Fundamental Limits in the Search for Less Discriminatory Algorithms—and How to Avoid Them

Benjamin Laufer
Cornell University
New York, NY
bdl56@cornell.edu

Manish Raghavan
Massachusetts Institute of Technology
Cambridge, MA
mragh@mit.edu

Solon Barocas
Microsoft Research
New York, NY
solon@microsoft.com

Abstract

Disparate impact doctrine offers an important legal apparatus for targeting unfair data-driven algorithmic decisions. A recent body of work has focused on conceptualizing and operationalizing one particular construct from this doctrine — the *less discriminatory alternative*, an alternative policy that reduces disparities while meeting the same business needs of a status quo or baseline policy. This paper puts forward four fundamental results, which each represent limits to searching for and using less discriminatory algorithms (LDAs). (1) Mathematically, a classifier can only exhibit certain combinations of accuracy and selection rate disparity between groups, given the size of each group and the base rate of the property or outcome of interest in each group. (2) Computationally, a search for a lower-disparity classifier at some baseline level of utility is NP-hard. (3) Statistically, although LDAs are almost always identifiable in retrospect on fixed populations, making conclusions about how alternative classifiers perform on an unobserved distribution is more difficult. (4) From a modeling and consumer welfare perspective, defining an LDA only in terms of business needs can lead to LDAs that leave consumers strictly worse off, including members of the disadvantaged group. These findings, which may seem on their face to give firms strong defenses against discrimination claims, only tell part of the story. For each of our negative results limiting what is attainable in this setting, we offer positive results demonstrating that there exist effective and low-cost strategies that are remarkably effective at identifying viable lower-disparity policies.

1 Introduction

Most scholarship on algorithmic discrimination over the past decade has focused on disparate impact doctrine, which allows plaintiffs to challenge facially neutral policies that nevertheless have an unjustified or avoidable disparate impact on legally protected groups. Disparate impact cases are initiated by a plaintiff who puts forward evidence that the decision-making process at issue has resulted in a disparate impact along the lines of a legally protected characteristic (e.g., race, gender, age, etc.). For example, a loan applicant must first demonstrate that there is a disparity in the approval rate for applicants according to their gender. The defendant can then put forward what is known as a "business necessity defense," where they claim that the observed disparities are required to meet some legitimate business goal — for example, that a lender's credit scoring model predicts default to some reasonable degree of accuracy. Finally, the plaintiff can rebut the firm's justification by pointing to what is commonly known as a less discriminatory alternative: an alternative decision-making process that would serve the firm's goals equally well, but with less disparate impact. If a rejected credit applicant can furnish a credit scoring model of equivalent accuracy that has a smaller gap in selection rates across gender groups, then the firm can be found liable for discrimination.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The final step in disparate impact doctrine and the notion of a less discriminatory algorithm (LDA) has attracted significant scholarly attention in recent years, as research suggests that there may be readily available less discriminatory alternatives to existing algorithms used in many contexts. One idea in particular that is attracting growing interest in both the technical and legal communities is the phenomenon of model multiplicity: it is often possible to develop many different models that all exhibit the same accuracy but make quite different predictions on individual points (i.e., a specific person) and groups of points (i.e., specific groups of people) [36, 5]. Two models might be equally accurate, but make opposite predictions for a specific person. Likewise, the first model might select members from one group more than members of another group, while the second model might select members from both groups at a similar rate, even though both models are equally accurate. Multiplicity therefore suggests that there is not always an inevitable trade-off between accuracy and fairness. Instead, practitioners will often be able to develop a large set of equally accurate models and then choose among these the one that happens to have less disparate impact across groups.

Model multiplicity speaks directly to the question of less discriminatory alternatives because it tells us that there will often be many equally accurate ways to select individuals based on some outcome of interest (e.g., loan default), some of which will have more of a disparate impact than others. In legal proceedings, less discriminatory alternatives are generally understood to serve as evidence of an avoidable disparate impact — and thus of illegal discrimination — and the burden for identifying them is today commonly thought to rest with the plaintiff [6]. Identifying a less discriminatory alternative is mainly a means for a plaintiff to hold a firm accountable for its past discrimination — that is, for the avoidable adverse impact already experienced by the plaintiff. But there is also growing belief that the phenomenon of multiplicity justifies placing the burden of searching for less discriminatory alternatives on firms themselves.[1] Given the fact of multiplicity, there is no reason to simply accept an algorithm that has a disparate impact without first asking whether a firm has attempted to find a less discriminatory alternative [5]. Unless a firm has specifically taken disparate impact into account while developing its algorithms, it is extremely unlikely to land on the least discriminatory means of achieving its goals [6]. Thus, a failure to even attempt to identify alternatives can be treated as a decision to accept an avoidable — and thus unjustified — disparate impact. Multiplicity thus puts pressure on firms to proactively test for disparate impact and, if found, to explore whether they can identify a less discriminatory means of achieving their goal equally well. This move positions LDAs as a way to prevent avoidable disparate impacts from occurring in the first place, not simply a way to establish that they have occurred in the past.

Legal scholars and computer scientists have worked together to capitalize on the promise of multiplicity for dealing with algorithmic discrimination, exploring the full implications of multiplicity for disparate impact doctrine [6] and proposing specific techniques for finding the least discriminatory alternative [22]. Civil society organizations have likewise sought to leverage this insight, calling on firms to take affirmative steps to find less discriminatory alternatives when developing their decision-making algorithms and asking regulatory agencies to clarify their expectations of regulated firms [38, 12, 37]. And a range of federal and state regulators have now issued statements instructing firms to search for less discriminatory alternatives while developing algorithms [9, 40, 39].

Despite this apparent enthusiasm, there remains considerable uncertainty around how much can be achieved by exploiting multiplicity and how easily this can be achieved [41]. Even the scholarship championing multiplicity recognized that it is generally not possible to find the *least* discriminatory possible algorithm [48] and that the process of finding LDAs in practice is not always self-evident and may require non-trivial effort and resources [6]. In this paper, we attempt to provide a more precise technical characterization of the limits of LDAs and we explore what these limits might mean for the law and for the practice of searching for LDAs. In so doing, we attempt to provide greater clarity about what might be reasonably expected of a firm in light of multiplicity.

This paper puts forward a slate of fundamental (negative) results, summarized below, which each represent limits to searching for and using LDAs.

1. Mathematical Limits (Section 2): We show that there are bounds on how much one can close the gap in selection rates between groups at a certain levels of accuracy, given the size of each group and the base rate of the property or outcome of interest in each group.

[1] Plaintiffs also generally lack the necessary information to be able to recognize when firms' practices are producing a disparate impact, let along the necessary information, expertise, and resources to identify a less discriminatory alternative that would actually serve firms' goals equally well.

