within ε are considered equivalent.

There may also exist alternative models that would significantly reduce disparate impact but whose performance falls somewhat outside ε, being either more or less accurate than the baseline model. These alternatives should also be

Algorithmic Fairness: Choices, Assumptions, and Definitions, *in* 8 ANNUAL REVIEW OF STATISTICS AND ITS APPLICATION 141 (2021).

39. For example, different notions of model performance must be used if the model's prediction is real-valued (e.g., predicting amount of tax avoidance), such as mean squared error (MSE); binary (e.g., predicting whether or not an individual's resume should progress to the next level in screening), such as accuracy; or real-valued between zero and one (e.g., predicting risk of credit default), such as area under the curve (AUC). Even for models that give predictions of the same type, there are many different metrics of performance that are used in practice. For example, with binary predictions, developers routinely use accuracy, precision, recall, and F1 score, among others, to measure performance. Multiplicity applies under all of these circumstances. Whatever the relevant notion of model performance or chosen performance metric, it will be possible to leverage multiplicity to find models of equivalent performance, but with less disparate impact.

40. Black et al., *supra* note 6, at 851.

considered legally relevant, and in many cases, it may make sense to adopt them. However, because this Article focuses on the implications of model multiplicity, we purposefully define LDAs narrowly such that they do not encompass these more accurate and less accurate models. In doing so, we do not foreclose arguments that other more and less accurate models that reduce disparities should sometimes also be legally required.

C. MODEL MULTIPLICITY IN PRACTICE

While model multiplicity exists in theory, how can equally accurate models be discovered in practice? Finding a *specific* equally accurate model—corresponding to a particular re-drawing of a model's decision boundary[41]—is difficult through the typical model development process.[42] At the same time, recent research has shown that models with equivalent accuracy that differ in other behavior (including disparate impact) can be easily discovered in practice and occur naturally throughout the model development process.[43] Likewise, because we cannot explore the infinite number of ways to develop models that all achieve a certain level of performance, identifying the *least* discriminatory model from this set is often not possible in practice. Yet, with some effort, a model that is *less* discriminatory than a baseline model can almost certainly be found in practice.

1. Searching Through the Pipeline

Machine learning systems are built through a series of iterative decisions, often called the machine learning pipeline. These decisions often involve subjective choices for which there is often no correct answer, requiring the developer to weigh competing values and make judgment calls. However, each decision can substantially affect the behavior of the resulting model.

---

41. In other words, a decision boundary could be drawn to achieve similar accuracy while distributing error differently to reduce disparate impact. *See* Black et al., *supra* note 6, at 853.

42. It is generally very hard to predict exactly how machine learning models will draw their decision boundaries (e.g., the line separating loan applicants who are predicted to repay from those who are predicted to default). These decision boundaries are contingent on the particularities of the training data, which may be sufficiently complex or subtle that the range of possible decision boundaries would not be obvious to developers. It is even harder to control exactly where a model ends up drawing its decision boundary because there may not be a clear relationship between changes that could be made in the model development process and the ultimate location of the boundary. However, as we explain in this section, the model development process on its own can naturally lead to several different models that all exhibit equivalent performance, some of which will have less disparate impact than others. *See, e.g.*, Hamid Karimi & Jiliang Tang, Decision Boundary of Deep Neural Networks: Challenges and Opportunities, *in* WSDM '20: PROCEEDINGS OF THE 13TH INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING 919 (Ass'n for Computing Mach. eds., 2020); Yu Li et al., On the Decision Boundary of Deep Neural Networks 10 (2019), https://arxiv.org/abs/1808.05385.

43. *See, e.g.*, Kit T. Rodolfa, Hemank Lamba & Rayid Ghani, *Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy*, 3 NATURE MACH. INTEL. 896, 902 (2021); Alexander D'Amour et al., *supra* note 35, at 3; Jiayun Dong & Cynthia Rudin, Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models 1 (2020), https://arxiv.org/abs/1901.03209; Amanda Coston, Ashesh Rambachan & Alexandra Chouldechova, Characterizing Fairness Over the Set of Good Models Under Selective Labels, *in* 139 PROCEEDINGS OF THE 38TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 2144, 2144–45 (Proc. Mach. Learning & Rsch., 2021), https://arxiv.org/pdf/2101.00352 [https://perma.cc/QH2P-PFGN].

The model development pipeline has been described by a variety of scholars across disciplines[44] so we offer only a high-level overview here. The pipeline consists of problem formulation, data collection, data preprocessing, feature selection, statistical modeling, testing and validation, and deployment and monitoring. Each step presents opportunities for practitioners to make decisions that lead to slightly different eventual models, each with different behavior.

One can imagine the model creation process as a tree, with every choice represented by a branch that leads to different sets of smaller and smaller branches until they reach the leaves—the individual models. Choices made early in the process—for example, what input features to use—cause the exploration process to flow down one set of branches, leaving large portions of the tree unexplored. Smaller choices, such as how many times the model goes through the training data to learn patterns from them (known as the number of "epochs"), further narrow the set of models under consideration.

Rather than making each decision and moving forward only on that branch, developers can identify a vast array of other models by instead exploring different branches along this tree. Not all of these potential models will have equivalent performance (i.e., performance within ε). But as research has proven, many will.[45] And of these equally accurate models, it is very likely that several of them will have less disparate impact.

The method of searching for LDAs in practice is identical to that of identifying multiplicitous models—i.e., exploring a wider range of models made through various decisions across the machine learning pipeline—with the addition of testing for disparity as well as accuracy. Although expanding the exploration process during the machine learning pipeline helps to uncover multiplicitous models, not every theoretically possible, equivalently accurate model is easy to discover in practice. Though it may not be possible to discover the *least* discriminatory algorithm, or every potential LDA, a search is extremely likely to discover some LDAs.

As we note in Part V, increased exploration through the machine learning pipeline will not only yield equally accurate and less discriminatory models but also models that differ in performance—in particular, models that are more

44. *See, e.g.*, Harini Suresh & John Guttag, A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle, *in* EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION 1, 2–4 (2021); David Lehr & Paul Ohm, *supra* note 9, at 670–701; Nil-Jana Akpina, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu & Hoda Heidari, A Sandbox Tool to Bias(Stress)-Test Fairness Algorithms 17 (2022), http://arxiv.org/abs/2204.10233; Black et al., *supra* note 6, at 853–54.

45. *See, e.g.*, Coston et al., *supra* note 43, at 2145; Paes et al., *supra* note 7, at 553; Chen et al., *supra* note 7.

accurate and less discriminatory and models that are less accurate and less discriminatory.[46]

2. Practical Examples of Searching for LDAs

Practical examples of discovering less discriminatory models exist. In one research setting, Coston et al. developed a tool to test the accuracy and various notions of disparity over a range of models developed by randomizing elements of the modeling process.[47] They identified alternative models to the baseline model that have equivalent accuracy (within 1%), yet have lower selection rate disparity across racial groups by over 10%.[48] While their tool does not search over the entire pipeline (i.e., does not explore the entire tree, but instead only one juncture), the work shows that searching across the pipeline can lead to models with similar accuracy but reduced disparity.[49]

Real-world practice also demonstrates that less discriminatory models can be discovered. One example is the Monitorship of Upstart, a financial technology company that relies on machine learning and non-traditional applicant data, including data related to borrowers' higher education, to underwrite and price consumer loans.[50] After civil rights groups raised concerns that Upstart's underwriting model might be racially discriminatory, the company agreed to allow an independent Monitor to assess its algorithm.[51] Testing of Upstart's model showed a racially disparate impact on Black borrowers as compared to non-Hispanic white borrowers, leading the Monitor to explore the availability of alternative models.[52] We discuss this example in greater detail in Part IV. The upshot is that the Monitor was able to identify multiple models that reduced disparate impact while still performing comparably to the original model.[53]

---

46. Exploration may also reveal models that reduce the relative disadvantage suffered by one group, while exaggerating the relative disadvantage suffered by another. For example, it may be possible to discover a model that reduces the disparities between white and Black loan applicants but that exaggerates the disparity between white and Hispanic loan applicants, or that reduces the disparity between racial groups while exaggerating the disparity between gender groups. For now, we leave these concerns aside, but return to them in Part IV, which offers a real-world example of how to deal with this challenge, and Part V, which addresses the degree to which attempts to minimize disparities across multiple, potentially intersectional groups may limit developers' ability to find LDAs.

47. *See* Coston et al., *supra* note 43, at 2150.

48. *Id.*

49. *See id.* at 1252. This research focused on a criminal risk assessment instrument. We surface this example because of the underlying phenomenon at issue—a practical example of discovering a less discriminatory model through dedicated exploration. Though the example involves a criminal risk assessment instrument, our argument does not reach the use of these tools in the criminal legal system. As previously noted, we believe those tools and that system raise distinct concerns and deserve separate analysis.

50. *See* RELMAN COLFAX PLLC, FAIR LENDING MONITORSHIP OF UPSTART NETWORK'S LENDING MODEL: INITIAL REPORT OF THE INDEPENDENT MONITOR 3 (2021), https://www.relmanlaw.com/media/cases/1088_Upstart%20Initial%20Report%20-%20Final.pdf [https://perma.cc/SZ2K-2TBN].

51. *See id.*

52. *See id.* at 13, 23–24; RELMAN COLFAX PLLC, FAIR LENDING MONITORSHIP OF UPSTART NETWORK'S LENDING MODEL: THIRD REPORT OF THE INDEPENDENT MONITOR 24–25 (2022), https://www.relmanlaw.com/media/cases/1333_PUBLIC%20Upstart%20Monitorship%203rd%20Report%20FINAL.pdf [https://perma.cc/C75T-VMPG].

53. RELMAN COLFAX PLLC, *supra* note 52, at 24–25.

These examples suggest that the theoretical guarantee of model multiplicity translates into practice. Through purposeful, broader exploration in the model-development pipeline, developers can find models with indistinguishable performance that have reduced levels of disparate impact.

However, finding an LDA with large reductions in disparate impact is not a guarantee.[54] There are three main reasons for this. First, there are technical limits to the extent to which error can be redistributed to reduce disparity. For example, if the baseline model has a low error rate and a significant disparate impact, there is only so much of the disparate impact that can be reduced through exploiting model multiplicity. However, research has shown that in many circumstances relevant to civil rights law, models exhibit high error rates and high disparity, which may give developers sufficient room to redistribute error to meaningfully reduce disparate impact.[55] Second, it is not clear a priori which interventions will have the greatest impact on reducing disparities, and thus how to perform a search for LDAs most efficiently. Recall, however, that the examples discussed above were successful *despite* their relatively limited searches—that is, they did not explore the entire pipeline to search for LDAs—which suggests that knowledge of the ideal path on which to intervene is not always necessary for a successful search.[56] Third, finding an LDA requires that developers specifically allocate resources to the task—that is, to exploration along the modeling pipeline. Although the available techniques vary in cost, many changes that would be trivial for most developers to integrate into their model development processes have proven very useful in finding LDAs.[57]

3. Joint Optimization of Fairness and Performance

At this point, some readers may wonder why we focus on exploring different branches along the machine learning pipeline rather than integrating concerns with fairness directly into the optimization process. Over the past decade, a robust literature has developed around "algorithmic fairness,"[58] much of which focuses on how to jointly

54. *See, e.g.*, Coston et al., *supra* note 43, at 2144–55; Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi & Cédric Archambeau, Fair Bayesian Optimization, *in* PROCEEDINGS OF THE 2021 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 854, 854–63 (Ass'n for Computing Mach. eds., 2021), https://arxiv.org/abs/2006.05109.

55. *See* Manish Raghavan, Solon Barocas, Jon Kleinberg & Karen Levy, Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices, *in* FAT '20: PROCEEDINGS OF THE 2020 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 469, 470–71, 478 (Ass'n for Computing Mach. eds., 2020), https://arxiv.org/abs/1906.09208.

56. This may suggest that it is even more likely that LDAs could be found in practice if more of the pipeline were explored.

57. *See* Perrone et al., *supra* note 54, at 854.

58. *See, e.g.*, Dana Pessach & Erez Shmueli, *A Review on Fairness in Machine Learning*, ACM COMPUTING SURVEYS, Feb. 2022, at l, 2; Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman & Aram Galstyan, *A Survey on Bias and Fairness in Machine Learning*, ACM COMPUTING SURVEYS, July 2021, at 1, 2, https://arxiv.org/abs/1908.09635.

optimize for two goals: minimizing disparate impact and maximizing performance.[59] Because the optimization process is automated, these approaches have the apparent advantage of automating the search for models that obtain the best possible accuracy while limiting disparate impact.

Despite the intuitive appeal of these approaches, we chose not to focus on them for two reasons. First, joint optimization only targets one part of the machine learning pipeline—the optimization process—where developers could intervene to find LDAs. There are many other points in the development process where alternative choices could help to reduce disparate impact with no practical effect on performance, ranging from adjusting the problem formulation to feature selection to hyperparameter tuning.[60] Thus, while developers should adopt methods for joint optimization when appropriate, using such techniques will not exhaust the many other ways that developers can find LDAs and may cause developers to forgo models with even less disparate impact. Our proposal should thus not be understood as a simple call for developers to jointly optimize for fairness and performance.

Second, joint optimization can raise unresolved legal questions. To jointly minimize disparate impact and maximize performance during model training, developers will often require access to and use of protected characteristics. How, when, and under what circumstances such demographic data can be used in contexts covered by civil rights law is a point of debate in the legal scholarship, and scholars disagree as to when certain uses would introduce disparate-treatment concerns.[61] Notably, however, the kind of broad pipeline search for LDAs that we call for does not require the use of demographic data during model training. Instead, this data would only be used for testing and evaluation of varying possible models—methods that should not trigger disparate-treatment concerns.[62] In short, by broadening the model intervention aperture beyond joint optimization methods, developers may not only be able to find LDAs that they would otherwise miss, but may also face less uncertainty as to the legality of their interventions.

We elaborate on concrete methods for conducting the search for LDAs in Part V, but first, we turn to the legal landscape regarding less discriminatory alternatives and the implications of model multiplicity for law.

---

59. *See, e.g.*, Emily Black, Rakshit Naidu, Rayid Ghani, Kit Rodolfa, Daniel Ho & Hoda Heidari, Toward Operationalizing Pipeline-Aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools, *in* PROCEEDINGS OF 2023 ACM CONFERENCE ON EQUITY AND ACCESS IN ALGORITHMS, MECHANISMS, AND OPTIMIZATION, art. 36 (Ass'n for Computing Mach. eds., 2023), https://dl.acm.org/doi/10.1145/3617694.3623259.

60. *Id.*

61. *See, e.g.*, Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 825–41, 852 (2020); Daniel E. Ho & Alice Xiang, *Affirmative Algorithms: The Legal Grounds for Fairness as Awareness*, U. CHI. L. REV. ONLINE 134, 136 (2020); Pauline T. Kim, *Race-Aware Algorithms: Fairness, Nondiscrimination, and Affirmative Action*, 110 CALIF. L. REV. 1539 (2022).

62. *See* Kim, *supra* note 61, at 1574–83.

II. DISPARATE IMPACT DOCTRINE AND LESS DISCRIMINATORY ALTERNATIVES

Before discussing the implications of model multiplicity for the law in Part III, this Part explains the relevant legal landscape, describing disparate impact doctrine and exploring the role of less discriminatory alternatives in that framework. This review helps us establish two key points. First, while many take as a given that it is a plaintiff's burden to demonstrate the existence of less discriminatory alternatives, some legal authorities also suggest that such alternatives are relevant to a defendant's burden of justification—specifically, showing that alternatives with less discriminatory impact are unavailable. Second, while the cases are not entirely consistent, existing authority suggests that a viable alternative to the challenged practice does not have to be exactly equally effective and may entail some additional costs to the defendant.

A. THE DISPARATE IMPACT FRAMEWORK

Existing civil rights laws prohibit discrimination in employment, housing, and credit. Title VII of the Civil Rights Act of 1964 (Title VII)[63] prohibits discrimination in employment; the Fair Housing Act (FHA)[64] similarly forbids discrimination when renting or buying a home, getting a mortgage, or seeking housing assistance; and the Equal Credit Opportunity Act (ECOA)[65] outlaws discriminatory financial practices. Each law prohibits not only disparate treatment, which emphasizes discriminatory intent, but practices that have a disparate impact as well. Disparate treatment might apply to algorithms in some circumstances.[66] However, because the concept of intent does not fit easily with algorithms, and many algorithms appear to be facially neutral, much of the legal analysis has focused on disparate impact.[67]

In broad strokes, a disparate impact case unfolds as follows. First, a plaintiff must establish a prima facie case by showing that a policy or practice has a disparate impact on a disadvantaged group. Second, the defendant has the burden of demonstrating a legitimate business justification for the practice. Third, even if a defendant meets its burden, it can still face liability if the plaintiff can show an alternative that would serve the same ends with less disparate impact. Thus, if an algorithm disproportionately screens out members of a disadvantaged group, it triggers legal scrutiny, and the practice would be considered discriminatory if the defendant was unable to prove a business justification for its use.

Disparate impact doctrine was first articulated in *Griggs v. Duke Power Co.*,[68] which involved a challenge to employment practices under Title VII. The

63. 42 U.S.C. § 2000e.

64. 42 U.S.C. §§ 3601–19.

65. 15 U.S.C. § 1691.

66. *See* Barocas & Selbst, *supra* note 3, at 710; Matthew U. Scherer, Allan G. King & Marko J. Mrkonich, *Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms*, 71 S.C. L. REV. 449, 496–99 (2019).

67. *See, e.g.*, Barocas & Selbst, *supra* note 3, at 673; Kim, *supra* note 3, at 920–21; Selmi, *supra* note 3, at 634–44.

68. 401 U.S. 424, 431 (1971).

employer's minimum education and testing requirements had the effect of disproportionately screening out Black workers from more desirable jobs.[69] The Supreme Court held that facially neutral practices with a disparate effect on a disadvantaged minority group violated Title VII unless justified by business necessity.[70] The Court reasoned that the civil rights laws are directed at "the consequences of employment practices, not simply the motivation,"[71] and therefore, proof of intent is not necessary to establish liability.[72]

Shortly after *Griggs*, lower federal courts began to apply its reasoning in housing discrimination cases. Eventually, every circuit that considered the issue agreed that the FHA prohibits disparate impact discrimination[73]—a conclusion confirmed by the Supreme Court in *Inclusive Communities* in 2015.[74] Similarly, agency interpretation of the ECOA has followed the reasoning in *Griggs* and adopted the disparate impact theory in the credit context.[75] Although the Supreme Court has not squarely addressed the issue, courts across the country have consistently found that disparate impact claims are cognizable under the ECOA.[76]

Although the three-step analysis outlined above applies across the relevant statutes, they each define the burden of justification on the defendant in the second step of the analysis differently. Title VII requires a defendant to demonstrate that the challenged practice "is job related" and "consistent with business necessity."[77] With respect to employee selection tools, like tests, this typically means

---

69. *Id*. at 427–28.

70. *See id*. at 431.

71. *Id*. at 432 (emphasis omitted).

72. *See id*.

73. *See, e.g*., Huntington Branch, NAACP v. Town of Huntington, 844 F.2d 926, 935–36 (2d Cir. 1988); Resident Advisory Bd. v. Rizzo, 564 F.2d 126, 146 (3d Cir. 1977); Smith v. Town of Clarkton, N.C., 682 F.2d 1055, 1065 (4th Cir. 1982); Hanson v. Veterans Admin., 800 F.2d 1381, 1386 (5th Cir. 1986); Arthur v. City of Toledo, 782 F.2d 565, 574–75 (6th Cir. 1986); Metro. Hous. Dev. Corp. v. Vill. of Arlington Heights, 558 F.2d 1283, 1290 (7th Cir. 1977); United States v. City of Black Jack, 508 F.2d 1179, 1184–85 (8th Cir. 1974); Halet v. Wend Inv. Co., 672 F.2d 1305, 1311 (9th Cir. 1982); United States v. Marengo Cnty. Comm'n, 731 F.2d 1546, 1559 n.20 (11th Cir. 1984).

