Emily Black[a1d1] John Logan Koepke, Pauline T. Kim, Solon Barocas & Mingwei Hsu, 113 Geo. L.J. 53 (2024)

LESS DISCRIMINATORY ALGORITHMS

***55**

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 ***56** 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 ***57** 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 ***58** 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 ***59** 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 ***60** 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 ***61** alternative is. […] 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.

[…]

***62 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.

[…]

***64**

[…]

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 ***65** 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 ***66** 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 ***67** 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.

***68** 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 ***69** 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]

***70** 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 ***71** 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.

[…]

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 ***86** 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.

[…]

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 ***87** 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 ***88** 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

[…]

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 ***89** 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.

[…]

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 ***90** 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.

[…]

***91** […] [A]lthough 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 ***92** 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 ***93** 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 ***94** 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]

[…]

[…]

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 ***100** 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 ***101** 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

[…]

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.

[…]

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 ***105** 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 ***106** 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]

***107** *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 ***108** 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]

***109** 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]

[…]

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.

*111 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 ***112** 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 ***113** "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.

[…]

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] ***116** 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 ***117** 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 ***118** 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 ***119** 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 ***120** 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.

Footnotes Omitted