2. Computational Limits (Section 3): We find that, given an initial classifier, determining whether an LDA exists is computationally intractable in general (i.e., NP-complete).
3. Statistical Limits (Section 4): Firms often design algorithms before accessing all relevant information about the particular population subjected to an algorithm, so higher performance on a fixed dataset may not mean there's an LDA that generalizes to new populations.
4. Modeling Limits (Section 5): We observe that LDAs' narrow focus on the welfare of firms (i.e., whether they can achieve their goals equally well) fails to account for the fact that certain LDAs might leave consumers strictly worse off, even if they narrow disparities in selection rates across groups.

Each of these results might seem to provide firms with reasons to reject calls to search for LDAs or with arguments to defend themselves in a disparate impact case. However, these claims only tell part of the story. For each of our negative results limiting what is attainable in this setting, we offer *positive* results demonstrating that there exist effective and low-cost strategies that are remarkably powerful, if not perfect. These strategies enable firms to reliably unearth less discriminatory models that generalize to new populations, even with very simple search methods. In Appendix E, we provide results from empirical tests of a particular group of methods that randomly generate alternative models in the same model class. In some instances, we observe evidence that these methods can reduce disparity out-of-sample, sometimes at no cost to utility.

1.1 Related Literature

Here we provide an overview of related literature on less discriminatory alternatives, accuracy and fairness, and multiplicity.

Less discriminatory alternatives and statistical approaches. The applicability of discrimination law to data-driven algorithmic decisions has received much recent attention [3, 25, 21]. Attempts to define and operationalize normatively or legally useful notions of fairness abound [23, 19, 24]. In recent work, Gillis et al. [22] put forward a definition formalizing the notion of a less discriminatory alternative. In their formulation, optimization occurs over a fixed dataset and both false positive rate and false negative rate are constrained. The authors use a mixed-integer program to identify LDAs in the case of linear classifiers. Black et al. [7] point out the issue that fairness improvements may not generalize, and requiring the reporting of fairness improvements on training data could lead to manipulations and faulty results which the authors term 'D-hacking.' Auerbach et al. [1], inspired by disparate impact law, develop a statistical method for comparing the fairness-improvability of an alternative classifier to a baseline.

Accuracy-fairness trade-offs and welfare. Fair machine learning (ML) is a vast area of research and finding satisfactory notions compatible with commonly held normative intuitions has proved curiously difficult [4, 14, 15, 29]. Scholars have developed impossibility results, trade-offs and welfare notions to frame the terms and achievable goals of defining algorithmic fairness [27, 13]. Liang et al. [32] put forward a model for understanding the achievable performance across two groups' error rate and overall accuracy. Pinzón et al. [43] analyze the feasible set of realized equal opportunity and accuracy and provide conditions where these are incompatible. Pleiss et al. [44], focused on trade-offs between fairness notions, find that satisfying welfare defined over different error rates is equivalent to treating some population members randomly.

Multiplicity. There is a growing interest in the idea that many classifiers can be similarly accurate but have different performance along other desirable attributes. This notion is referred to commonly as model multiplicity [5, 51, 52], the Rashomon effect [8], or under-specification [18]. Semenova et al. [49] find that the size of the set of good models can serve as a measure of model-class simplicity. Cooper et al. [16] find that variance in the design of classification algorithms can lead to individual people receiving different treatments depending on which model is chosen. In the context of disparate impact cases, Black et al. [6] argue that the onus should be placed on the firm — rather than the plaintiff — to conduct a reasonable search for a less discriminatory alternative in light of the flexibility afforded by multiplicity [17, 47].

1.2 A Working Formalism

Here we put forward a mathematical formalism for our setting, in which a firm makes a selection among a population. Our aim here is to introduce notation that will support inferences about what is attainable — and what is not attainable — in the search for a less discriminatory algorithm.

Population. Following convention [26, 1, 32], our population is described by the joint distribution $\langle \mathcal{X}, \mathcal{G}, \mathcal{Y} \rangle$, where each member (e.g., a loan applicant) is described by features $x \sim \mathcal{X}$, belongs to (categorical) group $g \sim \mathcal{G}$, and has an underlying true label $y \sim \mathcal{Y}$. In settings with binary labels, $y \in \{0, 1\}$ (interchangeably, we may refer to these class labels as negative '$-$' and positive '$+$'). A finite dataset, drawn from the distribution $\mathcal{X}$, is denoted X, with corresponding group membership vector Y and group belonging vector G. Finally, we call the size of the population $n = |X|$ and may refer to subsets of the population using subscript: $n_{g,y}$.

Probability distributions. Though the joint distribution of $\langle \mathcal{X}, \mathcal{G}, \mathcal{Y} \rangle$ captures the relevant information about the population in full generality, there are a few particular probability distributions that will be useful to define. First, we define $\sigma(x) := \mathbf{P}[y = 1 \mid x]$, the probability that an individual with data x has a positive outcome class. For simplicity, we assume that this unobserved outcome class $\mathcal{Y}$ that the firm wishes to base their selection on is independent of protected group status conditional on $\mathcal{X}$, but we note that our results do not require this assumption (i.e., they hold in settings that exhibit differential prediction). Second, we define $\rho_g(x) := \mathbf{P}[x \mid g]$, the group-specific probability distribution over the data features. If the set of features X is finite, then $\rho_g(x)$ represents the fraction of members of group g with data values x.

Classifiers. We define a classifier $h(x) : \mathcal{X} \to \{0, 1\}$ as a mapping from features to binary labels. We use $\mathcal{H}$ to denote the set of all possible classifiers. The particular classifier that a firm commits to is $h^0(x)$, and a candidate alternative classifier (which may or may not be a LDA) is $h'(x)$.

Selection rates and errors. For a given classifier h, denote the selection rate $\texttt{SR}(h) := \mathbf{P}_{x \sim \mathcal{X}}[h(x) = 1] = \mathbf{E}_{x \sim \mathcal{X}}[h(x)]$. For a particular group g, the group-specific selection rate would be $\texttt{SR}_g(h) := \mathbf{E}_{x \sim \mathcal{X}}[h(x) \mid g]$. Finally, because the firm is hoping to design their classifier $h(x)$ to mimic the value y, we define the standard notions of *false positive rate*, *true positive rate*, *false negative rate*, and *true negative rate*. We will refer to them using their 3-letter abbreviations. For example, the false positive rate is defined as $\texttt{FPR} := \mathbf{P}_{x \sim \mathcal{X}}[h(x) = 1 \cap y = 0]$, i.e. the total portion of the population that is positively classified but has $y = 0$. The group-specific false positive rate $\texttt{FPR}_g$ is accordingly defined as the FPR conditional on group membership g. We use analogous notation for TPR, FNR and TNR.