74. Tex. Dep't of Hous. & Cmty. Affs. v. Inclusive Cmtys. Project, Inc., 576 U.S. 519, 545–46 (2015).

75. Regulation B, 12 C.F.R. § 1002.2 (2023).

76. *See* Golden v. City of Columbus, 404 F.3d 950, 963 n.11 (6th Cir. 2005) ("Neither the Supreme Court nor this Court have previously decided whether disparate impact claims are permissible under ECOA. However, it appears that they are."); Miller v. Am. Exp. Co., 688 F.2d 1235, 1239–40 (9th Cir. 1982) ("The ECOA's history refers by analogy to the disparate treatment and adverse impact tests for discrimination which are used in employment discrimination cases under Title VII."); Haynes v. Bank of Wedowee, 634 F.2d 266, 269 n.5 (5th Cir. 1981) ("ECOA regulations endorse use of the disparate impact test to establish discrimination . . . ."); Smith v. Chrysler Fin. Co., No. Civ.A. 00-6003, 2003 WL 328719, at *6 (D.N.J. Jan. 15, 2003) ("It is clear from the above language that disparate impact theory is present in the ECOA . . . ."); Powell v. Am. Gen. Fin., Inc., 310 F. Supp. 2d 481, 487 (N.D.N.Y. 2004) ("The ECOA provides for a private cause of action based on disparate impact or disparate treatment."); Palmer v. Homecomings Fin., LLC, 677 F. Supp. 2d 233, 240 (D.D.C. 2010) ("There appears to be agreement among the federal courts that disparate impact claims are permissible under the ECOA."); Osborne v. Bank of Am., Nat'l Ass'n, 234 F. Supp. 2d 804, 812 (M.D. Tenn. 2002) (allowing plaintiffs to proceed on a disparate impact claim under the ECOA).

77. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

that a firm's selection procedures must be "validate[d]"—i.e., demonstrated through appropriate statistical analysis to actually measure job-relevant abilities or characteristics.[78] Under the FHA, the defendant's burden is to show that its practice is "necessary to achieve one or more substantial, legitimate, nondiscriminatory interests."[79] And pursuant to the ECOA, a defendant must establish that its practice or policy "meets a legitimate business need."[80]

Regardless of how the requirement of business justification is formulated, the existence of less discriminatory alternatives matters under all three statutes when determining whether a defendant is liable for a practice that produces disparate effects. Because model multiplicity implies that LDAs almost always exist when a model has disparate effects, we explore in detail the concept of alternative practices in disparate impact doctrine. In the next two Sections, we focus on how existing law addresses two important questions: Who bears the burden of showing the existence of less discriminatory alternatives? And what exactly is a less discriminatory alternative?

B. WHO BEARS THE BURDEN?

The availability of less discriminatory alternatives is clearly relevant to the third step of disparate impact analysis, where the plaintiff has the burden of proof. This Section explains that some legal authorities have found that the existence of alternatives also bears on the *defendant's* burden of justification at the second step.

After the Supreme Court's decision in *Griggs*, some lower federal courts in the early 1970s found that less discriminatory alternatives were relevant to the *defendant's* burden of justification. The leading case was *Robinson v. Lorillard Corp.*, in which the Fourth Circuit found that proving business necessity entailed showing that there are "no acceptable alternative policies or practices which would better accomplish the business purpose advanced, or accomplish it equally well with a lesser differential racial impact."[81] Other circuits followed *Lorillard* in holding that an inquiry into available alternatives was part of the defendant's burden of showing business necessity.[82]

The Supreme Court signaled a different course in a 1975 decision, *Albemarle Paper Co. v. Moody*, which examined whether an employer had properly validated a general intelligence test used to screen employees.[83] In passing, the Court

78. *See* 29 C.F.R. § 1607.14(B).

79. 24 C.F.R. § 100.500(c)(2).

80. 12 C.F.R. § 1002.6(a)-2.

81. 444 F.2d 791, 798 (4th Cir. 1971).

82. The Eighth Circuit adopted *Lorillard's* "no alternatives" framework in 1972. United States v. St. Louis-S.F. Ry. Co., 464 F.2d 301, 308 (8th Cir. 1972). The Sixth Circuit adopted this framework in 1973. Head v. Timken Roller Bearing Co., 486 F.2d 870, 879 (6th Cir. 1973) ("The court's error arose from its failure to take into account one necessary element of the [*Lorillard*] test."). In 1974, the Fifth Circuit held that the "nature and requirements of th[e] [business necessity] burden were correctly outlined in *Robinson v. Lorillard Corp.*" Pettway v. Am. Cast Iron Pipe Co., 494 F.2d 211, 244–45 (5th Cir. 1974).

83. 422 U.S. 405, 410–11, 425 (1975).

also described a new, third step in the disparate impact framework: if an employer met its burden of proving job-relatedness, "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." [84] The Court did not discuss this possibility further because it separately concluded that the tests were not job related and therefore unlawful,[85] and so there was no need to consider the role of less discriminatory alternatives.[86]

Even though the description of a third step was dicta, *Albemarle*'s three-step process nonetheless became the standard framework for analyzing Title VII disparate impact cases. Eventually, Congress codified disparate impact doctrine in the Civil Rights Act of 1991, adopting the *Albemarle* framework, albeit in an oblique way. As amended, Title VII says that disparate impact is established if the plaintiff shows that a practice has a disparate impact and the defendant fails to demonstrate business necessity.[87] Alternatively, the practice is unlawful if the plaintiff makes a demonstration "in accordance with the law as it existed on June 4, 1989, with respect to the concept of 'alternative employment practice'"[88] and the defendant "refuses to adopt such alternative employment practice."[89]

This odd formulation reflects a history of disagreement over the third step of the framework. On June 5, 1989, the Supreme Court decided *Wards Cove Packing Co. v. Atonio*,[90] which made it significantly more difficult for plaintiffs to prevail in disparate impact cases, including by placing a heavier burden on plaintiffs to demonstrate a less discriminatory employment practice.[91] The Civil Rights Act of 1991 was intended to undo the effect of *Wards Cove* and several other Supreme Court decisions that had narrowly interpreted civil rights statutes.[92] However, rather than clearly articulating what it means to demonstrate a less discriminatory alternative practice, Congress simply set the law back to the day before the *Wards Cove* decision. So, what was the state of less discriminatory

84. *Id.* at 425.

85. *Id.* at 425–26, 435–36.

86. The Court's brief discussion of the new third step was also curious because the Court suggested that a showing that less discriminatory practices were available "would be evidence that the employer was using its tests merely as a 'pretext' for discrimination," citing its earlier decision in *McDonnell Douglas Corp. v. Green*, 411 U.S. 792, 802 (1973). *Albemarle*, 422 U.S. at 425. But *McDonnell Douglas* was a *disparate treatment* case that dealt with the proof structure in cases involving intentional discrimination. 411 U.S. at 799–800. *Albemarle*'s reliance on *McDonnell Douglas* and reasoning about pretext conflated the disparate treatment and disparate impact theories. By definition, disparate impact cases are about discriminatory effects, not intent, and so proving "pretext" would seem to be beside the point.

87. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

88. 42 U.S.C. § 2000e-2(k)(1)(C).

89. 42 U.S.C. § 2000e-2(k)(1)(A)(ii).

90. 490 U.S. 642 (1989).

91. *See id.* at 660. The majority in *Wards Cove* also changed the defendant's burden of proving business necessity to a burden of production and put the burden of proof for this issue on the plaintiff. *Id.* That holding was unambiguously abrogated by Congress in the Civil Rights Act of 1991. 42 U.S.C. § 2000e-2(k)(1)(A)(i).

92. *See, e.g.*, Robert Belton, *The Unfinished Agenda of the Civil Rights Act of 1991*, 45 RUTGERS L. REV. 921, 926–27 (1993).

alternatives before *Wards Cove*? The answer depends partly on *Albemarle*'s effect on *Lorillard*.

But *Albemarle*'s effect on *Lorillard* was uncertain. It did not expressly disapprove of *Lorillard*'s approach of requiring defendants to show there are no alternatives with less racial impact. Rather, *Albemarle* added another route for plaintiffs to prove discrimination without addressing the approach taken in *Lorillard* and other cases—that a defendant's burden of demonstrating business necessity entails showing that no acceptable alternatives are available.[93] In other words, *Albemarle* did not foreclose the possibility that less discriminatory alternatives remain relevant to whether a defendant has shown that its practice is truly necessary.

After *Albemarle*, some legal authorities continued to find the availability of alternatives relevant at the second step in the analysis. In 1978, the federal agencies responsible for enforcing employment discrimination laws issued the Uniform Guidelines on Employee Selection Procedures,[94] laying out detailed principles for validating tests or other employee selection procedures found to have an adverse impact. The Guidelines specifically state that a validity study should include "an investigation of suitable alternative selection procedures and suitable alternative methods of using the selection procedure which have as little adverse impact as possible,"[95] and an employer should make "a reasonable effort to become aware of such alternative procedures . . . ."[96] The Guidelines thus indicate that the employer's burden in justifying a test or selection procedure through a validity study includes making a reasonable search for alternatives with less adverse impact.

The Civil Rights Act of 1991 expressly placed the burden of showing an alternative employment practice on the plaintiff, but courts have sometimes also found consideration of available alternatives relevant to the employer's burden at the second step. For example, in *Bradley v. Pizzaco of Nebraska, Inc.*,[97] the court held that to justify a practice with discriminatory impact, the employer must prove a "compelling need" for its policy *and* "the lack of an effective alternative policy that would not produce a similar disparate impact."[98] Similarly, in *EEOC*

93. The *Albemarle* Court was clearly aware of *Lorillard* and the other cases cited in note 82, *supra*, because it cited them in its opinion; however, it never addressed their reasoning about the relevance of less discriminatory alternative practices. *See* Albemarle Paper Co. v. Moody, 422 U.S. 405, 413 n.6–7, 414 n.8, 445 (1975).

94. *See* EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607 (2023).

95. *Id.* § 1607.3(B). The Guidelines also state that "[w]here two or more selection procedures are available which serve the [employer's] legitimate interest in efficient and trustworthy workmanship, and which are substantially equally valid for a given purpose," the employer should use the one with less adverse impact. *Id.*

96. *Id.* The Guidelines also state that the "alternative selection procedures investigated and available evidence of their impact should be identified" in documentation for validity studies. *Id.* §§ 1607.15(C)(6), (D)(8).

97. 7 F.3d 795 (8th Cir. 1993).

98. *Id.* at 797. The case involved a challenge to a strict no-beard requirement for delivery drivers on the grounds that it disproportionately excluded African-American males who suffer at higher rates from

v. Dial Corp.,[99] the court stated that "[p]art of the employer's burden to establish business necessity is to demonstrate the need for the challenged procedure," which includes showing that other measures "could not produce the same results."[100]

Under the FHA, there has been similar uncertainty about who bears the burden of showing a less discriminatory alternative. Like the early employment cases, some courts found that the availability of alternatives was part of the defendant's burden in disparate impact housing cases. In *Resident Advisory Board v. Rizzo*, the Third Circuit held that after a plaintiff established a prima facie case, a defendant must show that its challenged practice serves "in theory and practice, a legitimate, bona fide interest"[101] and that "no alternative course of action could be adopted that would enable that interest to be served with less discriminatory impact."[102] If a defendant satisfied that burden, the Court held, the burden switched back to the plaintiff, who would then have to demonstrate "that other practices [were] available."[103] Thus, the Third Circuit found the existence of alternative practices relevant both to the defendant's burden of justification and to the plaintiff's burden in the final third step of the analysis.

The Second Circuit followed the reasoning in *Rizzo*, holding that "a defendant must present bona fide and legitimate justifications for its action with no less discriminatory alternatives available."[104] Although other circuits placed the burden of proving a less discriminatory alternative on plaintiffs,[105] the Third Circuit reiterated its view in 2011 that an inquiry into alternatives was relevant at both the second and third steps of the analysis. It held that defendants in FHA cases bear the "initial burden of showing that there are no less discriminatory alternatives"[106] before the burden shifts back to the plaintiffs, "who must demonstrate that there is a less discriminatory way to advance the defendant's legitimate interest."[107]

---

pseudofolliculitis barbae (PFB), a skin condition which makes shaving difficult or impossible. *Id.* at 796. The court found that the defendant had not shown business necessity because it could have easily adopted the less discriminatory practice of permitting a medical exception for employees with PFB. *Id.* at 798–99.

99. 469 F.3d 735 (8th Cir. 2006).

100. *Id.* at 743. The court specifically rejected the defendant's argument that the burden regarding less discriminatory alternatives must be placed on the plaintiff. *Id.* The court found that because the defendant had not shown business necessity, the third step of the analysis was never reached and no burden shifted to the plaintiff. *Id.*

101. 564 F.2d 126, 149 (3d Cir. 1977).

102. *Id.*

103. *Id.* at 149 n.37.

104. Huntington Branch, NAACP v. Town of Huntington, 844 F.2d 926, 939 (2d Cir. 1988); s*ee also* Salute v. Stratford Greens Garden Apartments, 136 F.3d 293, 302 (2d Cir. 1998) (similarly pointing to *Rizzo* and *Huntington Branch* for the no less discriminatory alternative standard).

105. *See, e.g.*, Oti Kaga, Inc. v. S.D. Hous. Dev. Auth., 342 F.3d 871, 883–84 (8th Cir. 2003); Graoch Assocs. # 33, L. P. v. Louisville/Jefferson Cnty. Metro Hum. Rels. Comm'n, 508 F.3d 366, 374 (6th Cir. 2007); Mountain Side Mobile Ests. P'ship v. Sec'y of Hous. & Urb. Dev., 56 F.3d 1243, 1254 (10th Cir. 1995).

106. Mt. Holly Gardens Citizens in Action, Inc. v. Twp. of Mount Holly, 658 F.3d 375, 387 (3d Cir. 2011).

107. *Id.* at 382.

Agency guidance regarding who bears the burden of establishing a less discriminatory alternative in the housing context was inconsistent for many years. A policy statement issued by HUD and other federal agencies in 1994 noted the relevance of alternative policies with less discriminatory effect without indicating which party bore the burden on the issue.[108] Later that year, HUD indicated it would propose a rule that would "describe the standards required to demonstrate business necessity and the absence of alternatives with a less discriminatory impact,"[109] indicating that HUD believed a defendant must show that no less discriminatory alternatives were available. The following year, in a proposed rulemaking regarding certain mortgage lenders, HUD suggested that if these entities relied on factors that have a disparate impact, they were required to show both business necessity and that no less discriminatory alternatives exist.[110] Ultimately, HUD withdrew this language in the final rule.[111] The following year, in 1996, HUD also withdrew its "Methods of Proof of Discrimination Under the Fair Housing Act" rule from its list of regulatory priorities.[112]

In 2013, HUD at last promulgated a final rule regarding disparate impact liability under the FHA and squarely placed the burden to demonstrate a less discriminatory alternative on plaintiffs. After a defendant shows that its practice is necessary to meet a substantial, legitimate, nondiscriminatory interest, the plaintiff may still prevail by demonstrating that "the challenged practice could be served by another practice that has a less discriminatory effect."[113] In issuing the regulation, HUD stated that placing the burden on the plaintiff "makes the most

108. *See* Policy Statement on Discrimination in Lending, 59 Fed. Reg. 18266, 18269 (Apr. 15, 1994) ("Even if a policy or practice that has a disparate impact on a prohibited basis can be justified by business necessity, it still may be found to be discriminatory if an alternative policy or practice could serve the same purpose with less discriminatory effect.").

109. Statement of Regulatory Priorities, 59 Fed. Reg. 57087, 57102 (Nov. 14, 1994).

110. This proposed rulemaking involved government-sponsored enterprises (GSEs), Fannie Mae and Freddie Mac, under the Federal Housing Enterprises Financial Safety and Soundness Act of 1992. HUD proposed language that seemed to require the GSEs to prove, as part of a demonstration of business necessity, that no less discriminatory alternative exists. 60 Fed. Reg. 9154, 9190 (Feb. 16, 1995) ("[W]here such factors have a disparate result on the basis of race, color, religion, sex, handicap, familial status, age, or national origin . . . the factors cannot be considered unless they both are justified by business necessity and no less discriminatory alternative to such factors exists.").

111. In its final rule regarding Fannie Mae and Freddie Mac in December 1995, HUD withdrew its earlier language that would have placed the less discriminatory alternative burden on GSEs, citing objections from entities like Freddie Mac and the Mortgage Bankers Association. *See* 60 Fed. Reg. 61846, 61867 (Dec. 1, 1995).

112. OFF. OF INFO. & REGUL. AFFS., OFF. OF MGMT. & BUDGET, EXEC. OFF. OF THE PRESIDENT, METHODS OF PROOF OF DISCRIMINATION UNDER THE FAIR HOUSING ACT (FR-3534) (1996), https://www.reginfo.gov/public/do/eAgendaViewRule?pubId=199604&RIN=2529-AA67 [https://perma.cc/NB6Z-YWQD].

113. 78 Fed. Reg. 11482 (Feb. 15, 2013). Under the Trump administration, HUD attempted to issue a revised rule that would have not only confirmed that plaintiffs bear the burden of establishing a less discriminatory alternative, but also raised the standard for making such a showing. 84 Fed. Reg. 42854 (Aug. 19, 2019). A federal court stayed the revised rule and, after President Biden took office, HUD withdrew it and issued a final rule restoring its 2013 rule. 88 Fed. Reg. 19450 (Mar. 31, 2023) (to be codified at 24 C.F.R. pt. 100).

sense because it does not require either party to prove a negative" and is consistent with schemes under Title VII and the ECOA.[114]

The treatment of less discriminatory alternatives in the credit context is similarly inconsistent. Because of the paucity of ECOA disparate impact cases, most of the relevant history surrounds Regulation B, ECOA's implementing regulation.[115] Citing *Griggs* and *Albemarle*, early rulemaking efforts indicated that an "effects test" applied to creditworthiness decisions without resolving the role of less discriminatory alternatives.[116] In the 1990s, agencies enforcing ECOA issued inconsistent guidance, at times suggesting that to justify a practice with disparate impact a creditor was required to prove a business purpose *and* that no less discriminatory alternative was available,[117] and at other times appeared to backtrack from that position.[118] Separately, the official staff interpretation of Regulation B points to the doctrine as established in Title VII as well as the burdens of the Civil Rights Act of 1991, suggesting that the plaintiff bears the burden of demonstrating that the defendant's legitimate business need can "reasonably be achieved as well by means that are less disparate in their impact."[119]

Without a doubt, it is possible for a plaintiff to establish disparate impact liability—even if a challenged practice is justified by business necessity—by demonstrating the existence of a viable, less discriminatory alternative to the existing practice. But some legal authorities have also found that the existence of a less discriminatory alternative is part of the defendant's burden of justification to show that practices with less disparate impact are unavailable.

C. WHAT IS A LESS DISCRIMINATORY ALTERNATIVE?

Putting aside the question of who bears the burden, an entirely separate question remains: What exactly is a less discriminatory alternative for legal

114. Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460, 11474 (Feb. 15, 2013).

115. *See* Equal Credit Opportunity Act Amendments of 1976, Pub. L. No. 94–239, § 2, 90 Stat. 251 (codified at 15 U.S.C. § 1691).

116. Amendments to Regulation B to Implement the 1976 Amendments to the Equal Credit Opportunity Act, 42 Fed. Reg. 1242, 1246, 1255 (Jan. 6, 1977) (citing Griggs v. Duke Power Co., 401 U.S. 424 (1971); Albemarle Paper Co. v. Moody, 422 U.S. 418 (1975)).

117. In December 1994, the Federal Reserve Board (FRB) proposed amendments to the staff commentary to Regulation B. A proposed comment regarding disparate impact and empirically derived and other credit scoring systems suggested that "credit scoring systems that employ neutral factors could violate the act or regulation if there is a disparate impact on a prohibited basis, unless the practice is justified by business necessity with no less discriminatory alternative available." 59 Fed. Reg. 67235, 67237 (Dec. 29, 1994). Separately, the National Credit Union Administration sent a letter to unions with an "informational white paper" attached, which explains that for a lender to justify a practice that has a disparate impact, "a lender would be required to prove a business purpose for the policy and that no less discriminatory alternative is available." Nat'l Credit Union Admin., NCUA Letter to Unions, Letter No. 174 (Aug. 1995).

118. In June 1995, the FRB backtracked and deleted the proposed comment. 60 Fed. Reg. 29965, 29966 (June 7, 1995). Notably the Board pointed out that "commenters uniformly expressed concern . . . about the Board's articulation of the standards of proof and burdens of persuasion under a disparate impact analysis (sometimes referred to as the effects test)." *Id*. at 29967.

119. 12 C.F.R. § 1002.6(a)-2.

purposes?[120] As an initial matter, when plaintiffs argue that a defendant should have adopted a less discriminatory alternative, they must show that such an alternative actually exists (i.e., that adopting it would have less discriminatory impact than the challenged practice).[121] Such a requirement makes sense: if the goal is to reduce or remove disparate impact, an alternative is not truly viable if its effects on a disadvantaged group are unchanged or even worsen. Courts have also held that the alternative cannot be merely hypothetical or speculative.[122] Instead, there must be evidence that the proposed alternative will actually reduce the disparate impact.[123] What evidence is sufficient is often contested, and proof may be particularly challenging when the proposed alternative involves a wholly different choice or method. For example, in a case challenging a housing development as discriminatory, it may be very difficult

120. Legal scholars and practitioners have bemoaned for decades the absence of concrete legal guidance regarding less discriminatory alternatives. *See, e.g.*, David C. Hsia, *The Effects Test: New Directions*, 17 SANTA CLARA L. REV. 777, 784 (1977) (noting for Title VII purposes, "it is unclear just how much less discriminatory a suggested alternative employment practice must be in order to result in a finding of unlawful discrimination"); Note, *The Civil Rights Act of 1991 and Less Discriminatory Alternatives in Disparate Impact Litigation*, 106 HARV. L. REV. 1621, 1627 (1993) (noting that courts, Congress, and regulators provided little guidance for "when a proposed [less discriminatory alternative] is sufficiently less discriminatory to warrant imposition on an employer"); Peter E. Mahoney, *The End(s) of Disparate Impact: Doctrinal Reconstruction, Fair Housing and Lending Law, and the Antidiscrimination Principle*, 47 EMORY L.J. 409, 494 (1998) ("[L]ittle guidance exists in the case law regarding the showing required under the less discriminatory alternative prong."); Michael G. Allen, Jamie L. Crook & John P. Relman, *Assessing HUD's Disparate Impact Rule: A Practitioner's Perspective*, 49 HARV. C.R.-C.L. L. REV. 155, 189 (2014) (explaining that "[h]ow much of a decrease in predictive power, if any . . . a defendant [must] accept to achieve a less discriminatory effect" would "likely turn on a case-by-case inquiry" and "in a majority of cases the ultimate weighing of competing, nonquantifiable factors will rest with the fact finder"); Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 580 (2016) (arguing that a less discriminatory alternative and the performance of the alternative practice is unclear); Scherer et al., *supra* note 66, at 471 ("A key question that remains largely unresolved is how effective the plaintiff's proposed alternative must be to defeat an employer's showing of business necessity."); Jason Jia-Xi Wu, *Algorithmic Fairness in Consumer Credit Underwriting: Towards a Harm-Based Framework for AI Fair Lending*, 21 BERK. BUS. L.J. 65, 113 (2024) (arguing that for purposes of the ECOA, "the existing legal standard for assessing the sufficiency of 'less discriminatory alternatives' is nebulous").

121. *See* Jones v. City of Boston, 845 F.3d 28, 37 (1st Cir. 2016); Lopez v. City of Lawrence, 823 F.3d 102, 121 (1st Cir. 2016).

122. *See* Allen v. City of Chicago, 351 F.3d 306, 314 n.9 (7th Cir. 2003); Gillespie v. Wisconsin, 771 F.2d 1035, 1045 (7th Cir. 1985). HUD also notes that less discriminatory alternatives "may not be hypothetical or speculative." Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460, 11473.

123. *See, e.g.*, Finch v. Hercules Inc., 865 F. Supp. 1104, 1132 (D. Del. 1994) ("While plaintiff at summary judgment is not required to prove his proposed alternative is less discriminatory, he must at least introduce some evidence to support such a conclusion."); Johnson v. City of Memphis, 770 F.3d 464, 475–76 (6th Cir. 2014) ("[P]laintiffs' briefing offers no data showing that simulations provide equally valid and less discriminatory evaluations than other forms of practical tests."); Sw. Fair Hous. Council, Inc. v. Maricopa Domestic Water Improvement Dist., 17 F.4th 950, 971 (9th Cir. 2021) ("Appellants here provide arguments but fail to present evidence sufficient to allow a jury to conclude that any equally effective, less discriminatory alternatives exist." (emphasis omitted)).

to measure the reduction in disparate impact when the proposed alternative involves a different location or type of development.[124]

In addition to reducing disparities, the alternative must also advance the defendant's legitimate business purposes. Once again, however, there has been a great deal of disagreement about the details.[125] Uncertainty surrounds two related questions. First, how similar must the proposed alternative be to the defendant's current practice in meeting its goals? And second, are costs relevant to determining whether an alternative is a viable one? These questions are related because some courts have suggested that additional costs might make a proposal sufficiently less effective than the challenged practice such that it is not a viable alternative.

Much of the confusion around these issues started with Title VII. As explained above, the amended statute says that one route to disparate impact liability is to make a demonstration "in accordance with the law as it existed on June 4, 1989, with respect to the concept of 'alternative employment practice.'"[126] The difficulty is that the law regarding alternative employment practices was not clear on June 4, 1989.[127] Prior to the decision in *Wards Cove* on June 5, 1989, the Supreme Court had decided only a handful of disparate impact cases, none of which offered a definitive interpretation of what an alternative practice entailed.[128]

In *Watson v. Fort Worth Bank & Trust*,[129] decided the year before *Wards Cove*, Justice O'Connor tried to offer such an interpretation. She suggested that when

---

124. *See* N.Y.C. Env't Just. All. v. Giuliani, 214 F.3d 65, 72 (2d Cir. 2000) ("[C]onclusory statements about the assumed availability of other buildable City-owned lots in the vicinity of particular gardens does not suffice to establish a likelihood that the plaintiffs will meet their burden of showing that a less discriminatory option is available to achieve the City's legitimate governmental goals.").

125. Some courts have taken a fairly restrictive view of the kinds of evidence that would establish if a practice is similarly effective in advancing a corporation's goals, or if an alternative actually reduces a disparate effect. For example, some courts have rejected proposed alternatives that would simply have a defendant revert to an old practice, despite evidence it was just as effective. *See* Hardie v. Nat'l Collegiate Athletic Ass'n, 876 F.3d 312, 321 (9th Cir. 2017). Other courts have rejected proposed alternatives that point to policies used by defendants in slightly different contexts. *See* Contreras v. City of Los Angeles, 656 F.2d 1267, 1285 (9th Cir. 1981). Some courts have held that even where evidence supports the contention that an alternative practice in general tends to result in less adverse impact, that's not enough. *See* Lopez v. City of Lawrence, 823 F.3d 102, 120 (1st Cir. 2016). Still other courts have taken a less restrictive view and have approved of alternatives previously used by the defendant and other similarly-situated companies, or at least deemed them relevant. *See* Kilgo v. Bowman Transp., Inc., 570 F. Supp. 1509, 1521 (N.D. Ga. 1983), *aff'd*, 789 F.2d 859 (11th Cir. 1986); Christner v. Complete Auto Transit, Inc., 645 F.2d 1251, 1263 (6th Cir. 1981).

126. 42 U.S.C. § 2000e-2(k)(1)(C).

127. *See* 137 Cong. Rec. S29055 (daily ed. Oct. 30, 1991) (statement of Sen. Wallop) ("The terms 'business necessity' and 'job related' are intended to reflect the concepts enunciated by the Supreme Court in *Griggs v. Duke Power* and in the other Supreme Court decisions prior to *Wards Cove v. Atonio*. . . . [a] non-standard . . . ." (quoting L. Gordon Crovitz, *Bush's Quota Bill: (Dubious) Politics Trumps Legal Principle*, WALL ST. J., Oct. 30, 1991)).

128. Most of its cases did not reach the issue at all. *Dothard v. Rawlinson* did not discuss alternative employment practices at all. *See generally* 433 U.S. 321 (1977). In *New York City Transit Authority v. Beazer*, the Court found that the employer's practice was justified, and therefore discussion of possible alternatives was unnecessary. *See* 440 U.S. 568, 590–91 (1979). *Connecticut v. Teal* also failed to discuss any alternative employment practices. *See generally* 457 U.S. 440 (1982).

129. 487 U.S. 977 (1988).

determining whether a less discriminatory alternative exists, "[f]actors such as the cost or other burdens of proposed alternative selection devices are relevant in determining whether they would be equally as effective as the challenged practice in serving the employer's legitimate business goals."[130] This part of her opinion was joined by only three other Justices, and therefore did not represent the views of the Court or create binding precedent. Nevertheless, Justice O'Connor appeared to be pushing the view that any alternative must be "equally as effective" as the challenged practice, and that cost was relevant to that determination.[131]

In *Wards Cove*, a majority of the Justices finally agreed on Justice O'Connor's formulation,[132] but Congress soon nullified that holding in the Civil Rights Act of 1991.[133] While Congress failed to provide detailed guidance regarding alternative employment practices, it clearly intended to abrogate the reasoning of *Wards Cove* on that issue and thus arguably rejected the "equally effective" requirement for a less discriminatory alternative.[134] At a minimum, the legislative response indicates that a viable alternative practice need not perform identically to the employer's challenged practice and that some costs might have to be incurred in adopting the alternative.

Lower courts that have considered less discriminatory alternatives in the employment context have articulated the requirements in various ways. Some courts have held that a proposed alternative must be "equally effective" or "equally valid,"[135] while others have relied on a looser showing, such as "comparably effective."[136] EEOC Guidance also indicates that a viable alternative need not be identical in performance.[137] Given that Congress repudiated the reasoning

130. *Id.* at 998.

131. *See id.* Other courts, addressing the *Wards Cove* standard before the Civil Rights Act of 1991, held that a plaintiff "in a disparate impact suit bears some burden of proving that plaintiff's suggested alternative practice is no more expensive than the employer's current practice, or—at the very least—that the practice is *economically feasible for the employer*." MacPherson v. Univ. of Montevallo, 922 F.2d 766, 773 (11th Cir. 1991) (emphasis added).

132. 490 U.S. 642, 661 (1989).

133. *See supra* notes 90–92 and accompanying text.

134. *See* 42 U.S.C. § 2000e-2(k)(1)(C).

135. Chicago Tchrs. Union, Local 1 v. Bd. of Educ. of Chicago, 419 F. Supp. 3d 1038, 1053 (N.D. Ill. 2020); Hardie v. Nat'l Collegiate Athletic Ass'n, 876 F.3d 312, 321 (9th Cir. 2017).

136. *See, e.g.*, Adams v. City of Chicago, 469 F.3d 609, 613 (7th Cir. 2006) (stating the plaintiffs must show that proposed alternative "would be of substantially equal validity"); Fitzpatrick v. City of Atlanta, 2 F.3d 1112, 1118 (11th Cir. 1993) (explaining that plaintiff can prevail by showing an alternative policy "with lesser discriminatory effects that would be comparably as effective" at meeting the employer's business needs). *Cf.* Cureton v. Nat'l Collegiate Athletic Ass'n, 37 F. Supp. 2d 687, 713 (E.D. Pa. 1999), *rev'd*, 198 F.3d 107 (3d Cir. 1999) (noting that "'equally effective' means equivalent, comparable, or commensurate rather than identical" under Title VI); Elston v. Talladega Cty. Bd. of Educ., 997 F.2d 1394, 1407 (11th Cir. 1993) (articulating a "comparably effective" standard in Title VI cases).

137. *See* EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.3(B) (2023) ("substantially equally valid"); *Select Issues: Assessing Adverse Impact in Software, Algorithms, and Artificial Intelligence Used in Employment Selection Procedures Under Title VII of the Civil Rights Act of 1964*, U.S. EQUAL EMPL. OPPORTUNITY COMM'N (May 18, 2023), https://www.eeoc.gov/select-issues-assessing-adverse-impact-software-algorithms-and-artificial-intelligence-used [https://perma.cc/WWX9-TLXY] (using the phrase "comparably effective") [hereinafter *EEOC Select Issues*].

of *Wards Cove*, the better reading of Title VII is that a less discriminatory alternative should be comparable but not necessarily "equally effective."

Courts have also taken a range of approaches regarding costs. The cost of a proposed alternative remains relevant, and when the cost of adopting a new procedure is high, it has been grounds for finding that the alternative is not a viable one, particularly when it is unclear that it will produce a reduction in disparities.[138] On the other hand, courts have found that a less discriminatory practice may entail some administrative costs. As the Court in *Lorillard* explained, "some additional administrative costs may be imposed . . . . [A]voidance of the expense of changing employment practices is not a business purpose that will validate . . . an otherwise unlawful employment practice."[139]

In the housing context, HUD directly addressed what is required for a less discriminatory alternative in its regulations on disparate impact. The agency explicitly rejected an approach that would require plaintiffs' less discriminatory alternatives to be "equally effective."[140] It argued that such a heightened standard is "less appropriate in the housing context than in the employment area in light of the wider range and variety of practices covered by the Act that are not readily quantifiable."[141] Other regulatory guidance is consistent with this approach. In an advisory bulletin on fair lending, the Federal Housing Finance Authority (FHFA) suggested that a less discriminatory alternative need not be equally effective, but only comparably so,[142] and that alternatives

138. *See, e.g.*, Lopez v. City of Lawrence, 823 F.3d 102, 121 (1st Cir. 2016).

139. Robinson v. Lorillard Corp., 444 F.2d 791, 800 (4th Cir. 1971); *see also* Newark Branch, NAACP v. Town of Harrison, 940 F.2d 792, 804–05 (3d Cir. 1991) (rejecting defendant's argument that it should not have to eliminate residence requirement that caused a disparate impact because it would incur additional expense in processing an increased number of applications); NAACP v. N. Hudson, 665 F.3d 464, 481–82 (3d Cir. 2011) (following reasoning in *Town of Harrison* and ruling in favor of plaintiffs' disparate impact claim, in part because less discriminatory alternatives are available). Similarly, in a 1971 decision the EEOC explicitly noted that while it is not true that "there is no limit to the expense that Title VII requires an employer to incur in seeking reasonable alternatives to present employment criteria which operate to exclude minority groups. . . . the level of reasonable expense would increase in direct proportion to the extent of the impact." EEOC Dec. No. 72-0708, 4 Fair Empl. Prac. Cas. 437, 438 (1971) (emphasis omitted).

140. 78 Fed. Reg. 11460, 11473.

141. *Id.* Under the Trump administration, HUD sought to raise the bar by requiring a viable alternative to be "equally effective" "without imposing materially greater costs" on the defendant. 85 Fed. Reg. 60288, 60321 (Sept. 24, 2020). That change was reversed under President Biden, when HUD restored the original version of its regulations in 2023. In doing so, the agency explained that *Inclusive Communities* did not require that alternative policies be "equally effective." 88 Fed. Reg. 19450, 19491 (Mar. 31, 2023).

142. Federal Housing Finance Agency, Advisory Bulletin AB 2021-04 (Dec. 20, 2021) (available at https://www.fhfa.gov/sites/default/files/2023-07/AB%202021-04%20Enterprise%20Fair%20Lending%20and%20Fair%20Housing%20Compliance.pdf). The Bulletin provided the example of an automated underwriting model that includes a factor that leads to significantly lower disproportionate acceptance rates for Black borrowers. *Id.* at 8. If that factor only marginally improves the model's ability to predict risk, continued reliance on that factor "would be a violation because it has a significant disparate impact but the model without the factor would be a less discriminatory alternative." *Id.* In other words, a comparably, but not identically, effective model is a viable less discriminatory alternative.

may sometimes require entities to bear some minimal costs.[143]

In the credit context, interagency procedures have held variously that less discriminatory alternatives need to be "approximately equally effective," "equally effective,"[144] or "serve the same purpose with less discriminatory effect," but provide no discussion of how costs impact an alternative's viability.[145] Regulatory enforcement by the Consumer Financial Protection Bureau (CFPB) and DOJ against two indirect auto lenders sheds some additional light. In investigations into Honda Finance Corporation and Toyota Motor Credit Corporation, the CFPB found that each company's dealer markup policy was "not justified by legitimate business need and constitute[d] discrimination."[146] Those policies allowed automotive dealers to charge an interest rate above the interest rate at which the financing entity would provide financing.[147] Under those policies, dealers would receive extra compensation from the increased interest revenue derived from the dealer markup. In each enforcement action, the CFPB identified three potential less discriminatory alternative policies, even though each would force lenders to forego increased interest revenue.[148] Though not formally crafted as less discriminatory alternatives, the alternative policies suggest that defendants may need to forego substantial revenue in adopting an alternative.

Because such judgments are inherently contextual and fact-specific, the law cannot precisely define what exactly a less discriminatory alternative is. Nevertheless, it is clear that any such alternative must actually reduce the harmful impact on a disadvantaged group and meet the defendant's legitimate needs. And, though excessive costs may render an alternative nonviable, many authorities

143. The Bulletin also offers the example of an underwriting model that has a higher cutoff score for certain metro areas, which causes a disparate impact on Black and Latino applicants. *Id.* at 9. If the projected losses of not using the higher cutoff score for those metro areas are minimal, and the entity generally doesn't take metro-area differences into account in underwriting, the policy "would be a violation because a less-discriminatory alternative exists"—i.e., not taking into account metro-area differences. *Id.* In other words, if the losses from adopting a less discriminatory alternative are minimal, the entity should be required to do so.

144. OFF. OF COMPTROLLER OF THE CURRENCY, FED. DEPOSIT INS. CORP., FED. RSRV. BD., OFF. OF THRIFT SUPERVISION, NAT'L CREDIT UNION ADMIN., INTERAGENCY FAIR LENDING EXAMINATION PROCEDURES – APPENDIX 26–27 (2009), https://www.ffiec.gov/pdf/fairappx.pdf [https://perma.cc/9U5K-QBV2].

145. OFF. OF COMPTROLLER OF THE CURRENCY, FED. DEPOSIT INS. CORP., FED. RSRV. BD., OFF. OF THRIFT SUPERVISION, NAT'L CREDIT UNION ADMIN., INTERAGENCY FAIR LENDING EXAMINATION PROCEDURES, at iv (2009), https://www.ffiec.gov/PDF/fairlend.pdf [https://perma.cc/UH25-2LBC].

146. Consent Order, In re Am. Honda Fin. Corp., CFPB No. 2015-CFPB-0014, at 8 (Jul. 14, 2015), https://files.consumerfinance.gov/f/201507_cfpb_consent-order_honda.pdf; Consent Order, In re Toyota Motor Credit Corp., CFPB No. 2016-CFPB-0002, at 9 (Feb. 2, 2016), https://files.consumerfinance.gov/f/201602_cfpb_consent-order-toyota-motor-credit-corporation.pdf.

147. CFPB No. 2015-CFPB-0014, at 1; CFPB No. 2016-CFPB-0002, at 1.

148. CFPB No. 2015-CFPB-0014, at 9; CFPB No. 2016-CFPB-0002, at 9. *See CFPB and DOJ Reach Resolution with Toyota Motor Credit To Address Loan Pricing Policies With Discriminatory Effects,* CONSUMER FIN. PROT. BUREAU (Feb. 2, 2016), https://www.consumerfinance.gov/about-us/newsroom/cfpb-and-doj-reach-resolution-with-toyota-motor-credit-to-address-loan-pricing-policies-with-discriminatory-effects/ [https://perma.cc/FQP2-LJMG].

suggest that a defendant may not avoid implementing an alternative practice simply because it might entail some additional costs.