Disparity and Utility. We define the demographic disparity of a classifier h to be $\mathbf{\Delta}(h) := \texttt{SR}_1(h) - \texttt{SR}_2(h)$. We define this disparity assuming members of the population can be split into two groups, i.e. $\mathcal{G} \in \{1, 2\}$. If we assume group $g = 1$ to be 'advantaged' and group $g = 2$ to be 'disadvantaged,' then the demographic disparity would be non-negative. Sometimes, we'll be interested in absolute disparity, which we define as the absolute value of the demographic disparity. Finally, here we offer a broad notion of *utility* for a given classifier h, allowing entities (firms or consumers) to have particular weightings of preferences over true positive and false positive outcomes. We define utility as $\mathbf{U}(h; \lambda) := \texttt{TPR} - \lambda \texttt{FPR}$ where the value of $\lambda \in \mathbb{R}^+$ suggests the relative benefit of a true positive compared to the cost of increasing the probability of false positive classification. This particular formulation is discussed further in Section 5.

2 Mathematical Limits

In many cases, the firm has access to a single finite dataset with full information (X, G, Y). Any classifier $h(X)$, defined over this finite population, has some observable accuracy performance (e.g., utility, as we've defined it) and fairness performance (e.g., demographic disparity, as we've defined it). A natural starting question one might ask in this setting is, what accuracy and fairness measures are *(im)possible* to jointly achieve in the fixed sample? In other words, assuming we have access to group belonging, labels, and outcome information Y, what can we conclude from moving errors around freely, caring only about performance in-sample?

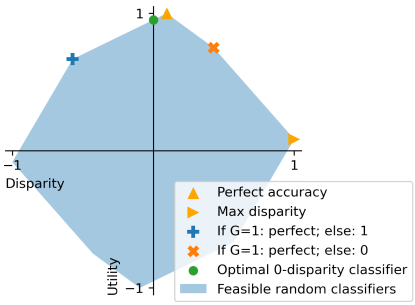
This inquiry uncovers a fundamental limit to the efficacy of any search for a less discriminatory alternative: at certain levels of accuracy, finding a 0-disparity classifier is impossible. In the remainder of this section, we will show that 0- or near-0 disparity classifiers are possible to achieve as long as the utility achieved by the starting classifier remains below a certain cutoff, which depends only on the sizes of the positive group 1 ($n_{1,+}$), negative group 1 ($n_{1,-}$), positive group 2 ($n_{2,+}$), and negative group 2 ($n_{2,-}$) populations.

Quickly we will see that achieving highly performant LDA classifiers on a fixed and labeled dataset is much easier than guaranteeing any accuracy or disparity property *in general*, and in future

(a) Example G, Y

	+	-
1	15	20
2	5	10

(b) Feasible $\mathbf{\Delta}, \mathbf{U}$ (random)



(c) Feasible $\mathbf{\Delta}, \mathbf{U}$ (deterministic)

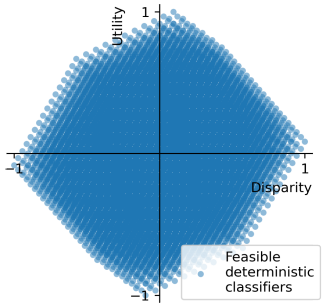


Figure 1: Consider a given, finite population broken down by group belonging and outcomes (a; left). If randomized decision rules are feasible, then a polygon depicts the convex set of feasible, in-sample utility and disparity values (b; center). If solutions are restricted to deterministic classifiers over the dataset (c; right), the polygon bounds the achievable values.

sections we will discuss questions about generalizability and inference. However, it can still be useful to know what is possible to achieve in-sample, because it serves to define the boundaries of what can be achieved more generally.

2.1 Polygon of Possibilities

Here we consider the set of all attainable accuracy and disparity values for a binary selection rule. The set of attainable accuracy and fairness values can be surmised given access to a dataset X with group membership G and outcomes Y. This should be no surprise: if we can peek at the outcomes corresponding to every data point in a dataset, then we can easily produce decisions that bound the attainable accuracy and/or disparity performance. For example, the perfectly accurate decision rule could be attained by simply using the outcome data as classification labels. The perfectly biased decision rule would only select based on protected attribute.The highest-accuracy, fair decision rule could be identified by starting with the perfect-accuracy classifier and swapping either 1) positive-labeled advantaged group members or 2) negatively-labeled disadvantaged group members until selection rates equalize (choose whichever swap optimally trades off utility for disparity reductions).

A representation of the feasible set of all decisions on the accuracy-disparity plane is provided in Figure 1. The blue shape in Figure 1(b) is the region encompassing all achievable accuracy (i.e., utility) and disparity values. Note, however, that as long as the data and decision are discrete, not all values within the polygons are achievable. With discrete data, the set of achievable decision rules is not connected. Instead, it is a dense lattice of points. Moving from one point to a neighboring point corresponds to switching the label on a single individual in the population. This lattice is depicted in Figure 1(c). For exposition, in much of the remainder our analysis, we allow randomized decisions, though the results should hold for the discrete case.

Recall that in this setting, the achievable accuracy and disparity values are computed on a single instance of a dataset with true outcome labels Y and group belonging G. This information allows us to make certain claims about what performance a classifier can achieve on a single dataset. For instance, any reasonable classifier should be in the upper right or upper left quadrant of this plane. If two of our goals are to maximize utility and minimize disparity, then an ideal classifier should be as far north as possible, without veering too far west or east. Notice that the most accurate classifier requires some non-zero level of disparity: if a classifier exhibits perfect accuracy on a dataset, then its disparity is the difference in group-specific base-rates (this point is plotted as a yellow pyramid). Among the set of perfectly fair classifiers, the most accurate is plotted as a green point. There are a number of questions we can ask about this feasible set. For example, how high does accuracy have to be such that there does not exist an alternative classifier with 0-disparity and the same (or higher) firm utility? In the remainder of this section, we make formal claims about the properties of the accuracy-disparity polygon, and discuss the reasonableness of conclusions based on it.

Theorem 1. *Assume group 1 is advantaged and randomized classifiers are feasible. There is a utility threshold* $\mathbf{U}^* = 1 - \min\left[\frac{n_1}{n_+}, \frac{\lambda n_2}{n_-}\right](BR_1 - BR_2)$ *such that there exists a zero-disparity alternative*

classifier h' with $\mathbf{U}(h') \geq \mathbf{U}(h^0)$ if and only if $\mathbf{U}(h^0) \leq \mathbf{U}^*$. Additionally, if $\mathbf{U}(h^0) > \mathbf{U}^*$, then the minimum disparity alternative classifier has $\boldsymbol{\Delta}(h') = 1 - \min\left[\frac{n_1}{n_+}, \frac{\lambda n_2}{n_-}\right]\left(\boldsymbol{\Delta}(h^*) - \boldsymbol{\Delta}(h^0)\right)$.

The proof of the above finding is provided in Appendix B. Intuition for this result can be found in Figure 1(b). The classifiers that optimally trade off between utility and disparity can be found along a (small) line segment between the perfect-accuracy classifier and the optimal, 0-disparity classifier. As long as a starting classifier has utility less than the optimal, zero-disparity classifier, then, there exists a zero-disparity LDA with equal or greater utility.

The fact that the accuracy-optimal classifier may, in many cases, have non-zero disparity could raise questions about whether a 0-disparity classifier is possible to achieve at a certain (starting) level of accuracy. In other words, a firm defending against charges of discrimination might claim that there is an inevitable trade-off between disparity and accuracy, and that therefore, favoring the advantaged group serves a business need. We find, however that the fundamental trade-off between these values on a fixed dataset does not 'kick in' unless the accuracy of the starting classifier is above a certain cutoff, and even still, there is room to maneuver to a floor on disparity. Further, as depicted in Figure 1, and as observed on empirical datasets in Figure 4, this fundamental bound is vacuous except at the most (often unreasonably) high levels of accuracy, so likely would not offer a sound defense in many realistic settings.

So far we have claimed that a selection policy with 0 disparity and the same level of accuracy exists (in sample) in many cases. But, in practice, we often want *deterministic* classifiers, and we cannot cleanly separate members of the dataset according to their group belonging and outcome label. If we stop allowing an algorithm designer to navigate the polygon in these ways, there may be no 0-disparity, deterministic classifier with limited data. Thus, in the next section, we ask when the firm can find the 'best' deterministic classifier.

3 Computational Limits

If it were easy to determine whether better alternatives exist, then the law would not have to rely on plaintiffs to tirelessly search for them. Hypothetically, an auditor could simply verify that any lending policy or hiring policy is the least discriminatory possible, given the set of alternative policies that might perform a similar function (at least as effectively) for the business. Better yet, if this were easy to do, the business itself could simply employ the procedure to find a minimally-discriminatory policy that performs at least as well as their existing policy. However, the task of arriving at less discriminatory alternatives is not so simple.

In this section, we offer findings on the computational complexity and algorithmic opportunities for certifying whether (reasonable) less discriminatory policies are possible. We show that certifying whether there exists a less discriminatory alternative algorithm is NP-hard. We demonstrate this finding even under special conditions where the firm accesses information and capabilities that, in realistic settings, further steps would be needed to estimate or execute. However, in these settings, we show that certain approximation and relaxation strategies enable a firm to arrive at an LDA as long as some *de minimis* difference in treatment is acceptable. These results suggest that defensive claims about the computational burden of identifying the least discriminatory policy are credible, but do not imply that reasonable search procedures are impossible.

A starting definition. Because we treat the problem of finding an LDA with some formality in this section, it is worth providing a technical definition for the sake of analysis. However, as we will quickly see, there are a number of ways to define this concept, and many intricacies and challenges arise depending on the definition used. Here we provide a starting formalism.

Definition 1 (Full-information LDA)**.** *Given a data distribution $\mathcal{X}$, a Bayes-optimal predictor $\sigma(x)$, a group-specific probability density function $\rho_g(x)$, a utility function $\mathbf{U}(h; \lambda)$, and a baseline classifier $h^0(x)$, a **full information LDA** is a classifier h' with utility at least $\mathbf{U}(h^0)$ and absolute disparity less than $|\boldsymbol{\Delta}(h^0)|$.*

We call this definition the 'full-information' case because we assume that we have full access to the the probability distribution $\mathcal{X}$ and the Bayes-optimal predictor $\sigma(x)$, which represents the true probability that members of our dataset with attributes x have positive outcome variable. We make this assumption for the sake of proving impossibility and hardness results: if we can show that a

problem is NP-hard or contains certain inherent impossibilities even in the case where our capabilities are unrealistically broad, then we can infer that these results hold in the (harder) scenario where we also have to predict labels. In future sections we consider cases where the firm does not have perfect access to σ. Finally, we note that Gillis et al. [22] offer an alternative formulation where the 'business need' condition is represented by upper bounds on both the FPR and FNR. Our utility notion is somewhat more flexible because it offers a *weighting* over errors, and focuses simply on the relative value of true positives compared to the cost of false positives.

Complexity result. Here we prove that the problem of certifying whether there exists an LDA, as defined in Definition 1, is NP-hard.

Theorem 2. *The full-information LDA problem $\langle \mathcal{X}, \sigma, \rho_g, h^0 \rangle$ is NP-complete.*

Appendix C.2 contains a proof of the theorem involving a reduction from the Subset Sum Problem.

Meaning and strategies. How should we interpret this result? On its face, it seems to provide firms with a natural defense: it is too computationally burdensome to find an LDA. A closer look suggests that NP-hardness may not be as much of an obstacle as it might seem. In Appendix C.3, we give a $(1+\epsilon)$-approximation algorithm for the full-information LDA problem. While it may be difficult to find the *least* discriminatory alternative, given full information, we can efficiently find a close approximation. We noted in Section 2 that mathematical constraints, while present, were unlikely to be binding in practice—the same is true here. If a plaintiff finds an LDA that a firm failed to use, it is unlikely to be because finding that LDA was intractable. Instead, it is more likely the firm lacked (or failed to collect) enough information. We turn next to the question of information and its limits.

4 Statistical Limits

The previous sections presented results in settings with unrealistic access to information. Section 2 allowed a finite sample of data to stand in for the population, and Section 3 presumed the distribution of data and the Bayes-optimal predictor were known. While these generous assumptions are valid for proving fundamental limits in the search for less discriminatory algorithms, they are not appropriate for positive results about the effectiveness of algorithms in realistic settings. The reality of ML-driven ('predictive') policies is that they are designed and deployed without perfect knowledge about the population subjected to them. Even if model A seems to perform better than model B on a training dataset, we cannot *guarantee* this performance out-of-sample on unseen data. These comparisons are particularly hard when A and B are sampled from different distributions (e.g., come from different model classes). The statistical task of understanding what inferences can be made from a limited dataset is a fundamental challenge for ML research. In what follows, we present two imperfect LDA definitions that illustrate the challenges introduced by statistical limits.