III. WHAT MODEL MULTIPLICITY MEANS FOR THE LAW

The overarching purpose of our civil rights laws is to remove arbitrary barriers to full participation in the nation's economic life for marginalized groups. Given the insights of model multiplicity, it makes little sense to permit the use of algorithms that arbitrarily exclude certain groups when less discriminatory models are available. Thus, our core argument is that entities that use decisionmaking algorithms should be legally required to search for less discriminatory alternatives before deploying them in critical domains like employment, housing, and credit.[149] First, under disparate impact doctrine, a defendant's burden of justifying a model with discriminatory effects should include a showing that it made a reasonable search for LDAs before implementing it. Although our argument regarding the application of disparate impact doctrine to algorithms is novel, our proposal does not entail a radical change in the law. Rather, it is consistent with and supported by numerous existing legal authorities. Second, as policymakers develop regulatory frameworks for the governance of algorithms, they should include a requirement that entities undertake a reasonable search for less discriminatory models. To be clear, these legal requirements would not impose a duty to find the least discriminatory model—a difficult, if not impossible, task under many circumstances. Rather, entities would have to show that their efforts to search for less discriminatory algorithms were reasonable.

A. DUTY TO SEARCH

Given the broad adoption of algorithmic systems in civil rights domains, legal duties must address the risks of discriminatory effects. What model multiplicity teaches is that whenever an algorithm has a disparate effect on a disadvantaged group, alternative models with less detrimental effects that are comparably effective in achieving the business purpose (in that they perform equally well) likely

149. Civil rights advocates have stressed the importance of interventions that would require companies to search for less discriminatory alternatives. In 2022 a range of civil rights groups released the Civil Rights Standards for 21st Century Employment Selection Procedures, which detail an auditing approach that would require exploring alternative selection procedures that might reduce or eliminate potential sources of discrimination. *See* CTR. FOR DEMOCRACY & TECH. ET AL., CIVIL RIGHTS STANDARDS FOR 21ST CENTURY EMPLOYMENT SELECTION PROCEDURES 8 (2022), https://cdt.org/insights/civil-rights-standards-for-21st-century-employment-selection-procedures/ [https://perma.cc/K94H-TXPD]. Groups have also called for financial regulators to "consider how lenders can be encouraged to develop less discriminatory alternatives to biased models" and urged the CFPB to suggest that it "expects entities to meaningfully, effectively, and routinely search for and adopt LDA models." *See* Letter from Nat'l Cmty. Reinvestment Coal., to Off. of the Comptroller of the Currency, Bd. of Govs. of the Fed. Rsrv. Sys., Fed. Deposit Ins. Corp., CFPB, Nat'l Credit Union Adm. 11 (Jul. 1, 2021), https://ncrc.org/request-for-information-and-comment-on-financial-institutions-use-of-artificial-intelligence-including-machine-learning/ [https://perma.cc/6CLV-BQLF]; Letter from Nat'l Cmty. Reinvestment Coal. to Rohit Chopra, Director, CFPB 2–3 (Mar. 11, 2022), https://ncrc.org/cfpb-should-encourage-lenders-to-look-for-less-discriminatory-models/ [https://perma.cc/428S-TMTF].

exist. Where such alternatives exist, organizations can adopt them with little or no negative effect on their performance goals.

However, as explained in Part I, the model building process will not inevitably or even likely happen upon a less discriminatory model unless developers pay attention to the issue.[150] Reducing unnecessary discriminatory effects will thus require entities to dedicate some effort and resources to looking for LDAs. To achieve the purposes of the civil rights statutes, the law must recognize a duty to take reasonable steps to identify and select models that reduce disparities.

Our proposal resonates with Richard Thompson Ford's argument that anti-discrimination law should be reoriented around requiring powerful actors "to meet a duty of care to avoid unnecessarily perpetuating social segregation or hierarchy."[151] His analysis, which focuses on actions that "perpetuate illegitimate hierarchies *in the run of cases*" rather than individual decisions,[152] is particularly apt for algorithmic systems. These models, by definition, implement decisionmaking systems intended to be applied across cases instead of individualized judgments. Other legal scholars have also called for imposing duties to avoid discrimination.[153] While our proposal shares similarities with these arguments, it crucially places *on the defendant* the responsibility of proving in court that it took reasonable steps to avoid unnecessary disparate impacts.

Legislators, courts, and regulators have expressed reluctance to place a burden on defendants to prove that they have adopted the *least* discriminatory way of achieving their goals, on the belief that it is impossible to know—and therefore to prove—that there are no less discriminatory alternatives.[154] But multiplicity introduces a remarkable degree of certainty into this calculus: Given that developers are extremely unlikely to happen upon the least discriminatory model, it is safe to assume that models that are less discriminatory than any baseline model exist and that these alternatives could be discovered with additional effort. Thus, in moving from traditional decisionmaking to algorithmic decisionmaking based on predictive models, the question shifts from whether any less discriminatory alternatives exist to what resources should be invested to find them. Placing the burden on the plaintiff becomes a test of its resources and wherewithal to find them. In contrast, entities that develop and use algorithms already search across potential models and are therefore much better positioned to look for LDAs. Under these circumstances, it makes sense as both a normative and practical matter to place a duty on the defendant to search for LDAs.

Placing a duty of reasonable search on the entities that develop and deploy predictive models makes sense because they are in the best position to search

150. *See supra* Part I.

151. Richard Thompson Ford, *Bias in the Air: Rethinking Employment Discrimination Law*, 66 STAN. L. REV. 1381, 1384 (2014).

152. *Id.* at 1385.

153. *See, e.g.*, David Benjamin Oppenheimer, *Negligent Discrimination*, 141 U. PA. L. REV. 899, 944–45 (1993).

154. *See supra* Section II.A.

efficiently and to find less discriminatory versions. Indeed, the model development process *inherently* entails exploration of alternatives, including assessing potential models for accuracy, robustness, and other characteristics. That exploration could easily (and should) be expanded to include a comparison of the disparate impacts of different models. This type of exploration is far less costly during the development process than after a model is in use.

In the absence of a duty to search, it would fall to individuals harmed by discriminatory models to challenge them after they have been implemented. Already, potential plaintiffs face significant obstacles to identifying practices that have a disparate impact and bringing successful legal challenges.[155] These obstacles are even more daunting when it comes to challenging discriminatory algorithms. Because comprehensive data that are needed to detect a pattern of disparate impact are unlikely to be available to affected individuals, they may not even realize that they have been subject to a discriminatory algorithm. Even if they did have access to this information, individual complainants likely lack the significant resources and technical skills to analyze them and discover discriminatory patterns.

Plaintiffs would face still greater obstacles to identifying less discriminatory alternatives. They would need access to the model itself, the training data, as well as information about its intended goals, documentation of the various choices made in the development process, and measures of its performance to assess the choices made by the developer and conduct their own search for alternative models. Entities that develop and use models are likely to resist disclosure of such information,[156] and even if plaintiffs obtained all the necessary information through discovery, they would still need significant technical expertise, effort, and computing time to identify possible alternatives that could have been uncovered more cheaply in the development process. And, even if an LDA were identified, the defendant might dispute its efficacy or object to the cost of implementing it—a cost the defendant could have avoided if it had sought out LDAs from the start.

This process not only involves significant monetary expenditure by both plaintiffs and defendants, it also takes a substantial amount of time—time in which an LDA could have been operative. Given these realities, the most effective approach is to put a duty on entities to incorporate a reasonable search for LDAs into the model development process. Doing so is more efficient and increases the likelihood of discovering such alternatives,[157] without imposing the unbounded

155. *See, e.g.*, Melissa Hart, *Disparate Impact Discrimination: The Limits of Litigation, the Possibilities for Internal Compliance*, 33 J.C. & U.L. 547, 551 (2007).

156. In some instances, companies might simply not have the relevant information. For example, the reasoning behind many of the choices made throughout the model development pipeline may not be documented.

157. Civil rights advocates might argue that imposing a duty to search does not go far enough. Requiring a reasonable search rather than mandating a particular result means that there is no guarantee that a viable LDA will be discovered and implemented. However, although there are strong reasons to believe LDAs exist, it is difficult to know ex ante which interventions in the model pipeline will yield less discriminatory models. Requiring entities to explore *every* possible branch in the pipeline to find the

costs of finding a globally optimal solution. It is also consistent with our civil rights laws, as it pushes companies to center anti-discrimination efforts when building models and to eliminate arbitrary barriers to equal opportunity. The next two Sections consider how such a duty should be incorporated into the law.

B. MODEL MULTIPLICITY AND DISPARATE IMPACT DOCTRINE

As explained in Part II, the idea of less discriminatory alternatives is already part of discrimination law.[158] Under the disparate impact framework, it is well recognized that plaintiffs may establish liability, even after a defendant has shown business necessity, by demonstrating that an alternative practice exists that would have a less disparate effect on a disadvantaged group.[159] What is underappreciated is that the existence of a less discriminatory alternative can also be relevant to the defendant's burden of justifying a practice with disparate impact.[160]

Under current law, after a plaintiff has established a prima facie case, the defendant bears the burden of justifying a practice with a disparate impact. It must show that the practice is "job related . . . and consistent with business necessity" under Title VII;[161] is "necessary to achieve one or more substantial, legitimate, nondiscriminatory interests" under the FHA;[162] or "meets a legitimate business need" under ECOA.[163] Although formulated differently, the burdens placed on the defendant under each law refer to need or necessity. "Necessity" implies that the entity cannot accomplish its goals another way.

It makes little sense to say that the defendant's chosen model is "necessary" if a reasonable search would have uncovered an equally effective model with less disparate effect. Thus, part of the defendant's burden should be to demonstrate that it made such a search and was unable to find an LDA.

Doing so does not involve a significant change in the law.[164] As discussed in Section II.A, multiple legal authorities have found that less discriminatory alternatives are relevant to assessing whether a defendant has met its burden of justification after the plaintiff has established a prima facie case of disparate impact.[165] To ensure that the practice is not arbitrarily excluding members of disadvantaged groups, the requirement of business necessity must have some teeth. Determining whether a practice is "necessary" inherently entails consideration of whether the

---

least discriminatory alternative is simply not feasible, as it would consume immense financial, computational, and environmental resources.

158. *See supra* Part II.

159. *See supra* Section II.A.

160. *See supra* Section II.B.

161. 42 U.S.C. § 2000e–2(k)(1)(A)(i); s*ee also* Regulations to Implement the Equal Employment Provisions of the Americans with Disabilities Act, 29 C.F.R. § 1630.10(a) (2023).

162. 24 C.F.R. § 100.500(c)(2) (2023).

163. 12 C.F.R. § 1002.6(a)-2 (2024).

164. We do not believe that requiring entities to make a reasonable search for LDAs will create any additional incentives for plaintiffs to sue. The decision to bring suit will continue to turn on whether potential plaintiffs are able to obtain evidence that an algorithm has a disparate impact.

165. *See supra* Section II.B.

business purpose could be met as well by some other means. Imagine a situation in which there are two comparable, well-known ways of achieving a business objective. If an entity chooses to implement a practice that has a disparate impact when the alternative would not impose similar disadvantages on a marginalized group, then it is difficult to characterize its reliance on the first practice as "necessary." In this context, the availability of a less discriminatory alternative is relevant to judging the defendant's business justification.

A defendant might object that requiring it to show that no other practice with less discriminatory effect exists puts it in the impossible position of proving a negative.[166] That argument has little force, however, in the case of discriminatory algorithms, where an entity might easily uncover less discriminatory alternatives by modestly expanding its search process. The defendant would not be expected to conduct an exhaustive search of all possible models.[167] Instead, as we explore in detail in Part V, there are multiple points of intervention in the machine learning pipeline that could lead to the discovery of LDAs, and some are relatively low cost to employ. Rather than requiring the developer to identify a globally optimal model or to prove a negative,[168] a duty to search entails *reasonable* measures to reduce disparate impacts by broadening the search to include nearby branches.

A 2019 article also recognized the phenomenon of model multiplicity but applied the insight at the final step of the disparate impact analysis.[169] Because there is no guaranteed way of finding the *least* discriminatory model among all models of equivalent accuracy, the authors thought it unfair to hold the employer retrospectively liable if plaintiffs later found a less discriminatory model that the company did not know of.[170] Thus, they asserted that plaintiffs should only be able to prevail at the third step by showing that the employer "actually considered and rejected" an alternative, less discriminatory algorithm.[171]

We draw a different lesson from model multiplicity, arguing that it is relevant to the *defendant*'s burden of showing business necessity at the second step of the analysis. Because the typical process of moving through the machine learning

166. *See supra* Section II.B.

167. *Cf.* 44 Fed. Reg. 11996, 120001 (Mar. 2, 1979) (explaining that the extent of a user's investigation of alternative procedures with less disparate impact should be "reasonable").

168. Scherer et al., *supra* note 66, at 510 (identifying a globally optimal model as "computationally complex for even a modestly large dataset, and wholly impractical for . . . high-dimensionality datasets").

169. *See id.*

170. *Id.* Scherer et al. agree that if a plaintiff discovers an alternative of equal performance and less adverse impact, it is "reasonable for a court to order the tool to be modified, going forward," but believe that no retrospective relief should be granted where the employer was unaware of the alternative. *Id.* As explained, our proposal differs because the defendant's lack of knowledge would not excuse it from liability unless it can show that it undertook a reasonable search to discover it. In either case, a grant of prospective relief should be sufficient to find the plaintiff a "prevailing party" for purposes of awarding attorneys' fees. *See* 42 U.S.C. § 2000e-5(k). The availability of attorneys' fees is important to support plaintiffs' successful efforts to search for less discriminatory alternatives that the defendant may have overlooked.

171. Scherer et al., *supra* note 66, at 510–11.

pipeline only considers a small range of branches in the tree of alternatives, it will likely fail to uncover LDAs that are easily discoverable with modest additional effort. Thus, we argue that companies should have a duty to reasonably search for these alternatives rather than only being liable if they failed to adopt LDAs that they actually considered in the development process.[172] This approach increases the likelihood of discovering LDAs because it creates incentives for developers, who are in the best position to discover them, to search for them in the first place.

Our approach entails no change in the third step of the disparate impact analysis. Recognizing a duty of reasonable search does not conflict with allowing the plaintiff to show that less discriminatory alternatives exist at step three of the disparate impact analysis. Even if the defendant establishes that it conducted a reasonable search, a plaintiff might still uncover a less discriminatory alternative. Particularly where the proposed alternative involves a wholly novel way of meeting a business purpose, it makes sense to put the burden on the plaintiff to describe the practice and demonstrate that it offers a workable alternative. And if the plaintiff does so, a defendant's refusal to adopt it might well call into question its motives, suggesting discriminatory intent.

At this juncture, several clarifying points are warranted. First, a duty to conduct a reasonable search for LDAs does not displace other requirements for justifying a practice with a disparate impact. The defendant still needs to show that its purpose in implementing an algorithmic tool advanced a legitimate business purpose. For example, an employer facing a disparate impact case must still show that its selection algorithm has been validated—i.e., that it measures relevant job skills or characteristics.[173] Demonstrating job-relatedness, however, should not be sufficient to justify an algorithm when the defendant failed to make a reasonable effort to identify and implement comparable models with less disparate effects.

Second, by focusing on the duty to conduct a reasonable search for LDAs, we do not foreclose broader duties for entities to search for other less discriminatory alternatives. For example, in addition to looking for LDAs, entities that rely on algorithms with disparate impacts might also have a duty to consider less discriminatory non-model alternatives. Similarly, when non-algorithmic practices have a disparate impact, perhaps entities should have to justify their failure to consider other readily available, less discriminatory practices. We are not making these broader claims here, but neither do we mean to foreclose them. Instead, this Article focuses more narrowly on the implications of model multiplicity for the law. And because the actions required for a reasonable search for LDAs are relatively unobtrusive and entail no significant change in defendants' business practices, a duty of reasonable search is easily justified in this context.

172. In recent guidance regarding algorithmic decisionmaking systems and compliance with Title VII, the EEOC suggested that "[f]ailure to adopt a less discriminatory algorithm that was considered during the development process . . . may give rise to liability." *EEOC Select Issues*, *supra* note 137.

173. *See* 42 U.S.C. § 2000e–2(k)(1)(A)(i).

Third, although we define LDAs narrowly to include algorithms "equivalent" in performance, as determined by some interval ε, we do not mean to suggest that "equally effective" should be the standard against which alternatives are assessed under disparate impact law. Although the issue is not definitively resolved, legal authorities have found that an alternative need not be "equally effective" to be considered a viable less discriminatory alternative.[174] Indeed, as previously discussed, Congress arguably rejected such a standard when it abrogated *Wards Cove*,[175] and HUD has explicitly disavowed requiring alternatives to be "equally effective."[176] We adopted a particularly high standard in this Article to show that, when it comes to models, LDAs are likely available under *even the most stringent definition* of a legally relevant alternative, and so defendants should have a duty to search for them. However, this does not mean that a stringent requirement of "equal accuracy" is the appropriate legal standard. There may be less discriminatory algorithms that differ somewhat in performance (i.e., fall outside the bounds of ε), but nevertheless reduce disparate impact to such a degree that an entity should be required to adopt them. And outside the algorithmic context, practices that affect access to housing, employment, and credit are so varied, and their performance may be so difficult to measure, that it makes little sense to insist on a strict requirement of equal effectiveness.

What would disparate impact litigation look like under the framework we propose? Imagine a challenge to an algorithm used in credit, where the plaintiff alleges that she was rejected for a loan because the lender relied on a racially discriminatory model. The plaintiff would first have to establish a disparate impact. She might, for example, demonstrate that the model used by the lender results in disproportionately fewer Black applicants receiving loans. At that point, the burden would shift to the defendant to demonstrate that its practice is necessary. It would not only have to demonstrate that its model actually advances a legitimate business need and justify its definition of model performance, but it would also need to show that it undertook a reasonable search for LDAs before adopting the model at issue. The lender might do so by producing evidence of the choices it made during the model building pipeline, such as testing the effect of changes to the combination of input features on group disparities. The plaintiff, of course, might contest whether the lender's efforts were actually reasonable—for example, by arguing that the lender failed to pursue readily available, low-cost explorations, such as comparing the disparate impact of alternative models during the development process.

A company may have found an LDA during its development process but decided against deploying it. In such a case, the rejected model might be evidence that the defendant knew of a model that would have advanced its business purpose with less disparate effects but that it refused to implement it. The defendant

174. *See supra* note 131 and accompanying text.
175. *See supra* Part II.
176. 78 Fed. Reg. 11460, 11473 (Feb. 15, 2013).

would bear the burden of showing that business necessity required it to implement the algorithm with greater disparate effects. Ultimately, a court would have to determine if, based on all the available evidence, the defendant's efforts to search for LDAs were reasonable and whether the availability of a viable LDA trumped a claim of business necessity. If the defendant satisfied its burden of showing a business necessity and that its efforts to search for an LDA were reasonable, then the plaintiff would still have the opportunity to identify a viable LDA that may have been overlooked by the defendant.

What would it take practically to incorporate a duty to search for LDAs? In the employment context, there are clear legal antecedents for requiring defendants to show that available, less discriminatory alternatives are infeasible as part of their burden of justification.[177] That approach is consistent with the text of Title VII, which puts the burden on defendants to demonstrate that a challenged practice is "consistent with business necessity,"[178] as well as the Uniform Guidelines, which specify that employers should investigate suitable alternatives with lesser adverse impact as part of validating their selection procedures.[179] Given these authorities, courts could simply adopt this approach in disparate impact cases involving algorithms under Title VII. Clarification by the EEOC on this point would be helpful. In recent guidance, the EEOC noted that "[one] advantage of algorithmic decisionmaking tools is that the process of developing the tool may itself produce a variety of comparably effective alternative algorithms. Failure to adopt a less discriminatory alternative that was considered during the development process therefore may give rise to liability."[180] Because LDAs are not likely to surface on their own, the EEOC could make clear that satisfying the business necessity test includes a showing that the employer undertook a reasonable search for LDAs during the development process.