What is knowable when? To what statistical standards should a purported LDA be held? An LDA must, of course, be (nearly) as good from the firm's perspective as the model in question, and it must also exhibit lower disparities. But on what data distributions should we evaluate these claims? In general, we can only expect to evaluate performance of a model on pre-deployment data collected by the firm, since post-deployment data typically lacks ground truth outcomes for those rejected by a model [28]. On the other hand, we *can* evaluate the disparity produced by a candidate LDA on post-deployment data, since this measure does not require access to ground truth labels; indeed, this is one of the reasons to use selection rate disparities as an indicator for discrimination [45]. It might seem that this is the natural way to evaluate an LDA: it must offer (1) comparable utility as measured on pre-deployment data, and (2) lower disparity as measured on post-deployment data.

Unfortunately, this definition introduces a critical asymmetry between the plaintiff and the defendant that renders it nearly vacuous. Whereas the plaintiff's search is entirely in-sample — they observe both pre-deployment and post-deployment data when looking for an LDA[2]— the defendant's search is out-of-sample. The defendant must (by definition) commit to a model before observing post-deployment data. Even if we restrict our attention to a set of models that achieve a certain accuracy, there is no guarantee that the lowest-disparity model on pre-deployment data will continue to exhibit minimal disparity on post-deployment data. In other words, the plaintiff will often be able to identify such an LDA after the fact, even if a defendant could never have done so in advance.

[2]Recall that plaintiffs initiate a disparate impact claim by using *post-deployment data* to establish that a policy has resulted in a disparity in selection rates.

To illustrate this point, we can consider the simpler task of increasing accuracy. Given labeled data in hindsight, producing the perfect-accuracy decision rule is easy (just use the outcome labels!). Similarly, producing a fair classifier in hindsight is intuitively an easier task than designing a classifier that must be fair on unseen data.

Optimizing for the future. Recognizing the mistake of evaluating a model on post-deployment data, we might revise our LDA definition as follows. As before, an LDA should (1) maintain comparable utility when evaluated on pre-deployment data. However, instead of evaluating its post-deployment disparity, we will instead require that (2) it has lower disparity as measured on the *pre-deployment* data. We no longer allow for a temporal asymmetry between plaintiffs and defendants; all evaluations are conducted solely on information the firm has at the time of model development.

While it is in principle feasible for a firm to meet this standard, doing so would be undesirable. A firm's goal is not (and should not be) to optimize for performance on some pre-deployment dataset; they instead seek to perform well out-of-sample when the model is deployed. A model that overfits to the pre-deployment data (e.g., by only assigning positive classification to feature values with positive labels in the pre-deployment data) is unlikely to generalize well. Similarly, while a firm can achieve any utility-disparity combination shown in Figure 1 in-sample, we would not expect these metrics to generalize out-of-sample [7]. Instead, firms will in practice restrict the complexity of the models they train to trade variance for bias. Models trained this way will in general be far away from the boundaries of the feasible polygon (see Figure 4). Under the LDA definition discussed here, the plaintiff could effectively choose a model with (near-)perfect pre-deployment performance. But it would be undesirable and unreasonable to expect the firm to deploy such a model, since it would almost certainly have poor post-deployment performance.

Reasonable and unreasonable searches. If neither of these definitions captures the principle behind LDAs, what could we do instead? Our proposal builds on Black et al. [6] to put forth a standard of a *reasonable* search. While we cannot provide a comprehensive definition of "reasonble" in this context, we offer several concrete criteria. First, when sampling multiple models from the same distribution (e.g., changing the random seed for a random split of the dataset or initialization of a randomized training procedure), it is reasonable for a firm to choose the model that optimizes its stated utility-disparity trade-off as measured pre-deployment. Intuitively, while in-sample performance does not provide an unbiased estimate for out-of-sample performance, classical results from learning theory tell us that empirical risk minimization yields good models [50]. Second, a "reasonable" search cannot require the defendant to know about the *realization* of data post-deployment. A plaintiff may, however, question whether the firm's pre-deployment data was collected so as to be distributed similarly to the post-deployment data — if, for example, the defendant trained models on data from one country and deployed in another, the plaintiff might question whether the firm's model selection process was "reasonable." Finally, the defendant's choice of model class is necessarily based on heuristics. Without knowing the true data distribution, there is no "right" model class. A plaintiff might well question whether, e.g., the defendant could have reduced disparities while preserving utility by using a more complex model class; such questions must be litigated on a case-by-case basis.

5 Consumer Welfare and Modeling Limits

So far, we have predominantly focused on methods for arriving at an LDA with a single utility function, parameterized by λ. One might reasonably assume this utility function represents the interests of the firm, since disparate impact law harps on finding an alternative policy that meets the firm's business needs. However, it is well known that firm and consumers can have divergent interests [2]. When firms select whom to offer a loan, their policies aim to avoid granting credit to people who are likely to default, because the firm is unlikely to turn a profit in these cases. Granting risky loans could have deleterious consequences for the consumer, too, because they can worsen the consumer's financial outlook in the long-term [34]. In this section, we explore the possibility that consumers and firms might have different preferences over outcomes. In certain cases where the utilities of consumers and firms differ, we show that there exist alternative classifiers that jointly serve business needs and lower selection rate disparities but that leave consumers strictly worse off. In other words, there exist classifiers satisfying our prior definitions of LDA that *harm* the consumers, including those belonging to the disadvantaged group.

However, we show that just because someone can identify the existence of a consumer-harming LDA, this does not necessarily mean that consumer-benefiting LDAs do not exist. Additionally, we find that requiring that consumers should not be harmed does not impose significant computational burden or effort on the LDA search. For any number of additional utility considerations beyond 2, or in settings where the set of utility considerations is unknown, the optimization is easily reduced to the case where $|\vec{\lambda}| = 2$.

Consumer-harming LDAs exist. Here, we show that if firm and consumer welfare differ (as specified further in Appendix D.1), it is possible to identify an LDA that strictly harms consumers.

Claim 1. *There exists a full-information LDA setting $\langle \mathcal{X}, \sigma, \rho_g \rangle$ with two utility parameters $\lambda_f \neq \lambda_c$, and a pair of classifiers h^0, h' such that $\mathbf{\Delta}(h') < \mathbf{\Delta}(h^0)$ and $\mathbf{U}_f(h') \geq \mathbf{U}_f(h^0)$, but h' strictly harms consumers: $\mathbf{U}_c(h') < \mathbf{U}_c(h^0)$.*

The above finding (proven in Appendix D.2) suggests that there are certain instances where it is only possible to achieve consumer-benefiting or firm-benefiting alternatives, but not both simultaneously. An LDA that serves the firm's interest will not, necessarily, benefit consumers, and vice versa.

The existence of consumer-harming LDAs points to the fact that simplistic definitions of this concept can lead firms to arrive at classifiers that do not meaningfully improve conditions for consumers. It also suggests that an LDA might reasonably be refuted, if it fails to meet the condition that consumers (especially, those from the disadvantaged group) should be better off as the result of a proposed LDA. Thankfully, as we will see, adding stipulations and considerations for consumer welfare does not pose a significant computational challenge as it has nice convexity properties.

Utility behaves nicely. In practice, how should a firm account for consumer utility, particularly when consumers don't necessarily know their utility functions and firms don't have the ability to ask at scale? Here we show that as long as an LDA is able to satisfy two utility requirements, defined by two values λ_1, λ_2, it satisfies all utility values for lambda in the interval between them.

Claim 2. *If a firm is able to identify a region $[\lambda_1, \lambda_2]$ such that the utility consideration(s) are within this region, an LDA that improves welfare for each of the endpoints of the region also improves utility for every utility function in that region.*

A formal proof is provided in Appendix D.3. Another way of thinking about the above result is that the utility as we've defined it is convex, so it isn't too burdensome to add constraints that require additional players or utility functions are satisfied in the LDA search. Further, in a world where firms do not know, exactly, what consumer welfare is, they should still be able to search for LDAs that satisfy a range of (likely) utility formulations. Broadly, even though it is possible that naïvely-defined LDAs leave consumers worse off, it is easy to stipulate that an LDA should satisfy additional utility considerations, and this is unlikely to considerably increase the firm's computational burden.

6 Conclusion

This paper is concerned with a legal concept that predates the era of pervasive machine learning. The less discriminatory alternative has emerged as an enticing notion because it could offer a way for old rules to apply to new tools. We find, however, that simple and elegant operationalizations of this concept run up against fundamental limits, trade-offs, and impossibilities that circumscribe the attainable properties and attributes of an LDA search. Mathematically, a classifier can only exhibit certain combinations of accuracy and disparity – a near-perfect-accuracy model cannot have 0 disparity unless a dataset already exhibits equal base rates. Computationally, a search for the least-discriminatory-possible model at some baseline level of utility is generally intractable in polynomial time. Statistically, although LDAs are almost always identifiable in retrospect on fixed populations, making conclusions about how alternative classifiers perform on an unobserved distribution is much more difficult. Finally, from a modeling and consumer welfare perspective, defining an LDA only in terms of 'business needs' can lead to LDAs that leave consumers strictly worse off. Each of these results seems to offer ammunition to firms defending their (potentially unfair) practices against accusations of discrimination. However, they only tell part of the story. While each of these results do fundamentally limit the search, they do not preclude firms and plaintiffs from adopting available, low-cost, and effective methods for identifying LDAs.

References

[1] Eric Auerbach, Annie Liang, Max Tabord-Meehan, and Kyohei Okumura. Testing the fairness-improvability of algorithms. *arXiv preprint arXiv:2405.04816*, 2024.

[2] Oren Bar-Gill and Elizabeth Warren. Making credit safer. *University of Pennsylvania Law Review*, 157(1):101, 2008.

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[5] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

[6] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. Less discriminatory algorithms. *Available at SSRN*, 2023.

[7] Emily Black, Talia Gillis, and Zara Yasmine Hall. D-hacking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 602–615, 2024.

[8] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[9] Consumer Financial Protection Bureau. Fair lending report of the consumer financial protection bureau. *Consumer Financial Protection Bureau*, June 2024. URL `https://files.consumerfinance.gov/f/documents/cfpb_fair-lending-report_fy-2023.pdf`.

[10] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209, 2022.

[11] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[12] Jennifer Chien and Adam Rust. Urgent call for regulatory clarity on the need to search for and implement less discriminatory algorithms. *Consumer Federation of America*, 2024. URL `https://22.org/wp-content/uploads/2024/06/240625-CR-CFA-Statement-on-Obligation-to-Search-fo`

[13] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[14] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[15] A Feder Cooper, Ellen Abrams, and Na Na. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 46–54, 2021.

[16] A Feder Cooper, Katherine Lee, Solon Barocas, Christopher De Sa, Siddhartha Sen, and Baobao Zhang. Is my prediction arbitrary? measuring self-consistency in fair classification. *arXiv preprint arXiv:2301.11562*, 2023.

[17] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, pages 2144–2155. PMLR, 2021.

[18] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[20] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1): 5–47, 2022.

[21] Talia B Gillis and Jann L Spiess. Big data and discrimination. *The University of Chicago Law Review*, 86(2):459–488, 2019.

[22] Talia B Gillis, Vitaly Meursault, and Berk Ustun. Operationalizing the search for less discriminatory alternatives in fair lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 377–387, 2024.

[23] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[24] Michael Kim, Omer Reingold, and Guy Rothblum. Fairness through computationally-bounded awareness. *Advances in neural information processing systems*, 31, 2018.

[25] Pauline T Kim. Data-driven discrimination at work. *William & Mary Law Review*, 58(3):857, 2017.

[26] Jon Kleinberg and Sendhil Mullainathan. Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808, 2019.

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[28] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[29] Benjamin Laufer, Sameer Jain, A Feder Cooper, Jon Kleinberg, and Hoda Heidari. Four years of facct: A reflexive, mixed-methods analysis of research contributions, shortcomings, and future prospects. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 401–426, 2022.

[30] Benjamin Laufer, Thomas Gilbert, and Helen Nissenbaum. Optimization's neglected normative commitments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 50–63, 2023.

[31] Benjamin Laufer, Jon Kleinberg, Karen Levy, and Helen Nissenbaum. Strategic evaluation. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–12, 2023.

[32] Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 58–59, 2022.

[33] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31, 2018.

[34] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[35] Michael Lohaus, Matthäus Kleindessner, Krishnaram Kenthapadi, Francesco Locatello, and Chris Russell. Are two heads the same as one? identifying disparate treatment in fair neural networks. *Advances in Neural Information Processing Systems*, 35:16548–16562, 2022.

[36] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/marx20a.html.

[37] John Merrill1, Mark Jones, Mark Eberstein, Kareem Saleh, Dana Lockwood, Lusine Petrosyan, and Michael Akinwum. Improving mortgage underwriting and pricing outcomes for protected classes through distribution matching. *National Fair Housing Alliance and FairPlay*, 2024. URL https://nationalfairhousing.org/wp-content/uploads/2024/04/Unlocking-Fairness-Final_April-202

[38] NCRC. Cfpb should encourage lenders to look for less discriminatory models, ncrc, April 2022. URL https://ncrc.org/cfpb-should-encourage-lenders-to-look-for-less-discriminatory-models/.

[39] New York State Department of Financial Services. Insurance circular letter no. 7. *New York State Department of Financial Services*, July 11, 2024. URL https://www.dfs.ny.gov/industry-guidance/circular-letters/cl2024-07.