In the housing and credit contexts, requiring entities that rely on predictive algorithms to search for LDAs would provide beneficial regulatory clarification.[181] CFPB officials have already noted in public remarks that "[r]igorous searches for less discriminatory alternatives are a critical component of fair lending compliance management."[182] While such remarks align with our argument, they are no

177. *See supra* Section II.B.

178. 42 U.S.C. § 2000e–2(k)(1)(A)(i).

179. *See supra* notes 94–96. The Guidelines contemplate that entities that use selection procedures make a reasonable effort to become aware of and investigate alternatives.

180. *EEOC Select Issues*, *supra* note 137.

181. *See, e.g.*, MICHAEL AKINWUM, JOHN MERRILL, LISA RICE, KAREEM SALEH & MAUREEN YAP, AN AI FAIR LENDING POLICY AGENDA FOR THE FEDERAL FINANCIAL REGULATORS, BROOKINGS 7–11 (2021), https://www.brookings.edu/articles/an-ai-fair-lending-policy-agenda-for-the-federal-financial-regulators/ [https://perma.cc/YZ3S-FFCV].

182. Brad Blower, *CFPB Puts Lenders & Fintechs on Notice: Their Models Must Search for Less Discriminatory Alternatives or Face Fair Lending Non-Compliance Risk*, NAT'L CMTY. REINVESTMENT COAL. (Apr. 5, 2023), https://ncrc.org/cfpb-puts-lenders-fintechs-on-notice-their-models-must-search-for-less-discriminatory-alternatives-or-face-fair-lending-non-compliance-risk/ [https://perma.cc/RMG5-KDQR]. Former CFPB Director Patrice Alexander Ficklin made similar remarks during a webinar several months later. *See Getting Ahead of the Curve: Emerging Issues in the Use of AI and Machine Learning in*

substitute for formal, detailed guidance.[183] To formalize this expectation, the CFPB could consider updating its examination manual to specifically ask questions regarding the search for LDAs, or issue an advisory opinion.[184] Depending on ongoing supervisory activities related to the expectation that lenders affirmatively search for LDAs, the CFPB could use its Supervisory Highlights to discuss expectations regarding the search for LDAs. The CFPB could also consider amending Regulation B in several locations to clarify what burdens plaintiffs and defendants bear.[185] Currently, Regulation B does not explicitly do so. Such amendments might make plain that when a plaintiff challenges a creditor's use of an algorithmic system and alleges unlawful disparate impact, the creditor must show as part of its demonstration of business necessity that it took reasonable steps to search for and implement LDAs. Relatedly, separate guidance on model risk management would need to be updated to incorporate searches for less discriminatory alternatives as part of the model development, implementation, and validation process.[186]

In the housing context, regulatory clarification is likely necessary, given the specificity of HUD's existing disparate effects regulation.[187] First, HUD would

Financial Services, FINREGLAB (Jan. 17, 2024, 2:30 PM), https://finreglab.org/events/getting-ahead-of-the-curve-emerging-issues-in-the-use-of-artificial-intelligence-and-machine-learning-in-credit-underwriting/ [https://perma.cc/6WAN-ECKX].

183. *See, e.g.*, FINREGLAB, EXPLAINABILITY & FAIRNESS IN MACHINE LEARNING FOR CREDIT UNDERWRITING 63 (2023) [https://perma.cc/EC7V-LD7E] (noting that "the agency has not issued formal guidance on LDA topics to date despite urging by advocates and some industry stakeholders"); *see also* RELMAN COLFAX PLLC, FAIR LENDING MONITORSHIP OF UPSTART NETWORK'S LENDING MODEL: FOURTH AND FINAL REPORT OF THE INDEPENDENT MONITOR 19 (2024), https://www.relmanlaw.com/media/cases/1511_Upstart%20Final%20Report.pdf [https://perma.cc/6TG6-QRDX] (arguing that "more formal and detailed guidance is necessary").

184. While this Article was written, the CFPB noted that it had recently directed lenders to develop a process for the consideration of a range of less discriminatory models and again reiterated that robust fair lending testing of models should include searches for and implementation of less discriminatory alternatives using manual or automated techniques. *See* CFPB, FAIR LENDING REPORT OF THE CONSUMER FINANCIAL PROTECTION BUREAU 8, 36 (2024), https://files.consumerfinance.gov/f/documents/cfpb_fair-lending-report_fy-2023.pdf [https://perma.cc/5JXC-JDCR]. Such a directive aligns with our proposal. Notably, however, several groups have called upon the CFPB to provide even more specificity and clarity on how financial institutions should search for and implement LDAs. *See* Letter from Jennifer Chien, Sr. Pol'y Couns., Consumer Reps. & Adam Rust, Dir. Fin. Servs., Consumer Fed'n. of Am., to Rohit Chopra, Dir., CFPB (June 26, 2024), https://advocacy.consumerreports.org/wp-content/uploads/2024/06/240626-CR-CFA-Statement-on-Less-Discriminatory-Algorithms-FINAL.pdf [https://perma.cc/9CRJ-UMMA].

185. For example, the CFPB would need to update 12 C.F.R. §§ 1002.2, 1002.5, 1002.6, as well as the Official Interpretations, throughout.

186. *See generally* BD. OF GOVERNORS OF THE FED. RSRV. SYS. & OFF. OF THE COMPTROLLER OF THE CURRENCY, SUPERVISORY GUIDANCE ON MODEL RISK MANAGEMENT (2011), https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf [https://perma.cc/SSM3-GRVG].

187. In two recent guidance documents, HUD took steps to clarify this obligation in relation to tenant screening services and advertising through digital platforms. In one document, HUD suggested that one guiding principle for non-discriminatory tenant screening entailed "[e]valuating the data used by a model, creating test data, and rigorously testing to validate the model can reveal whether it has disparate outcomes and what less discriminatory alternatives might exist." *See* HUD, GUIDANCE ON APPLICATION OF THE FAIR HOUSING ACT TO THE SCREENING OF APPLICANTS FOR RENTAL HOUSING 14 (2024), https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Guidance_on_Screening_of_Applicants_for_Rental_Housing.pdf

need to expand its definition of "legally sufficient justification" to include reasonable efforts to search for and implement less discriminatory alternatives when the challenged practice is a housing algorithm.[188] Second, HUD would need to expand upon a defendant's burden of proof in a discriminatory effects case challenging an algorithmic system, requiring a defendant to demonstrate it took reasonable steps to search for and implement less discriminatory alternative algorithms.[189]

C. MODEL MULTIPLICITY AND REGULATORY GOVERNANCE

While existing civil rights laws clearly apply to covered entities when they rely on algorithms to make decisions, many have argued that these laws are inadequate to meet the challenges of algorithmic discrimination because the doctrine was developed with human decisionmakers in mind and their gaps in coverage mean that not all relevant actors are covered by the laws.[190] In addition, these laws primarily rely on a model of retrospective liability and depend on individual victims of discrimination to sue and to prove after-the-fact that discrimination occurred. This model is unlikely to be effective in preventing the sorts of pervasive, systemic harms that algorithmic decisionmaking can cause.

Given the limitations of a backward-looking, liability-focused regime, many researchers and policymakers have called for new regulatory tools to address the risk that algorithms may discriminate.[191] Scholars have proposed a variety of

[https://perma.cc/WTF3-AWFC]. In another document, HUD explicitly suggested that ad platforms should "[p]roactively identify and adopt less discriminatory alternatives for AI models and algorithmic systems, including by assessing data used to train AI models and verifying that the technologies measure lawful attributes that predict valid outcomes." See HUD, GUIDANCE ON APPLICATION OF THE FAIR HOUSING ACT TO THE ADVERTISING OF HOUSING, CREDIT, AND OTHER REAL ESTATE-RELATED TRANSACTIONS THROUGH DIGITAL PLATFORMS 12 (2024), https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Guidance_on_Advertising_through_Digital_Platforms.pdf [https://perma.cc/26HF-8NT2].

188. *See* 24 C.F.R. § 100.500(b)(1) (2017).

189. *See id.* § 100.500(c)(2).

190. *See, e.g.,* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1673–74 (2020); Talia B. Gillis, *The Input Fallacy*, 106 MINN. L. REV. 1175, 1239 (2022); Kim, *supra* note 3, at 866; Scherer, et al., *supra* note 66, at 494.

191. *See, e.g.*, DILLON REISMAN, JASON SCHULTZ, KATE CRAWFORD & MEREDITH WHITTAKER, AI NOW INST., ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY 15 (2018) [https://perma.cc/A2NX-LMEW]; Margot E. Kaminski, & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-Layered Explanations*, 11 INT'L DATA PRIV. L. 125, 138 (2021); Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. & TECH. 117, 140–47 (2021); ANNETTE BERNHARDT, LISA KRESGE, & REEM SULEIMAN, DATA AND ALGORITHMS AT WORK: THE CASE FOR WORKER TECHNOLOGY RIGHTS 2 (U.C. Berkeley Lab. Ctr. eds., 2021), https://laborcenter.berkeley.edu/wp-content/uploads/2021/11/Data-and-Algorithms-at-Work.pdf [https://perma.cc/5VZB-SRGS]; WHITE HOUSE OFF. OF SCI. & TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE 3 (2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [https://perma.cc/8DFG-WTEF]; Alex Engler, *A Comprehensive and Distributed Approach to AI Regulation: Proposing the Critical Algorithmic Systems Classification (CASC)*, BROOKINGS (Aug. 31, 2023), https://www.brookings.edu/articles/a-comprehensive-and-distributed-approach-to-ai-regulation/

regulatory interventions, including pre-market licensing regimes,[192] as well as new regulatory instruments and subpoena power.[193] To that effect, legislation has been introduced at the federal, state, and local levels seeking to use various policy levers to create new mechanisms to hold algorithms accountable and to require some level of transparency.[194] As with disparate impact doctrine, there remains a great deal of uncertainty about the details regarding how these regulatory tools should operate. Many proposals have been floated, but the best approach for governing algorithmic-decision systems remains contested and very much in flux.[195]

Two commonly proposed policy solutions are audits and algorithmic-impact assessments.[196] Scholars have frequently mentioned auditing as a necessary tool for diagnosing discriminatory impacts.[197] Algorithmic-impact assessments have also received significant scholarly and policymaker attention.[198] While many of these interventions are designed, at least in part, to reveal when systems lead to disparate effects on a prohibited basis, they do not offer clear guidance regarding what to do once an assessment or audit surfaces disparate impact. At bottom, these proposals mostly only require documentation and evaluation and critically stop short of requiring further action. Requiring a duty to search for LDAs helps answer the questions, (1) "*What* are we auditing for?" and (2) "Algorithmic impact assessments, in service of *what*?".

Practically, these regulatory requirements could be implemented through new legislation or rulemaking under existing authority, though rulemaking appears more likely. For example, the Federal Trade Commission is considering whether to promulgate a new trade rule as part of its ongoing Commercial Surveillance and Data Security rulemaking.[199] A new rule could require entities designing and deploying algorithmic systems in sensitive civil rights domains to maintain

---

[https://perma.cc/E5A4-VGNK]; Laws. Comm. for Civ. Rts. Under Law, *Online Civil Rights Act* (Dec. 2023), https://www.lawyerscommittee.org/online-civil-rights-act/ [https://perma.cc/VNT9-WQNL].

192. *See* Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83, 111 (2017).

193. *See* Engler, *supra* note 191.

194. *See, e.g.*, Algorithmic Accountability Act, H.R. 6580, 117th Cong. (2022); American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022); Stop Discrimination by Algorithms Act, B24-0558, 2021 Council, Reg. Sess. (D.C. 2021); Automated Decision Tools, A.B. 331, 2023–24 Leg., Reg. Sess. (Cal. 2023); S.B. 5356, 2023–24 Leg., Reg. Sess. (Wash. 2023).

195. *See* Kaminski, *Binary Governance*, *supra* note 4, at 1550–53.

196. *See* N.Y.C. ADMIN. CODE § 20–871 (2023).

197. *See, e.g.*, Kim, *supra* note 1, at 190; Ajunwa, *supra* note 4, at 661; Bryan Casey, Ashkon Farhangi & Roland Vogl, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 143, 182 (2019). *But see* Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633, 660–61 (2017) (describing the limitations of auditing as a tool for ensuring accountability of algorithms).

198. *See supra* note 194. For a broader discussion on how states are approaching the issue, see Sorelle Friedler, Suresh Venkatasubramanian & Alex Engler, *How California and Other States are Tackling AI Legislation*, BROOKINGS (Mar. 22, 2023), https://www.brookings.edu/articles/how-california-and-other-states-are-tackling-ai-legislation/ [https://perma.cc/978N-G63N].

199. Trade Regulation Rule on Commercial Surveillance and Data Security, 87 Fed. Reg. 51273, 51275–76 (Aug. 22, 2022).

reasonable and appropriate measures to address algorithmic discrimination.[200] An entity's failure to do so would be an unfair trade practice, while following such procedures would be considered compliant with the regulation.

IV. A CASE STUDY: THE UPSTART FAIR LENDING MONITORSHIP AND MODEL MULTIPLICITY

We now consider whether model multiplicity can be exploited in practical settings, describing a case study that involved a successful search for LDAs.

Upstart is a financial technology company that relies on machine learning and non-traditional applicant data, including data related to borrowers' higher education, to underwrite and price consumer loans. Due to civil rights concerns,[201] Upstart voluntarily agreed to submit its underwriting model to scrutiny by an independent Monitor,[202] which was tasked with assessing whether Upstart's model had an adverse impact on any protected group, and "if so, whether there are less discriminatory alternative practices that maintain the model's predictiveness."[203]

First, the Monitor tested Upstart's then-existing model, which it referred to as the Baseline Model, and determined that it exhibited practically and statistically significantly lower approval rates for Black applicants as compared with non-Hispanic white applicants.[204] In response, Upstart maintained that the model advanced its valid business need to predict the probability of default or prepayment.[205] The Monitor then turned to identifying whether a less discriminatory alternative existed.[206]

---

200. *See, e.g.*, Selbst & Barocas, *supra* note 4, at 1025; STEPHEN HAYES & KALI SCHELLENBERG, DISCRIMINATION IS UNFAIR: INTERPRETING UDA(A)P TO PROHIBIT DISCRIMINATION 5–6 (Student Borrower Prot. Ctr. eds., 2021), https://protectborrowers.org/wp-content/uploads/2021/04/Discrimination_is_Unfair.pdf [https://perma.cc/T6CD-NVKN].

201. *See* STUDENT BORROWER PROT. CTR., EDUCATIONAL REDLINING 4 (2020), https://protectborrowers.org/wp-content/uploads/2020/02/Education-Redlining-Report.pdf [https://perma.cc/YY53-U5J5]; Letter from Senators Sherrod Brown, Elizabeth Warren, Robert Menendez, Cory Booker & Kamala D. Harris to Dave Girouard, CEO, Upstart (Feb. 13, 2020), https://www.brown.senate.gov/imo/media/doc/2020-02-13%20Senate%20letter%20to%20Upstart.pdf [https://perma.cc/7N8X-C2ZK]; S. COMM. ON BANKING, HOUS. & URB. AFFS., USE OF EDUCATIONAL DATA TO MAKE CREDIT DETERMINATIONS 1, https://www.banking.senate.gov/imo/media/doc/Review%20-%20Use%20of%20Educational%20Data.pdf [https://perma.cc/N7Z3-K68H].

202. The agreement was between Upstart, the Student Borrower Protection Center (SBPC), and the NAACP Legal Defense Fund (LDF). Two authors of this Article (Logan Koepke and Mingwei Hsu) served as technical advisors to NAACP-LDF during the Monitorship. All information in this Article is drawn exclusively from the public Monitorship reports. No confidential information has been shared or is reproduced here in any way.

203. RELMAN COLFAX PLLC, *supra* note 183, at 7.

204. RELMAN COLFAX PLLC, FAIR LENDING MONITORSHIP OF UPSTART NETWORK'S LENDING MODEL: SECOND REPORT OF THE INDEPENDENT MONITOR 3 (2021), https://www.relmanlaw.com/media/cases/1180_PUB_LIC%20Upstart%20Monitorship_2nd%20Report_FINAL.pdf [https://perma.cc/3E9Y-4CDB].

205. *Id.* at 18.

206. *Id.*

The Monitor relied on two main procedures to explore alternatives. First, it searched through every possible subset of the original model's input features to identify combinations that yielded reductions in disparate impact.[207] In other words, supposing that Upstart's model included features A, B, and C, the search process examined models trained on every possible combination of these inputs: A and B, A and C, and B and C. Second, the Monitor turned to hyperparameter tuning.[208] Hyperparameters are inputs provided to the machine learning algorithm that guide the learning process. Hyperparameter tuning is the process of finding optimal values of hyperparameters that govern the model architecture or training process. For example, hyperparameter tuning of a random forest model would include optimizing the number of decision trees in the forest, the number of features considered by each tree, the number of levels in each decision tree, and so on.

To evaluate whether an alternative performed comparably to the Baseline Model, the Monitor turned to Upstart's primary performance metric,[209] which predicts both default risk and prepayment risk (i.e., the risk that a loan recipient might pay off a loan on a shorter repayment schedule, thereby reducing the total interest that could be charged) and is reported as an average.[210] This performance metric inherently has some degree of uncertainty. This uncertainty can be computed as a probability range: the Uncertainty Interval. By the Monitorship's reasoning, if the performance of an alternative model yields a value that falls within that Uncertainty Interval, there is a strong argument that it reasonably serves the same purpose as the Baseline Model, since the Baseline Model was already deemed acceptable to Upstart's business purposes. Because "there is inherent and measurable uncertainty around how a model will perform[, this] suggests that . . . alternative models within those bounds of uncertainty are likely to be similarly effective at meeting business needs."[211] Based on statistical testing, the Monitor developed two Uncertainty Intervals: a 95% Uncertainty Interval band of +/− 20 performance metric points from the average performance, and a 68% Uncertainty Interval band of +/− 10 performance metric points from the average performance.[212] In other words, a narrower boundary of equivalent performance

207. *Id*. at 19.

208. *Id*. at 19–20.

209. While the Monitor refers to the performance metric as the "Error Metric" in its reports, we simply refer to this metric as the performance metric for simplicity.

210. Upstart trains and validates its model using 5-fold cross validation, so for each of the five cross validations, a performance metric is calculated. Consider a toy example in which five cross validations return performance metric values of 1000, 1010, 980, 1012, and 1018. The average of the performance metric would be 1004. *See* RELMAN COLFAX PLLC, *supra* note 52, at 16.

211. *Id.*

212. *Id.* at 17–18. The first band (95% Uncertainty Interval, +/−20 performance metric points) is "wider" in the sense that it allows for a greater difference in accuracy, with higher confidence. The second band (68% Uncertainty Interval, +/− 10 performance metric points) is "narrower" in the sense that it allows for less difference in accuracy, albeit with less confidence. In other words, the first band is higher-confidence, but potentially more variance, whereas the second band is lower-confidence, but less variance. When referring to "+/− 20 performance metric points" or "+/− 10 performance metric

(68% Uncertainty Interval) and a wider boundary (95% Uncertainty Interval). As we described in Part I, identifying equally performing models requires defining some bound ε within which differences in performance should be considered equivalent. The Monitorship's use of the Uncertainty Interval is an example of establishing this bound of ε, within which differences in performance should be considered equivalent.[213]

The Monitor ultimately identified multiple viable LDAs. One LDA resulted in statistically significant reductions in disparities in approval rates for Black and Hispanic applicants compared with white applicants, and its performance metric fell within the narrower 68% uncertainty range.[214] Another LDA resulted in more statistically significant improvements in approval disparities for Black and Hispanic applicants, but its performance metric fell only within the 95% Uncertainty Interval.[215]

However, Upstart claimed that the Monitor's alternative "would unacceptably compromise the accuracy of its models"[216] and "declined to adopt the Monitor's recommended approach"[217] towards finding LDAs. The parties' disagreement centered on what the appropriate and legally required methodology is to assess whether a less discriminatory algorithm performs comparably to an existing baseline model in meeting a company's asserted legitimate business need. As a result, the Monitorship reached an impasse. Despite this impasse, the Monitorship's search for an LDA is a real-world demonstration of the promise of model multiplicity. Further, the parties agreed that "[r]egulators should provide guidance on how effective fair lending testing should be conducted, including clarifying expectations regarding identifying, assessing, and adopting less discriminatory alternative models."[218]

Separately, while the Monitorship was successful in identifying viable LDAs, it took almost two years[219] to complete this search, even in a cooperative, non-litigation setting. By the time the Monitorship presented its findings and furnished a less discriminatory alternative, Upstart had already changed its model, meaning that the proffered alternatives were less discriminatory than a Baseline M

points" for each Uncertainty Interval band, these intervals are expressed in units of the primary performance metric of "points," not as a percentage range or basis points. *Id.*

213. The Monitor imposed several additional criteria that an alternative would have to satisfy for the Monitor to recommend it as a viable alternative. For example, the Monitor would not recommend an alternative model if it introduced new disparities for other protected classes that were not present in the original model. *Id.* at 19. Such additional constraints likely reduced the total number of alternatives that could have been considered viable.