[40] U.S. Department of Housing and Urban Development. uidance on application of the fair housing act to the screening of applicants for rental housing. *U.S. Department of Housing and Urban Development*, April 29, 2024. URL https://www.hud.gov/sites/dfiles/FHEO/documents/FHEO_Guidance_on_Screening_of_Applicants_for_R

[41] Richard Pace. Fool's gold? assessing the case for algorithmic debiasing. *Pace Analytics, LLC.*, 2023.

[42] Eike Petersen, Sune Holm, Melanie Ganz, and Aasa Feragen. The path toward equal performance in medical machine learning. *Patterns*, 4(7), 2023.

[43] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the incompatibility of accuracy and equal opportunity. *Machine Learning*, 113(5):2405–2434, 2024.

[44] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

[45] Manish Raghavan and Pauline T Kim. Limitations of the" four-fifths rule" and statistical parity tests for measuring fairness. *Geo. L. Tech. Rev.*, 8:93, 2024.

[46] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10):896–904, 2021.

[47] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*, 2024.

[48] Matthew U Scherer, Allan G King, and Marko J Mrkonich. Applying old rules to new tools: Employment discrimination law in the age of algorithms. *SCL Rev.*, 71:449, 2019.

[49] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1827–1858, 2022.

[50] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[51] Jamelle Watson-Daniels, Solon Barocas, Jake M Hofman, and Alexandra Chouldechova. Multi-target multiplicity: Flexibility and fairness in target specification under resource constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 297–311, 2023.

[52] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10306–10314, 2023.

[53] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

A Further Related Work

The so-called fairness-accuracy trade-off has been discussed by a number of scholars, and there are instances where improving both is attainable algorithmically [46, 53, 11, 42]. The extent to which some form of disparate treatment is required to achieve demographic parity on datasets with base rate differences is discussed by [35, 10]. Reasoning about what notions of fairness (e.g. selection rates) and utility are attainable relate to, broadly, the opportunities and limits of phrasing the LDA concept as an optimization problem, and face constraints about what is easily measurable and accessible (see e.g., [30, 45]). Fuster et al. [20] have explored the distributional effects *between-group* of introducing machine learning to credit markets.

B Supplementary Materials on Mathematical Limits

Here we provide supplementary matrials and proofs for claims made in Section 2.

B.1 Proof of Theorem 1

Proof. The proof relies on reasoning about the Pareto-efficient region of achievable $\mathbf{U}$ and $\mathbf{\Delta}$ values. We say an alternative classifier **Pareto-dominates** a starting classifier if 1) the alternative's performance measures (e.g., utility and disparity) are greater than or equal to the starting classifier's, and 2) at least one of the alternative's performance measures is strictly greater than the starting classifier's. We define the set of **Pareto-efficient** classifiers as the set of classifiers that are not Pareto-dominated.

A classifier in our setting is specified by four quantities: the proportion of the population of type (g, y) selected, for $g \in \{1, 2\}$ and $y \in \{+, -\}$. Phrased this way, we can use the fact that we're searching over a convex decision space (a cube representing four proportion values) and any perturbation or 'swap' affects disparity and utility as a function of the g and y value of the swapped data, as well as the fixed population sizes $n_{g,y}$.

Observe that the perfect classifier, $h^*(x) = y$, is in the Pareto-efficient set because it uniquely maximizes utility. We can use this classifier as an 'anchor point' and find the alternative classifiers that optimally trade off utility and disparity. Notice there are exactly four possible ways to perturb h^*: a) Decrease the proportion of $(1, +)$ members with positive label, b) decrease the proportion of $(2, +)$ members with positive label, c) increase the proportion of $(1, -)$ members with positive label, or d) increase the proportion of $(2, -)$ members with positive label. We can immediately rule out two of these perturbations by noticing they strictly harm both accuracy and disparity: (b) and (c) introduce errors that strictly leave group 2 with fewer selections or group 1 with more selections (respectively). This leaves two possible candidates for introducing errors that optimally trade-off accuracy and disparity.

Two claims are necessary for the remainder of the proof. **First**, observe that *there are sufficient swaps of each type (a) and (d) to achieve zero-disparity using only one of these types*. This observations follows directly from our definitions of demographic disparity and 'advantaged group.' If group 1 is advantaged (given), then $\texttt{SR}_1(h^*) > \texttt{SR}_2(h^*) \rightarrow \frac{\sum_x h^*(x|g=1)}{n_1} > \frac{\sum_x h^*(x|g=2)}{n_2}$. Adding given constraints, we can say $1 \geq \frac{\sum_x h^*(x|g=1)}{n_1} > \frac{\sum_x h^*(x|g=2)}{n_2} \geq 0$. However, based on the feasible set of classifiers, we know that by exhausting all of swap (a) we'd have $\texttt{SR}_1 = 0$ without any change to $\texttt{SR}_2$. Similarly, if we exhaust all swaps of type (d) we'd have $\texttt{SR}_2 = 1$ at no change to $\texttt{SR}_1$. Because there are sufficient swaps to change the ordering of the group selection rates, we know that there are enough of each swap type to equalize selection rates across groups.

Second, observe that with full information, each swap type has a constant ratio between utility reduction and disparity reduction, given as follows:

- Swap type (a): Reducing the number of positively-labeled members of group 1 corresponds to a decrease in the true positive rate of $\frac{1}{n_+}$ meaning the unit change in $\mathbf{U}$ as a function of swap (a) is $-\frac{1}{n_+}$. The effect on disparity is $\frac{1}{n_1}$. So the ratio between utility reduction and disparity reduction is $\frac{n_1}{n_+}$.
- Swap type (d): Increasing the number of positively-labeled members of group 2 corresponds to an increase in the false positive rate of $\frac{1}{n_-}$ meaning the unit change in $\mathbf{U}$ as a function of swap (d) is $-\lambda\frac{1}{n_+}$. The effect on disparity is $\frac{1}{n_2}$. So the ratio between utility reduction and disparity reduction is $\frac{\lambda n_2}{n_-}$.

Taken together, the above two claims suggest that a constant (real, not necessarily integer) number of 'swaps' draw the Pareto-efficient region trading off between utility and disparity, and this region is a line segment defined by two points in utility-disparity space: utility-optimal classifier h^* and the utility-optimal classifier with zero disparity h^f, representing by $(\mathbf{\Delta}, \mathbf{U}) = (0, u^f)$, where $u^f = 1 - \min\left[\frac{n_1}{n_+}, \frac{\lambda n_2}{n_-}\right]\mathbf{\Delta}(h^*)$.

For any starting classifier with utility less than or equal to u^f, the classifier h^f represents a zero-disparity alternative meeting the utility requirements. Otherwise, the minimum-disparity classifier at the a given utility value corresponds to the point in the Pareto region at the same starting utility value.

□

C Supplementary Materials on Computability

Here we provide supplementary materials and proofs for claims made in Section 3.