214. *See* RELMAN COLFAX PLLC, *supra* note 52, at 24.

215. *Id.* at 24–25.

216. Press Release, Legal Defense Fund, LDF, SBPC, and Upstart Announce Final Monitorship Report on AI and Fair Lending (Mar. 27, 2024), https://www.naacpldf.org/press-release/ldf-sbpc-and-upstart-announce-final-monitorship-report-on-ai-and-fair-lending/ [https://perma.cc/F5ZG-6GHK].

217. RELMAN COLFAX PLLC, *supra* note 183, at 14.

218. *Id.* at 4.

219. On December 1, 2020, the parties agreed to appoint an independent Monitor. RELMAN COLFAX PLLC, *supra* note 51, at 25. On September 16, 2022, the Third Report of the Independent Monitor, which identified viable LDAs, was published. *See* RELMAN COLFAX PLLC, *supra* note 52.

odel that had been in use, but was no longer relevant.[220] Had Upstart been required in the first instance to take reasonable steps to search for and implement LDAs, it's possible, even likely, they would have discovered an LDA more quickly and at less cost and avoided unnecessary disparate impacts on real borrowers. Finally, while offering a helpful example of a search for LDAs, the Monitorship only relied on a limited set of techniques to perform its search.[221] In Section V, we survey a larger variety of potential interventions available to practitioners throughout the model development pipeline that can be explored to search for LDAs.

V. THE DUTY TO SEARCH FOR AND IMPLEMENT LDAS IN PRACTICE

In this section, we examine the specific steps that firms should be expected to take to fulfill their duty to search for and implement LDAs in practice. We first describe the processes that firms need to put in place as a basic requirement of the duty. We then consider the degree to which costs may limit what firms are expected to do as part of these processes. In so doing, we argue that firms must take *reasonable steps* to search for and implement LDAs, where reasonableness is determined by the costs of interventions, evidence-based best practices, and the severity of the disparate impact at issue. Building on this discussion and drawing on a range of interventions, we then describe the specific kinds of exploration that one might reasonably expect to see covered entities perform in practice.

A. BASIC REQUIREMENTS OF THE DUTY

The capacity to fulfill *any* kind of duty to search for and implement an LDA depends on four related processes, each of which is a basic requirement of the duty. First, firms must have a process in place for collecting or inferring the demographic information necessary to perform a disparate impact analysis.[222]Absent

220. *See* RELMAN COLFAX PLLC, *supra* note 52, at 29 (noting that while the Monitor would likely have recommended Upstart adopt Model 2, "Upstart updated its model prior to completion of these analyses").

221. As mentioned, the Monitor searched through subsets of the original model's inputs and engaged in hyperparameter tuning. *Id.* at 30.

222. Companies will either need to collect this information in a legally permissible fashion or rely on inference methodologies. As a result, existing civil rights law that either expressly prohibits companies from collecting demographic information—or is otherwise silent as to the collection of demographic data for anti-discrimination testing purposes—will need to change. For example, Regulation B, which implements the ECOA, says that a "creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction," with certain exceptions. *See* 12 C.F.R. § 1002.5(b). This general prohibition stands in stark contrast to mortgages. Under Regulation C, which implements the Home Mortgage Disclosure Act, lenders are required to collect certain demographic information. *See* 12 C.F.R. § 1003.4(a)(10). More recently, the Dodd-Frank Wall Street Reform and Consumer Protection Act amended the ECOA, such that covered financial institutions are required to collect and report to the CFPB data on applications for credit for small businesses, including those that are owned by women or minorities. *See* Dodd-Frank Wall Street Reform and Consumer Protection Act, Pub. L. No. 111-203, 124 Stat. 2056 (2010) (codified as amended at 15 U.S.C. §1691(c)(2)). For more discussion, see Miranda Bogen, Aaron Rieke & Shazeda Ahmed, Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination, *in* FAT '20: PROCEEDINGS OF THE 2020 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND

information about, for example, the gender of the people whose data are being used to evaluate a model's performance, firms will be unable to establish whether the model's performance and selection rate differs by gender. Additionally, given the sensitivity of these data, firms must also adopt appropriate policies and procedures to protect the data and limit their use for unrelated purposes. Second, firms must have a process for actually performing the disparate impact analysis itself. Notably, this must include a process for evaluating a model for disparate impact both before deployment and on an ongoing basis once it has been deployed. Third, firms must establish a process for searching for LDAs. Firms should apply this process when developing a model, where the search for LDAs should be incorporated into the model development process from the start, and after a model has been developed or deployed, when addressing an identified disparate impact. A key part of this process also includes documenting the point at which the firm decides to bring its search to a close—that is, why the firm believes it has done enough, under its particular constraints, to search for an LDA.[223] Finally, firms must have a process for determining when they will adopt an LDA and for implementing the LDA in practice.

If firms do not have these processes in place or cannot explain how they go about each process, they will have failed to fulfill their duty. Without these

---

TRANSPARENCY, *supra* note 55, at 492–93. Limited availability of protected attribute data in government and private organizations frequently makes assessing algorithmic fairness and training fairness-constrained systems difficult. However, there is a growing literature around developing methods for measuring and mitigating discrimination in machine learning models with noisy, incomplete, or even no access to protected attribute information. *See, e.g.*, Hadi Elzayn et al., *Measuring and Mitigating Racial Disparities in Tax Audits*, STAN. INST. FOR ECON. POL'Y RSCH. (Jan. 30, 2023), https://siepr.stanford.edu/publications/measuring-and-mitigating-racial-disparities-tax-audits; Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li & Yang Liu, Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes, *in* ICML '22: PROCEEDINGS OF THE 40TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (Ass'n for Computing Mach. eds., 2022), https://arxiv.org/abs/2210.03175; Aaron Rieke, Vincent Southerland, Dan Svirsky & Mingwei Hsu, Imperfect Inferences: A Practical Assessment, *in* FACCT '22: PROCEEDINGS OF THE 2022 5TH ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, *supra* note 6, at 767, https://dl.acm.org/doi/10.1145/3531146.3533140; Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov & Jonathan Tannen, *How Meta Is Working to Assess Fairness in Relation to Race in the U.S. Across Its Products and Systems*, META AI (Nov. 2021), https://ai.meta.com/research/publications/how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems/ [https://perma.cc/T5XS-Z8KM]; Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta & Michael Jordan, Robust Optimization for Fairness with Noisy Protected Groups, *in* NEURIPS '20: PROCEEDINGS OF THE 34TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (2020), https://proceedings.neurips.cc/paper/2020/hash/37d097caf1299d9aa79c2c2b843d2d78-Abstract.html; Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha & Madeleine Udell, Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved, FAT* '19: PROCEEDINGS OF THE 2019 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, *supra* note 30, at 339, https://arxiv.org/abs/1811.11154; CFPB, USING PUBLICLY AVAILABLE INFORMATION TO PROXY FOR UNIDENTIFIED RACE AND ETHNICITY: A METHODOLOGY AND ASSESSMENT (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf [https://perma.cc/X6GG-SEBK].

223. By "bring its search to a close," we simply mean the current search process at hand, not a permanent close. We expect firms to perform another search each time they re-train a model, observe a change in disparate impact, or there is reason to believe that the science of LDA search may have new insights to offer.

processes, firms cannot take reasonable steps to determine whether a disparate impact might be avoided by searching for and adopting an LDA. Of course, satisfying these basic requirements alone may not fulfill the duty because the processes actually adopted may still fall short of what is reasonable. These processes could be far from robust—poorly thought through, poorly resourced, and poorly executed. Beyond these basic requirements, we argue that firms must take *reasonable steps* to search for and implement LDAs.

B. REASONABLE STEPS

Crucial to what counts as *reasonable steps* is what resources regulated entities should be expected to devote to the search for a less discriminatory alternative. As discussed in Section II.C, the law is ambiguous about how costs factor into the determination of whether a less discriminatory alternative is available. While courts have rejected proposed alternatives that would impose significant costs on defendants,[224] courts and regulators have also endorsed alternatives that are clearly not costless.[225] These decisions suggest that there is an expectation that regulated entities incur *reasonable* costs in seeking to avoid disparate impact.

In this Article, we have adopted a relatively conservative definition of an LDA—namely, it must exhibit accuracy equal to the challenged algorithm. This definition means that LDAs will not impose on firms the cost of a loss in model performance. Aside from performance costs, however, there remains a question about the degree to which firms should be expected to shoulder the monetary costs involved in developing or administering LDAs. In what follows, we therefore set aside questions about *performance* costs—the cost imposed on regulated entities if compelled to adopt an LDA of lower performance—and focus on *development* and *administrative* costs—the expense involved in finding and fielding an LDA.

While there may be limits to how much disparate impact can be reduced by searching among alternative models of equal accuracy, the major bottleneck will likely be the cost of that search, not what the search itself can uncover. In other words, the question moves from one about the comparative effectiveness of proposed alternatives to one about the costs that defendants should be reasonably expected to incur in searching for and implementing the alternatives.[226]

This is an important question because exploiting multiplicity is not a costless undertaking. To identify models that exhibit equivalent accuracy but differ with respect to their disparate impact, regulated entities will need to take additional steps beyond those that they would take if they were only concerned with

224. *See supra* note 138 and accompanying text.

225. *See supra* note 139.

226. Penny Crosman, *Consumer Groups to CFPB: Make Banks Fairness-test AI Lending Software*, AM. BANKER (June 27, 2024, 3:25 PM), https://www.americanbanker.com/news/consumer-groups-to-cfpb-make-banks-fairness-test-ai-lending-software [https://perma.cc/P2R5-KN2N] (noting that a vendor working with lenders, FairPlay, did not discover suitable LDAs when the vendor did 100 searches, but did find LDAs that "were both accurate and fairer" when performing 300 searches, suggesting that the primary constraint is resources).

accuracy. Given the costs of these interventions, how far should defendants be expected to go in performing such a search?

Furthermore, models of equal accuracy may not cost the same amount to administer. For example, models purposefully designed to take in a larger number of features to reduce disparate impact can be more difficult to execute in practice than simpler models because they require collecting and inputting more data each time a firm wishes to make a decision. In effect, each decision is more expensive. What are the additional expenses that firms should be reasonably expected to shoulder?

An easy starting point is that firms should not be allowed to invoke sunk costs. Sunk costs are investments in algorithms that would never have been deployed had the firm taken concerns with disparate impact into account from the start. Under existing law, firms have a clear responsibility to avoid practices with unjustified disparate impacts. And as understanding about model multiplicity and the possibilities of uncovering LDAs diffuses, a firm's decision to press ahead with an algorithm that results in an avoidable disparate impact becomes less and less justifiable. If it does so, it should not be allowed to invoke the costs of this foolish investment as a basis for refusing to adopt a less discriminatory alternative.

But how might we determine what counts as reasonable costs beyond this one easy case? At a minimum, it would be reasonable to expect firms to make any investment to reduce disparate impact that covers its own costs. For example, firms should be obligated to collect more features if doing so helps to reduce disparate impact—and if the cost of doing so is covered by the additional benefits that firms enjoy from the resulting improvements in the accuracy of the decisionmaking process.[227] In theory, firms should have incentives to do this already, but it is possible that they are not doing nearly as much as they could.[228]

Further still, firms should be expected to incur the costs of following evidence-based best practices—that is, the processes and procedures that independent research demonstrates to be effective and that experience suggests are possible for firms to execute in practice. While best practices are not precise standards, they have the benefit of evolving alongside advances in our understandings and capabilities. They can also be responsive to norms regarding the amount of resources that industries are devoting to the problem, helping to

227. This would exceed our stricter definition of an LDA because the model would be *more* accurate, not just equally accurate. We nevertheless include this in our set of reasonable steps because it would result in regulated entities adopting a less discriminatory alternative at no additional expense.

228. This might be true for a number of reasons. First, firms might not bother to make such investments if the benefits of doing so are modest, which is especially likely if the improvements are limited to a minority population that represents a small number of customers. Second, firms might incur an opportunity cost in allocating resources to address this issue rather than other issues that would be more profitable to address. Finally, firms might struggle to estimate the expected benefits of these investments and thus disfavor interventions that are not certain to cover their costs. *Cf.* Kim, *supra* note 3, at 894–97 (explaining why market forces are unlikely to eliminate biased employment selection algorithms).

catch individual firms that invest noticeably less in addressing avoidable disparate impact than their competitors. Best practices might require more from firms as the market for mitigations grows and costs drop.[229] While following such practices may impose costs on firms, defining reasonable steps according to prevailing evidenced-based best practices helps to ensure that no firm is at a competitive disadvantage and companies cannot out-compete each other by simply refusing to shoulder these costs.

Finally, what counts as reasonable steps should also depend on the magnitude of the disparate impact: the more severe the disparity, the more costs firms should be expected to bear in searching for and implementing an LDA. This principle follows certain courts' reasoning that it may be reasonable to expect firms to bear some costs if doing so significantly reduces disparate impact. Similarly, firms should invest more resources in the search for an LDA where an algorithm's harmful effects on a disadvantaged group are more pronounced.

C. SEARCHING FOR LDAS IN PRACTICE

In this Section, we survey a variety of potential interventions available to practitioners throughout the model development pipeline that could form part of a reasonable search for an LDA. The methods listed here are in no way exhaustive and serve partially as a jumping-off point to various related literatures.

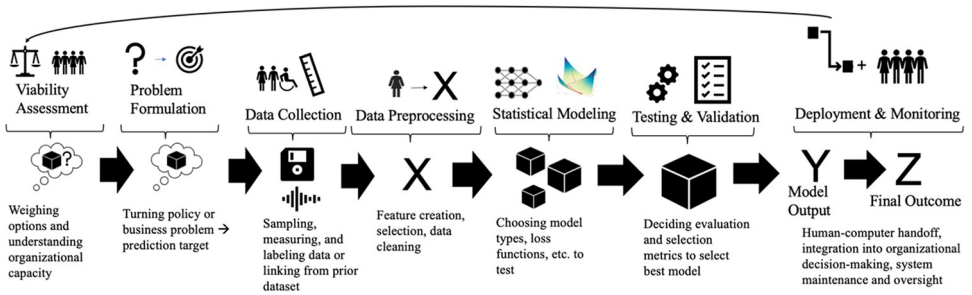
We follow our brief exploration of potential interventions with discussion of their cost. The extent of exploration deemed reasonable will depend on the resources available to a given company; some techniques may only be viable for well-resourced firms, while others may be appropriate for poorly-resourced firms. We also note that research in this space is rapidly evolving. Methods and interventions that are costly today may become more viable in the near future—even for companies with few available resources.

1. General Methodology to Search for LDAs

As discussed in Part I and depicted in Figure 3 below, predictive models are developed through an iterative, cyclic series of decisions along the model-development pipeline. The primary way to search for LDAs is to intervene along this pipeline, exploring alternative design choices in order to create and evaluate alternative models. Ideally, this search should occur during the initial model-development process to minimize cost and harm. Developing a machine learning model already involves exploring alternative models and testing their performance, so testing for discrimination as part of this process should not create a significant additional burden.

229. Of course, as discussed in Section III.B, this would not limit individual plaintiffs' ability to proffer alternatives that go well beyond current best practices. Even though firms would only be found liable of discrimination if they failed to follow best practices, they would nevertheless be compelled to adopt the alternative put forth by plaintiffs.

Figure 3. An illustration of some of the main components of the machine learning development pipeline.



Several other works in the algorithmic-fairness and legal literatures have outlined interventions aimed at reducing disparity in the model building process.[230] In particular, one recent article offers a wider survey of methods which leverage design choices made along the model-development pipeline to reduce disparate impact.[231] These methods are also useful for searching for LDAs. While searching for LDAs, practitioners will look to reduce disparity and maintain accuracy within ε. However, it is difficult to know a priori if a given intervention will worsen, improve, or not affect accuracy. Thus, to search for LDAs, a variety of interventions should be used, and the resultant alternative models with varying performance should be evaluated for model accuracy and disparity to see if they qualify as an LDA.

With this framing in mind, we step through examples of potential intervention points and particular interventions throughout the model development pipeline as depicted in Figure 3 that could be used to reduce disparate impact.

2. Examples of Interventions

Problem Formulation: One often overlooked point of intervention to reduce disparate impact is the problem formulation stage of model creation (i.e., the translation of a real-world problem into a machine learning task).[232] How an organization chooses a numerical proxy to stand in for its overall goals can profoundly affect model behavior, including disparate impact.[233] Recent research by Black et al. and Elzayn et al. into models used by the IRS to predict the likelihood of tax

230. *See, e.g.*, Mehrabi et al., *supra* note 58; Suresh & Guttag, *supra* note 44; Black et al., *supra* note 59. For a discussion of various approaches to bias detection and mitigation for a computer science audience, see generally Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J.L. ANALYSIS 113 (2018).

231. Black et al., *supra* note 59.

232. *See* Passi & Barocas, *supra* note 30, at 41.

233. *See, e.g.*, Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCIENCE 447, 454 (2019).

noncompliance provides a clear example.[234] They show that changing the problem formulation of IRS audit selection models from whether individuals are likely to be noncompliant at all (e.g., with binary labels, describing if they were compliant or not) to predicting the amount of money they failed to report (i.e., with continuous labels of the amount of taxes owed), the distribution of those recommended for audit by the algorithm shifted from lower-income and Black individuals towards higher-income and more white individuals, thus reducing disparate impact.[235]

While we are unaware of any current, off-the-shelf tools to guide developers' consideration of a variety of prediction tasks for a given business problem, it should be rather obvious to developers how to go about exploring alternatives. They can consider a variety of different prediction targets, train models with those targets, and then compare their performance and disparate impact. For example, a lending firm could compare the disparity induced by models that define default as 12 weeks of non-payment, 16 weeks of non-payment, or 20 weeks of non-payment.

Data collection: Interventions to reduce disparate impact during data collection (i.e., the process of gathering data with which to train and evaluate a model) are particularly well-studied. A famous example is differential accuracy across faces with different skin tones in facial-analysis models due to a dearth of representation of darker faces in training datasets. [236] For interventions during the data collection process, we refer readers to the large literature outlining measures to identify and prevent disparities from arising in AI systems because of bias in data collection.[237]

However, at the very least, robust testing of data quality across demographic groups throughout data collection is of paramount importance—for example, testing demographic representation, testing rates of feature missingness across groups, testing for spurious correlations that may impact prediction, and testing the predictiveness of their data across demographic groups can help identify if new data should be collected in hopes of reducing disparities. Though adding

234. *See generally* Emily Black, Hadi Elzayn, Alexandra Chouldechova, Jacob Goldin & Daniel Ho, Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models, *in* PROCEEDINGS OF 2022 5TH ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, *supra* note 6, at 1479, https://dl.acm.org/doi/10.1145/3531146.3533204; Elzayn et al., *supra* note 222.

235. Elzayn et al., *supra* note 222, at 5; *see also* Black et al., *supra* note 234, at 1486.

236. *See, e.g.*, Joy Buolamwini & Timnit Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, *in* PROCEEDINGS OF 2018 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1, 10 (2018).

237. *See, e.g.*, Mehrabi et al., *supra* note 58; Mora Geva, Yoav Goldberg & Jonathan Berant, Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets, *in* PROCEEDINGS OF THE 2019 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE 9TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING 1161 (Ass. Computational Linguistics, 2019), https://arxiv.org/abs/1908.07898; Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh & Lora M. Aroyo, "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI, *in* PROCEEDINGS OF THE 2021 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS Paper 39 (Ass'n for Computing Mach. eds., 2021), https://dl.acm.org/doi/10.1145/3411764.3445518.

new data, especially that of underrepresented demographic groups, can be costly and challenging, recent work outlines how to choose datapoints to search for and add to a machine learning system to maximize decreases in selection rate disparity and other notions of discrimination.[238] Finally, especially in the case of novel data collection, using a dataset documentation system—such as Datasheets for Datasets[239]—can unearth some less-than-ideal choices made during the data collection process that could be remedied to reduce discrimination in alternative models, and potentially even increase performance.

Data preprocessing: Data preprocessing concerns the steps machine learning practitioners take to make data usable by the machine learning model—for example, turning images into numbers or filling in missing values in credit application forms.

Deciding how to prepare data for algorithm consumption can have a large impact on disparities. Interestingly, in one example, Wan et al. point out that accuracy differences across different languages in language translation models are partially related to the size of the base piece of language that the model works with—which is usually a word.[240] They show that longer words degrade model performance, and that breaking down long words into sub-words so that the smallest language units across languages are roughly the same size can mitigate the performance disparities across translations for different languages. While some automated preprocessing frameworks offer suggestions for how to clean data to maximize predictive performance,[241] few tools can automatically determine what changes in preprocessing are most helpful for reducing disparities. However, some preliminary interventions to explore include experimenting with different data preprocessing choices and testing their impact on desired disparity metrics, such as various imputation and dropping techniques, data transformations, and methods of encoding data as features, as these have been shown in a few works to impact fairness performance.[242]

---

238. *See, e.g.*, Irene Y. Chen, Fredrik D. Johansson, & David Sontag, Why Is My Classifier Discriminatory?, 32ND CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 3539, 3544 (Curran Assocs. eds., 2018), https://arxiv.org/abs/1805.12002.

239. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III & Kate Crawford, *Datasheets for Datasets*, 64 COMMC'NS ACM 85 (2021).

240. Ada Wan, Fairness in Representation for Multilingual NLP: Insights from Controlled Experiments on Conditional Language Modeling, *in* INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS 1, 2 (2022), https://openreview.net/pdf?id=-llS6TiOew [https://perma.cc/5MQ4-RZVD].

241. *See, e.g.*, Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang & Martin Zinkevich, Data Validation for Machine Learning, *in* PROCEEDINGS OF MACHINE LEARNING AND SYSTEMS 1, 10 (Proc. Mach. Learning & Rsch., 2019), https://mlsys.org/Conferences/2019/doc/2019/167.pdf.

242. *See, e.g.*, Vincent Jeanselme, Maria De-Arteaga, Zhe Zhang, Jessica Barrett & Brian Tom, Imputation Strategies Under Clinical Presence: Impact on Algorithmic Fairness, PROCEEDINGS OF THE 2ND MACHINE LEARNING FOR HEALTH 12 (2022), https://arxiv.org/abs/2208.06648; Sumon Biswas & Hridesh Rajan, Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline, *in* PROCEEDINGS OF THE 29TH ACM JOINT MEETING: EUROPEAN SOFTWARE ENGINEERING CONFERENCE AND SYMPOSIUM ON THE FOUNDATIONS OF SOFTWARE ENGINEERING (Ass'n for

Feature selection: Feature selection, often intertwined with data collection and preprocessing, refers to the process of selecting which features from the available data will be inputs to a machine learning model. As highlighted in Part IV, feature selection was the main intervention used in the Upstart Monitorship. This method involved creating models that had various combinations of features as inputs and comparing their discriminatory effects.[243]

Testing a variety of feature combinations from available data is a relatively low-cost and low-effort method of searching for an LDA. That said, doing an exhaustive search over *all* possible feature combinations is not the most efficient way to search the feature space for alternative combinations leading to less discriminatory impact. As has been explored in prior work,[244] there are automated methods of exploring feature combinations that reduce the search space to combinations of features that have a high chance of producing a less discriminatory impact. Especially given the availability of such tools, companies could easily consider various permutations of input features to the model as a part of the search for an LDA.

Statistical modeling: Decisions surrounding what type of model will be used and how it will be trained refer to decisions about statistical modeling. Decisions here include, for example, choosing the model type (e.g., a simple linear model or a more complex neural network) used to generate predictions; how exactly the model will be trained (i.e., the particular learning rule that determines how a model responds to information during training); and a loss function (i.e., a precise definition of what behaviors a model will be penalized and/or rewarded for during training).

Much of the work in fairness-in-machine-learning considers the effects of changing the model's incentives during training (e.g., adding a constraint to a model's loss function that explicitly prioritizes various equity goals, such as equalizing selection rates across demographic groups).[245] However, each decision made during the statistical modeling process can have an impact on model

Computing Mach. eds., 2021), https://dl.acm.org/doi/10.1145/3468264.3468536 [https://perma.cc/T2A4-939B].

243. *See* RELMAN COLFAX PLLC, *supra* note 52, at 24.

244. *See* Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou & Baowen Xu, Training Data Debugging for the Fairness of Machine Learning Software, *in* ICSE '22: PROCEEDINGS OF THE 44TH INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING, 2215 (Ass. for Computing Mach., 2022), https://dl.acm.org/doi/abs/10.1145/3510003.3510091; FINREGLAB, LAURA BLATTNER & JANN SPIESS, MACHINE LEARNING EXPLAINABILITY & FAIRNESS: INSIGHTS FROM CONSUMER LENDING (July 2023), https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-07-13_Empirical-White-Paper_Explainability-and-Fairness_Insights-from-Consumer-Lending.pdf [https://perma.cc/P9KX-734C].

245. *See, e.g.*, Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford & Hanna Wallach, A Reductions Approach to Fair Classification, *in* PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 1 (Proc. Mach. Learning & Rsch., 2018), https://arxiv.org/abs/1803.02453 [https://perma.cc/826K-Z8GL]; Moritz Hardt, Eric Price & Natan Srebro, Equality of Opportunity in Supervised Learning, *in* 30TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1, 2 (2016), https://arxiv.org/abs/1610.02413 [https://perma.cc/5VXX-K3E8]. The extent to which such approaches are legally admissible is a matter of debate. *See generally* Kim, *supra* note 61, at 1564–66, 1574–83 (reviewing debate and arguing that many race-conscious debiasing techniques are permissible under anti-discrimination law).

prediction behavior, including outcome disparities. For example, prior work has shown that the choice-of-learning rule (the procedure by which the model learns from data) can influence the extent to which a model amplifies underlying differences in base rates in its selection rates.[246]

Model practitioners should test a variety of statistical modeling choices to the extent possible to search for models with lower disparity. Practitioners are already in the habit of testing several different combinations of modeling decisions to find the best performing combination. As we discuss below, a disparity metric displayed alongside accuracy as another number to consider during statistical modeling could easily be added to this process to aid the search for LDAs. When possible, practitioners should make use of the wide array of methods available to explicitly reduce disparity during, for example, model training.[247]

Model testing and validation: The processes by which a model is determined to be performing well, both in relation to other models in the training set and on unseen data, are referred to as model testing and validation. If fairness behavior is not explicitly added as a part of the model selection criteria, it is unlikely that the fairest model among those under consideration will be chosen.

One immediate intervention at this stage is to add a disparity metric as a method of choosing between (at least equally accurate) models. Common machine learning model development software offers built-in methods to choose between a variety of similar models based on accuracy. If disparity metrics were to be added as a secondary metric or tiebreaker, it would be easy to automatically explore whether it's possible to reduce disparate impact without sacrificing accuracy.[248] Recent work in the machine learning literature lends credence to the fact that tuning hyperparameters for fairness goals—even after the rest of the model is trained to maximize accuracy—is effective at discovering models with very similar performance, but with reduced disparity.[249] Importantly, however, the reductions in disparity must be robustly tested to ensure that they will generalize to unseen data—this can be accomplished by assessing fairness behavior with the same rigor as accuracy or other notions of performance, such as, for example, using cross-validation or extra hold-out sets.[250]

---

246. Klas Leino, Emily Black, Matt Fredrikson, Shayak Sen, & Anupam Datta, Feature-Wise Bias Amplification, *in* INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS 5 (2019), https://arxiv.org/abs/1812.08999 (demonstrating how gradient descent can lead to increased bias in certain data regimes).

247. We note the legal concerns around this process in Part I.

248. Automated hyperparameter tuning methods in this package could be easily modified to add a disparity metric in their search. *See, e.g.*, Fabian Pedregosa et al., *Scikit-Learn: Machine Learning in Python*, 12 J. MACH. LEARNING RSCH. 2825, 2825–30 (2011).

249. *See, e.g.*, Robin Schmucker, Michele Donini, Valerio Perrone, Muhamad Bilal Zafar & Cédric Archambeau, Multi-Objective Multi-Fidelity Hyperparameter Optimization with Application to Fairness, *in* 34TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 1 (2020), https://www.amazon.science/publications/multi-objective-multi-fidelity-hyperparameter-optimization-with-application-to-fairness [https://perma.cc/3Z3H-7GRP].

250. Emily Black, Talia Fillis & Zara Yasmine Hall, D-Hacking, *in* FACCT '24: PROCEEDINGS OF THE 2024 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 602, 610 (Ass'n for Computing Mach. eds., 2024), https://dl.acm.org/doi/10.1145/3630106.3658928.

We offer one further suggestion on how to leverage the machine learning development pipeline to search for LDAs: monitoring for changes in selection rate or accuracy rates across demographic groups after model deployment. Although *monitoring* (i.e., examining a model's behavior to ensure there is no degradation in performance over time) does not directly lead to the discovery of LDAs, it can signal the need to search for an LDA if a model's predictive behavior starts to be discriminatory after deployment. Research has shown that machine learning systems can become more discriminatory over time, often because the population to which it is applied drifts further from the population over which it was trained.[251] Entities that use machine learning models to allocate opportunities or resources can and should monitor for disparate impact throughout deployment of the model, with a plan to investigate disparate impact should it be found. There are many automated monitoring pipelines available that check for degradations in predictive performance.[252]

D. COSTS

While the actual cost of any given intervention will depend heavily on the circumstances under which a firm is operating, interventions that disrupt less of the existing pipeline are likely to be cheaper than those that involve greater change. Exploring the effects of different decisions further along in the modeling process can be very low cost. For example, it would be practically trivial to add a disparity metric to hyperparameter tuning, and doing so would help to automate the process of exploring alternatives. This addition would add to the cost of the computation (i.e., the number of calculations that need to be performed and the resources that need to be consumed to do so), but these costs are likely to be modest for the relatively simple models commonly used in domains subject to discrimination law. Likewise, interventions that take the available set of features as a given but explore whether different subsets of these features might lead to models of comparable performance but with less disparate impact can be rather inexpensive, especially when the process of exploring different combinations of features is automated using available computational methods.

---

251. *See, e.g.*, Harvineet Singh, Rina Singh, Vishwali Mhasawade & Rumi Chunara, Fairness Violations and Mitigation under Covariate Shift, *in* FAccT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 3 (Ass'n for Computing Mach. eds., 2021), https://dl.acm.org/doi/10.1145/3442188.3445865; Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan & Supriyo Chakraborty, Fair Transfer Learning with Missing Protected Attributes, *in* AIES '19: PROCEEDINGS OF THE 2019 AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 91 (Ass'n for Computing Mach. eds., 2019), https://dl.acm.org/doi/pdf/10.1145/3306618.3314236. For a discussion of the phenomenon of how model behavior more generally can change over time in deployment, see Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer & Neil D. Lawrence, DATASET SHIFT IN MACHINE LEARNING xi (MIT Press, 2008) [https://perma.cc/YDF5-W4K7]; Pang Wei Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, *in* PROCEEDINGS OF THE 38TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 4 (Proc. Mach. Learning & Rsch., 2021) [https://perma.cc/L9XQ-4RQ7].

252. *See, e.g.*, *Monitor Data and Model Quality*, AMAZON: AWS, https://docs.aws.amazon.com/sagemaker/latest/dg/model-monitor.html [https://perma.cc/SGZ3-Y8SP] (last visited Sept. 2, 2024).

As a general matter, interventions that involve a greater disruption to the model development pipeline or take fewer of the inputs to the model development process as a given are likely to be more costly. But under certain circumstances even these can be more economical than one might expect. For example, while additional data collection is often viewed as a costly undertaking, there are many situations where the cost of obtaining access to additional data is trivial. Notably, firms often train their models using a small sample of a much larger dataset that they have already collected because doing so reduces the cost of the computation. Under these circumstances, firms can purposefully sample a slightly larger fraction of the dataset to compensate for issues with the original sample thought to contribute to disparate impact, only slightly increasing the computational cost.

Similarly, while it can be quite costly to collect the data necessary to test a range of possible target variables, it is often possible to engineer different target variables by combining or transforming existing data in different ways, thereby avoiding the cost of additional data collection. For example, as described above, the IRS was able to reduce disparate impact by changing the target variable from a binary outcome (Is there or is there not tax fraud?) to a continuous outcome (How much money will be obtained from an audit?) without having to collect any new data.[253]

Finally, high-cost interventions, such as collecting additional training data or collecting additional features, are also those most likely to improve the overall accuracy of the model and may thus pay for themselves. If disparate impact is reduced by improving overall accuracy instead of redistributing the same amount of error, these interventions could be not just costless, but on net profitable.

In general, what constitutes a reasonable search depends upon the state of current research and available tools, the steps taken by other companies in an industry, the resources available to the company, and the extent of discrimination in a baseline model, among other case-by-case factors. That said, considering disparate impact at different decision points in the pipeline is a minor change from current modeling practices, and companies should explore the tree of potential models as expansively as possible under their circumstances.

VI. LIMITATIONS AND POTENTIAL OBJECTIONS

We have argued that the law should take account of model multiplicity by placing a duty to search for LDAs on entities that use algorithms in domains covered by civil rights laws. We have also explored how such a duty might be implemented practically. In this Part, we address some of the limitations of our proposal and consider potential objections. We first explain how context-specific needs might complicate the search for LDAs. Next, we explore the limitations of relying on accuracy as the relevant metric of performance. And finally, we address potential concerns about the legality of requiring a search for LDAs.

253. Black et al., *supra* note 234, at 1486.

A. CONTEXT-SPECIFIC CONSIDERATIONS

In this Article, we have talked generally about model performance, taking performance to mean the fraction of the model's predictions that it gets right. In practice, firms will often have context-specific reasons to favor a more precise definition of performance. For example, lenders may recognize that it may be more costly to incorrectly predict that an applicant is creditworthy when they will actually default than to incorrectly predict that an applicant is not creditworthy. Evaluating the performance of a model by only looking at the overall fraction of correct predictions would not capture this difference in the costs of different errors because each type of error counts against accuracy in the same way: a model can be 90% accurate overall whether the 10% of errors are false positives (incorrectly predicting that an applicant will repay) or false negatives (incorrectly predicting that an applicant will default). A lender might therefore want to stipulate that models only exhibit equal performance if they get the same fractions of predictions correct *and* if they also have the same false positive rates.[254]

Similarly, lending decisions are commonly based on estimated probabilities of default (e.g., person X has a 25% chance of defaulting, person Y has a 50% chance of defaulting, etc.) rather than a binary prediction of default or repayment (e.g., person X will repay, person Y will default, etc.). Lenders are often willing to extend loans to applicants with non-zero probabilities of default so long as they offset the risk of default with appropriate interest rates and maintain a manageable level of overall risk. As a result, an equally accurate model for lenders might not be one that maintains the same false positive rate; instead, it might be one that maintains a similar level of overall credit risk.[255]

Even under this stricter definition of model performance, model multiplicity continues to apply. There will be many models that can achieve equivalent performance even if equivalent performance is defined to include equal false positive rates or equal overall credit risk. For example, some models will spread true positives (correctly predicting that an applicant will repay) more equitably across different parts of the population than others, thereby generating less disparate impact without affecting the overall accuracy rate or the false positive rate. Likewise, researchers have shown that it is possible to develop many different models with comparable accuracy that assign different risk estimates to different people—implying that it should be possible to find models that maintain the same overall level of risk while increasing the selection rate for disadvantaged groups.[256] That said, as discussed in Part V, adding this (or any other) additional

254. Loan recipients might also prefer this more precise measure of performance since they might be harmed by receiving loans that they are ultimately unable to repay.

255. *See* Richard Pace, *Fool's Gold? Assessing the Case for Algorithmic Debiasing*, PACE ANALYTICS CONSULTING (Sept. 13, 2024), https://www.paceanalyticsllc.com/post/fools-gold-algorithmic-debiasing [https://perma.cc/8ZFQ-B5XZ].

256. *See generally* Jamelle Watson-Daniels, David C. Parkes & Berk Ustun, Predictive Multiplicity in Probabilistic Classification, *in* 37TH PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE 10306 (2023), https://arxiv.org/abs/2206.01131.

requirement to the definition of performance will reduce the total number of models of equivalent performance that are likely to be found.

It is equally possible to adopt different or more precise notions of fairness or to focus on fairness with respect to multiple, possibly intersectional groups,[257] not just between two groups. As with performance, these additional constraints will limit how many LDAs can be found in practice, but they will not completely foreclose the possibility of making such discoveries.[258]

There is a risk that a company might take context-specific considerations to an extreme, asserting that what matters is not any particular property of the model, but the models' effect on a business metric like net revenue. And because model properties may have a complex relationship with ultimate business impact, accepting this argument could effectively give firms free rein to reject models that are less discriminatory but equally effective in terms of model accuracy because of some loosely specified business justification. In light of this risk, courts and regulators should not defer to a business's explanation of its performance requirements, including any downstream business impact. Instead, the company should be required to justify its definition of model performance as part of its burden of showing business necessity so that viable LDAs are not arbitrarily ruled out.[259] This requirement would also address the related risk that companies subject to a duty to search might *artificially* impose additional requirements for evaluating model performance to limit the possible set of LDAs.

It is also important to remember that identifying models that perform equally well requires determining the degree to which models can differ from one another across the relevant measure of performance and still be considered equivalent. As the reader will recall, it's necessary to decide on a threshold level of difference beyond which models are considered meaningfully different. The Upstart Monitorship contemplates the importance of this decision in its final report.[260] When this specific threshold or bound ε is so narrowly defined, and where entities take an extremely strict approach, it is likely that an "entity would rarely if ever adopt less discriminatory models."[261] In such a scenario, where companies require functionally equivalent performance, then "elaborate model testing protocol risks simply becoming window-dressing" as the testing protocol is designed

257. For a discussion of intersectionality, see Kimberlé Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, 1989 U. CHI. LEGAL F. 139 (1989).

258. *See* Michael Kearns, Seth Neel, Aaron Roth & Zhiwei Steven Wu, Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness, *in* PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING 2564, 2569 (Ass'n for Computing Mach. eds., 2018) (proposing a method for finding models that are fair with respect to a large number of subgroups defined by the available set of features).

259. *See* CFPB, *supra* note 184, at 8 (directing lenders to "document the specific business needs the models serve, as well as document specific standards for assessing whether a model serves each stated business need").

260. *See* RELMAN COLFAX PLLC, *supra* note 183, at 4.

261. *See id*. at 16.

"such that less discriminatory alternatives are rarely if ever adopted."[262] To overcome this issue, a firm should also be expected to justify its choice of ε alongside its definition of model performance.

Finally, firms need to be careful when selecting an LDA that they pick a model that is robust to differences between the development and deployment contexts. Simply selecting the model with the lowest disparity in selection rates among all models of equivalent accuracy runs the serious risk of "over-fitting" to the data: selecting a model that happens to exhibit a specific selection rate disparity on the exact data in the training set, but not necessarily on the slightly different data that might be encountered in deployment.[263] To estimate the selection rate disparity that a model is likely to exhibit in deployment, developers can check to see how well the model performs on a sample of data that has been purposefully withheld from the model development process. The selection rate disparity exhibited by the model on previously unseen data will be a more reliable indicator of its likely performance in deployment—and developers should choose the model whose selection rate disparity generalizes well to unseen data.

B. QUESTIONS OF ACCURACY

In arguing for a duty to search for LDAs, we defined an LDA narrowly in order to explore the implications of model multiplicity. However, "equal accuracy" is not always an appropriate or desirable measure of a less discriminatory alternative.

First, as explained in Part II, equal accuracy is not necessarily required by law. Some legal authorities have articulated the standard more loosely—for example, looking for "comparable" alternatives. Looking for equally accurate algorithms should not preclude investigation of other LDAs. When disparate impact is severe and can be substantially reduced at a modest cost in accuracy, regulators should consider such a model to be a viable less discriminatory alternative.

Second, the fact that an entity made a reasonable search for LDAs should not excuse the use of fundamentally flawed algorithms that should not be used at all. Sometimes, the best achievable accuracy of a model may be so low and the discriminatory effects so pronounced that it should be irrelevant whether a further search would make the model marginally fairer. In certain high-stakes domains, like those regulated by discrimination laws, there should be performance levels below which models should not be considered for use at all. The correct intervention in such cases would be to seek to bring the model's overall performance up to a reasonable level or, if the model's best achievable performance continues to fall below some floor even after such an intervention, to develop an entirely different way of making these decisions.

---

262. *See id.*

263. Frances Ding, Moritz Hardt, John Miller & Ludwig Schmidt, Retiring Adult: New Datasets for Fair Machine Learning, *in* NIPS '21: PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON INFORMATION PROCESSING SYSTEMS 6478, 6490 (Curran Assocs. ed., 2021) https://dl.acm.org/doi/10.5555/3540261.3540757.

A third caveat is that accuracy is not necessarily a stable and reliable metric. There are many well-known weaknesses in the way models are commonly evaluated. To start, model evaluations are premised on the idea that the data used to evaluate the model are representative of both the population and the circumstances to which the model will be applied.[264] For example, to have any faith in an evaluation of a model developed to predict job performance, one must believe that the specific population of employees whose data comprise the evaluation dataset are a good reflection of the population that will apply for jobs in the future. One must also believe that the nature of the job to be performed by job applicants will be identical to the job performed by the employees in the evaluation dataset. But both things can easily change over time. The specific people that end up applying for a job might be quite different from those people who have occupied the role in the past. In particular, the part of the population in the evaluation dataset for which the model performs poorly may become a larger fraction of the overall population that applies for the job, thereby increasing the overall error rate. And the nature of that job might evolve, either due to changes in the workplace itself or due to changes in the overall business environment, thereby changing which features best predict future job performance and causing degradations in model performance. Either change would undermine the reliability of the reported accuracy of the model at the time of evaluation.

The reliability of the evaluation also depends entirely on the reliability of the data used to perform it. Models are generally evaluated by testing to see how they perform on data that have been "held out" of the training process, on the belief that performance on data that the model has not yet seen is a good indication of how it will perform on future data. Unfortunately, because the "held out" data come from the same pool of data used to train the models, the evaluation dataset is likely to share all the problems with the training dataset, including potentially inaccurate labeling of the outcome of interest—and the evaluation data will not help to surface those problems. Incorrect labels in the evaluation dataset are a particularly pernicious problem because these labels are treated as ground truth, the "true" outcome against which models' predictions will be evaluated. As a result, if a model correctly predicts an incorrectly labeled outcome, this will be counted as an accurate prediction, not an error. The reported accuracy of a model thus does not represent how well a model has predicted the actual outcome, but how well it has predicted the labeled outcome, even if that outcome has been labeled incorrectly.

If the underlying data are structured by discriminatory practices and policies, performing "well" on the data may be meaningless. Consider tenant-screening systems. These systems primarily rely on three data sources: criminal-history

264. *See* SOLON BAROCAS, MORITZ HARDT, & ARVIND NARAYANAN, FAIRNESS AND MACHINE LEARNING: LIMITATIONS AND OPPORTUNITIES 42 (MIT, 2023).

records, eviction records, and credit scores.[265] As has been documented, these three kinds of records are fundamentally tainted by race, gender, and disability discrimination in the criminal, housing, and credit systems.[266] The appropriate response to algorithmic systems like tenant-screening systems—which are dependent on these data sources—is not to tinker with the process by which a model is learned using these records to potentially discover an LDA. Instead, more fundamental interventions are necessary. One would need to find completely different features or develop an entirely new problem formulation. When the target of prediction and the data used to make a prediction are so deeply intertwined with discriminatory practices, firms using such models should not be allowed to defend an algorithm with adverse impact by asserting that they performed a reasonable search for an LDA. In these cases, they should forego the use of such flawed algorithms altogether.[267]

Our proposal to require a search for LDAs does not displace the need to attend to these risks. To justify a model with disparate effects, the entity relying on such a model would still need to justify its use by meeting other standards relevant to judging its validity. For example, as part of its burden of justification, a defendant should still be required to demonstrate the appropriateness of the target, the representativeness and accuracy of the training data, the reliability and validity of the chosen measure, the statistical robustness of the model, and so on.[268] In other words, searching for LDAs to reduce unnecessary disparate impacts would be just one part of ensuring that algorithmic systems avoid unjustified discriminatory effects.

C. LEGAL CONCERNS

One potential objection to recognizing a duty to search for LDAs is that it may run afoul of anti-discrimination law if it constitutes unlawful disparate treatment under statutory law or prohibited "race-consciousness" under the Constitution.[269]

---

265. *See* TINUOLA DADA, NATASHA DUARTE & UPTURN, HOW TO SEAL EVICTION RECORDS: GUIDANCE FOR LEGISLATIVE DRAFTING 9 (2022). Lydia X. Z. Brown, *Tenant Screening Algorithms Enable Racial and Disability Discrimination at Scale, and Contribute to Broader Patterns of Injustice*, CTR. FOR DEMOCRACY & TECH. (July 7, 2021), https://cdt.org/insights/tenant-screening-algorithms-enable-racial-and-disability-discrimination-at-scale-and-contribute-to-broader-patterns-of-injustice/ [https://perma.cc/ZX6L-LBY5]; Peter Hepburn, Renee Louis & Matthew Desmond, *Racial and Gender Disparities Among Evicted Americans*, EVICTION LAB (Dec. 16, 2020), https://evictionlab.org/demographics-of-eviction/ [https://perma.cc/8YB8-7N8M].

266. *See* DADA ET AL., *supra* note 265, at 18.

267. Rashida Richardson, Jason M. Schultz, and Kate Crawford offer a similar example in their study of policing algorithms built using police data at a time when those entities were under investigations or court supervision for discrimination. *See* Rashida Richardson, Jason M. Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. ONLINE 15, 20 (2019).

268. *See, e.g.*, Kim, *supra* note 3, at 921.

269. Note that only consideration of race or ethnicity even arguably triggers strict scrutiny under the Equal Protection Clause. Sex-based classifications are subject to intermediate scrutiny, and the use of some demographic characteristics like age or disability is not subject to any heightened scrutiny at all. *See* Marcy Strauss, *Reevaluating Suspect Classifications*, 35 SEATTLE U. L. REV. 135, 146 (2011).

The latter concern has been heightened by the Supreme Court's recent decision in *Students for Fair Admissions, Inc. (SFFA) v. President & Fellows of Harvard College*, which struck down certain race-conscious college admissions policies.[270] This objection misapprehends both our proposal and anti-discrimination law, which permits efforts to reduce discriminatory effects even after the *SFFA* decision.

For most private entities, statutory law, not the Constitution, is the principal source of regulation. Civil rights laws like Title VII, the FHA, and the ECOA prohibit not only disparate treatment (commonly described as intentional discrimination) but also disparate impact (facially neutral practices that have unfairly disparate effects on disadvantaged groups).[271] Disparate impact doctrine aims at "the removal of artificial, arbitrary, and unnecessary barriers" that have unjustified discriminatory effects.[272] Imposing a duty to search for LDAs comports with this purpose because an arbitrary and unnecessary barrier would be created if a company chose a model with significant racial disparities when an alternative model would perform as well and have less disparate effect.

Some might worry, however, that because searching for LDAs entails paying attention to characteristics forbidden under the civil rights laws, undertaking such a search might itself trigger liability for disparate treatment. There is nothing per se unlawful about examining the discriminatory effects of a selection system and seeking to reduce the bias in that system.[273] The Supreme Court has repeatedly stated that one of Congress's purposes in passing civil rights laws was to spur "self-examin[ation]" and "self-evaluat[ion]," with the goal of eliminating arbitrary discriminatory practices.[274] Recognizing that voluntary compliance is key, courts have approved proactive efforts to remove sources of bias, for example, when employers expand their recruiting efforts to attract a more diverse pool of candidates or stop relying on unnecessary tests that have discriminatory effects.[275] Of course, much depends upon the specific ways entities go about trying to reduce disparate impact. Relying on explicit racial quotas will run afoul of anti-discrimination law, and some techniques may fall into a gray area of legal uncertainty because of their novelty. However, many of the techniques currently available to search for LDAs are clearly permissible under existing law.

Concerns about taking affirmative steps to reduce racial impacts may stem from a misreading of the Supreme Court's decision in *Ricci v. Stefano*.[276] In that case, the City of New Haven discarded a promotional examination for firefighters because it would have produced a nearly all-white promotional class, and the

270. *See* Students for Fair Admissions, Inc. (SFFA) v. President & Fellows of Harvard Coll., 600 U.S. 181, 217–18 (2023).

271. *See* Mahoney, *supra* note 120, at 422 & n.38.

272. *See* Griggs v. Duke Power Co., 401 U.S. 424, 431 (1971).

273. *See generally* Kim, *supra* note 61.

274. United Steelworkers of Am. v. Weber, 443 U.S. 193, 204 (1979) (quoting Albemarle Paper Co. v. Moody, 422 U.S. 405, 418 (1975)).

275. *See* Kim, *supra* note 61, at 1561–66, 1563 n.10 and cases cited therein.

276. 557 U.S. 557 (2009).

City feared a disparate impact suit by minority firefighters.[277] According to the Court, discarding the results constituted disparate treatment against the successful test takers because of "the high, and justified, expectations of the candidates who had participated in the testing process," some of whom invested considerable time and expense to do so.[278] While *Ricci* found that discrimination law protected the interests of the individual firefighters who had taken the exam, it does not prohibit entities from taking steps *prospectively* to reduce the disparate impact of their practices. When entities seek to design fair selection procedures going forward, no settled expectations are disrupted, and no harm is done to individuals based on their race. The Court in *Ricci* recognized as much, noting with apparent approval several race-conscious strategies taken prospectively to reduce disparate impact.[279] Thus, while the Court found the City's actions under the specific circumstances in *Ricci* to be unlawful,[280] nothing in the decision suggests that choosing a less discriminatory alternative among equally effective models prior to implementation is disparate treatment.[281]

Another source of concern may be the Supreme Court's recent decision in *SFFA*. There, the Court disapproved of the undergraduate admissions policies of Harvard and the University of North Carolina, which, according to the Court, used race as "a determinative tip" for some applicants, such that admissions decisions turned on race.[282] The Court's majority criticized the universities' policies as involving racial stereotypes, demeaning applicants by judging them based on their ancestry rather than as individuals.[283] Further, the majority implicitly assumed that the universities' policies were overriding measures of "merit" that should and would have otherwise governed the decision.[284]

Because *SFFA* was decided under the Constitution, it is not directly relevant to private entities covered by civil rights laws, which are the focus of this Article. Even for public entities, however, the opinion has limited application because the

277. *Id.* at 593.

278. *Id.*

279. For example, the Court appeared to view favorably strategies such as oversampling minority firefighters when designing the written test and ensuring that each of the panels assessing candidates on the oral part of the exam contained minority members. *See id.* at 585. It noted that an employer is permitted to "consider[], before administering a test or practice, how to design that test or practice in order to provide a fair opportunity for all individuals, regardless of their race." *Id.*

280. *Id.* at 593.

281. Some commentators speculated after *Ricci* that the Supreme Court might eliminate the disparate impact theory altogether. *See, e.g.*, Richard Primus, *The Future of Disparate Impact*, 108 MICH. L. REV. 1341, 1362 (2010); Lawrence Rosenthal, *Saving Disparate Impact*, 34 CARDOZO L. REV. 2157, 2162–63 (2013). However, the doctrine is well-established. It has been codified in the statutory text of Title VII, and, post-*Ricci*, the Supreme Court confirmed its availability under the Fair Housing Act. While some advocates on the far right are claiming that disparate impact doctrine is unconstitutional, such a ruling by the Supreme Court would not only defy decades of precedent, it would also destabilize vast swaths of civil rights laws. Given the speculative nature of these developments, our analysis stays within the confines of well-established doctrine.

282. *See* Students for Fair Admissions, Inc. (SFFA) v. President & Fellows of Harvard Coll., 600 U.S. 181, 195 (2023).

283. *See id.* at 220.

284. *See id.*

search for LDAs is an entirely different process that does not raise the same concerns as college admissions. By examining racial impacts when choosing among models, a developer is not making any decisions about individual applicants that turn on their race, nor does the process involve stereotyping individuals based on their ancestry. Rather, the search for LDAs will typically involve comparing multiple models that differ in terms of their racial impacts (as well as their impact on other protected classes), but do not rely on protected class status as an input feature to make decisions about specific individuals. And, as explained earlier, the existence of multiple equally performing models means that selecting one with less discriminatory effect does not entail displacing a superior model. Because there is no single objectively "correct" model, no one can claim that they have a legitimate expectation that a particular model will be used. Thus, searching for LDAs entails an entirely different factual scenario from the college-admissions process that the Court struck down.[285]

Although the Court disapproved of Harvard's and UNC's admissions policies, it did not say that all forms of race-consciousness in government decisionmaking are forbidden under the Equal Protection Clause. Justice Roberts's majority opinion specifically allowed for consideration of race on an individual basis.[286] This is consistent with decades of prior court opinions, which have distinguished between race-consciousness, which is permissible in some circumstances, and racial classifications, which trigger strict scrutiny.[287] As many scholars have noted, the government often acts in race-aware ways outside of the affirmative action cases without triggering strict scrutiny.[288] And even the conservative

285. Practically speaking, the method by which models are evaluated and judged against one another involves historic examples—for example, previous loan applicants—not individuals who are currently subject to a decision process, which is a markedly different factual pattern than the process at issue in *SFFA*. For example, one practice that seemed to especially trouble the majority in *SFFA* involved steps by the universities to understand the racial makeup of the tentatively admitted class, at which point race may be a factor that causes some students to move from a provisional acceptance to a rejection. *Id.* at 221 n.7. Putting aside the merits of this approach to college admissions, it does not resemble how the models we are concerned with are tested or evaluated. A creditor, for example, may have an ongoing monitoring scheme to assess how its underwriting model performs. If the creditor determines that a certain disparity metric has passed an unacceptable threshold, they might engage in a more robust search for an LDA. Assuming they do discover a viable LDA, and that model is placed into production, that model would affect *future* applicants, but not by making decisions about them based on protected class characteristics.

286. The majority opinion stated that nothing "prohibit[s] universities from considering an applicant's discussion of how race affected his or her life, be it through discrimination, inspiration, or otherwise," so long as any benefit is "tied to *that student's* unique ability to contribute to the university." *Id.* at 231. The majority also left open the possibility that race-based admissions programs are permissible at our nation's military academies, specifically declining to address the issue "in light of the potentially distinct interests that military academies may present." *Id.* at 213 n.4.

287. Kim, *supra* note 61, at 1566–73.

288. *See, e.g.*, Kim, *supra* note 61, at 1566–68; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 856–58 (2020); R. Richard Banks, *Race-Based Suspect Selection and Colorblind Equal Protection Doctrine and Discourse*, 48 UCLA L. REV. 1075, 1108 (2001); R. Richard Banks, *The Color of Desire: Fulfilling Adoptive Parents' Racial Preferences Through Discriminatory State Action*, 107 YALE L.J. 875, 904 (1998); Samuel R. Bagenstos, *Disparate Impact and the Role of Classification*

Justices have acknowledged that the government may appropriately take goals such as increasing racial integration into account when making policy decisions like selecting a site for a new school.[289] Similarly, Justice Scalia wrote that states may seek to undo the effects of past discrimination without relying on racial classifications, for example, by adopting programs and policies that would make it easier for those excluded for racial reasons in the past to compete.[290] The *SFFA* decision did not change this existing law.[291] Searching for LDAs during the model-building pipeline does involve an awareness of race, but it is akin to decisions like where to locate a school or what procedures to adopt to lower barriers for small or new businesses to compete.[292] That kind of race consciousness alone should not trigger strict scrutiny where the resulting models do not make decisions about specific individuals that turn on their race.

To the extent that there are legitimate legal concerns about a duty to search under current law, they are more limited. For example, one might ask whether such a duty will result in an endless focus on racial considerations, or drive entities to engage in crude "racial balancing" to avoid liability. These concerns, however, misapprehend our proposal, which does not require entities to use models that are perfectly racially balanced. Nor does it require them to search endlessly to identify the *least* discriminatory model. The search required is only a reasonable one. It is not an open-ended obligation, but a component of establishing the business necessity of relying on a model with disparate effects. In that sense, our proposal is relatively modest: it simply requires entities that rely on decisionmaking algorithms to make reasonable efforts to identify LDAs as part of the model-development process.

CONCLUSION

Model multiplicity has profound ramifications for the legal response to discriminatory algorithms. The fact that multiple equally performing models exist that differ in their predictions means that there is not always a trade-off between a model's performance and the disparate impact it might cause. Indeed, the promise

and Motivation in Equal Protection Law After Inclusive Communities, 101 CORNELL L. REV. 1115, 1119 (2016).

289. *See, e.g.*, Parents Involved in Cmty. Schs. v. Seattle Sch. Dist. No. 1, 551 U.S. 701, 789 (2007) (Kennedy, J., concurring in part and concurring in judgment); Tex. Dep't Hous. & Cmty. Affs. v. Inclusive Cmtys. Project Inc., 576 U.S. 519, 545 (2015).

290. *See* City of Richmond v. J.A. Croson Co., 488 U.S. 469, 526 (1989) (Scalia, J., concurring in judgment).

291. As Sonja Starr has argued, the *SFFA* Court did not endorse "ends-colorblindness"—the "radical proposition[] that it is unconstitutional for government actors to notice a racial disparity and try to reduce it, even with race-neutral tools." Sonja Starr, *The Magnet School Wars and the Future of Colorblindness*, 76 STAN. L. REV. 161, 165, 169 (2024). Rather, the decision addressed only "means-colorblindness"—the requirement that goals such as racial diversity be pursued through race neutral means. *Id.* at 164. It left untouched the "distinction between means- and ends-colorblindness" which is "well supported by doctrine." *Id.* at 169.

292. *Cf. Inclusive Cmtys.*, 576 U.S. at 544–45; *Parents Involved*, 551 U.S. at 789 (Kennedy, J., concurring in part and concurring in judgement); *J.A. Croson*, 488 U.S. at 526 (Scalia, J., concurring in judgement).

of model multiplicity is that an equally accurate, but less discriminatory, alternative algorithm almost always exists. But without dedicated exploration, it is unlikely developers will discover potential LDAs. Thus, we have argued that to advance the purposes of the civil rights laws, entities that develop and deploy predictive models in covered domains have a duty to make a reasonable search for LDAs.

For decades, less discriminatory alternatives have been a legal backwater. Plaintiffs were poorly positioned to determine if they existed. Defendants have had little incentive to look for or implement them—and they've resisted efforts that would require them to prove they don't exist. Model multiplicity turns the situation on its head by suggesting that in nearly all cases there are less discriminatory alternatives to algorithms that have a disparate impact. And a number of relatively low-cost interventions in the development process offer promising avenues for exploration, and new techniques are constantly emerging which may make it easier and less costly to uncover models with comparable accuracy and less disparate impact. With reasonable efforts to search for LDAs, businesses that rely on algorithmic decision systems can avoid unnecessary disparate impacts. Recognizing a duty to make such efforts is critical to fulfilling the promise of the civil rights laws.