C.1 Intuition for Hardness Result

Here, we provide some intuition for the proof of the computational complexity of the full information LDA. Consider the specific case where our dataset is made up of discrete and finite data values $\mathcal{X} = \{x_1, x_2, ..., x_{|\mathcal{X}|}\}$. At its core, the mathematical task of finding the LDA is to 'grab' (i.e., positively label) a subset of data values in $\mathcal{X}$ that meet certain utility and disparity requirements. Intuitively, we can imagine our data set as a collection of points. Every point $x_i \in \mathcal{X}$ has a disparity value d_i and a utility value u_i. We are looking to find the subset $S^* \subseteq \mathcal{X}$ such that the classifier $h'(x) = \mathbf{1}[x \in S^*]$ minimizes disparity $\mathbf{\Delta}(h') = \left|\sum_{i \in S} d_i\right|$ subject to a single constraint $\mathbf{U}(h') = \sum_{i \in S^*} u_i \geq \mathbf{U}(h^0)$, where we can think of $u(h^0)$ as some constant utility cutoff or threshold that constrains the search. Notice that if we switch the sign on utility, we can think of each point as having a *cost* $c_i := -u_i$, and our utility constraint can be thought of as a budget $B := -\mathbf{U}(h^0)$ that constrains how much we are able to spend.

We're left with the following optimization problem formulation for finding the LDA:

$$\min_{S \subseteq \mathcal{X}} \left| \sum_i d_i \mathbf{1}[x_i \in S] \right| \quad \text{s.t. } \sum_{i=0}^{N} c_i \mathbf{1}[x_i \in S] \leq B.$$

If there exists a solution to the above optimization problem, then it would qualify as an LDA and solve the problem put forward in Definition 1. However, solving this problem may be computationally onerous. Notice the optimization above looks almost identical to the 0-1 Knapsack Problem, which is known to be NP-hard. A noticeable difference in our case, however, is that we take an absolute value over the objective function. The full proof (in Appendix C.2) involves a reduction from a related problem, the Subset Sum Problem.

C.2 Proof of Theorem 2

Proof. We provide a polynomial-time reduction from the subset sum problem (known to be NP-complete) to the LDA problem. We use a particular case of the subset-sum problem (preserving the NP-hard property), defined below:

Definition 2 (Subset sum problem)**.** *Given a set $\mathcal{W}$ of integers $\{w_1, w_2, ..., w_{|\mathcal{W}|}\}$, find a subset whose values sum to* 0*.*

Suppose we have a black box that computes any LDA problem. We're given an instance $\langle \mathcal{W} \rangle$ of the subset sum problem. Our goal is to build an instance of the LDA problem where the solution enables us to solve the subset sum problem efficiently.

To start the proof, we will build a population of people that collectively compose our dataset. We can construct this population however we want, and at the end, we'll have an LDA search over this population where the LDA classifier (if one exists) will precisely identify the subset of integers that solve subset sum! The people will be split cleanly into categories, denoted by a categorical data value $x \in \mathcal{X}$, which takes values $\mathcal{X} \in \{1, 2, ...\}$. We construct the population as follows: cycle through the values in our subset sum problem $\{w_1, ..., w_{|\mathcal{W}|}\}$. For each element i, if the integer w_i is positive, add $2w_i$ people of group 1 to our population, each of type $x = i$. Otherwise, if integer w_i is negative, add $2|w_i|$ people of group 2 to our population, each of type $x = i$. We end up with a population that looks something like this:

$\mathcal{X}$	G	$n_g(x)$
1	1	$2w_1$
1	2	0
2	1	0
2	2	$2\|w_2\|$
$\vdots$	$\vdots$	$\vdots$
$\|\mathcal{W}\|$	1	$2w_{\|\mathcal{W}\|}$
$\|\mathcal{W}\|$	2	0

In the above example, the first subset sum value w_1 is positive, the second is negative, and the last is positive, which means that the people are assigned to groups 1, 2, and 1, respectively. We use $n_g(x)$ to refer to the number of people of data type x and group g.

To complete our population, we add two more data values, x^* and x^{**}. The first data value, x^*, contains a *single* person of group 1. The second data value, x^{**}, will equalize the total number of people of each group: we take the absolute difference between all the people we've created in group 1 and all the people we've created in group 2, which is equal to $|2\sum_i w_i + 1|$. We'll add this many people with data type x^{**} to whichever group has fewer people to equalize the overall number of people in each group. Finally, if we want, we can add additional people to the population, but they must have data type x^{**} and an equal number must be added from each group, to preserve balance. So we'll say formally that 2α people may be added for some non-negative integer α as long as exactly α are added from each of groups 1 and 2. The total number of people in our population is therefore $N := 2\sum_i |w_i| + 1 + |2\sum_i w_i + 1| + 2\alpha$.

Now, our population is fully specified. Since we know the total number of people (N), we are now able to specify the probability density function of our population, both overall and within group — the distribution is specified by $\rho_g(x)$. Our last step is to specify the underlying base rates for each data point in our population. For this, we simply assign $\sigma(x) = 1$ for all $x \neq x^{**}$, and $\sigma(x^{**}) = 0$.

Our completed population looks like this:

$\mathcal{X}$	G	$n_g(x)$	$\rho_g(x)$	$\sigma(x)$	$U(x)$	$h^0(x)$
1	1 2	$2w_1$ 0	$\frac{4}{N}w_1$ 0	1	$\frac{2}{N}w_1$	0
2	1 2	0 $2\lvert w_2\rvert$	0 $\frac{4}{N}\lvert w_2\rvert$	1	$\frac{2}{N}\lvert w_2\rvert$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\lvert\mathcal{W}\rvert$	1 2	$2w_{\lvert\mathcal{W}\rvert}$ 0	$\frac{4}{N}w_{\lvert\mathcal{W}\rvert}$ 0	1	$\frac{2}{N}w_{\lvert\mathcal{W}\rvert}$	0
x^*	1 2	1 0	$\frac{2}{N}$ 0	1	$\frac{1}{N}$	1
x^{**}	1 2	α $\lvert 2\sum_i w_i + 1\rvert + \alpha$	$\frac{2\alpha}{N}$ $\frac{2}{N}\left(\lvert 2\sum_i w_i + 1\rvert + \alpha\right)$	0	$-\lambda\frac{1}{N}\left(\lvert 2\sum_i w_i + 1\rvert + 2\alpha\right)$	0

Based on the way we've set up this population, we want the LDA search to select exactly the values in $\mathcal{X}$ that correspond to the indices of $\mathcal{W}$ that solve the subset sum instance (if and only if a solution exists). Therefore, we want the LDA search to never select our slack parameters x^*, x^{**}. The first slack parameter will never be selected because the density of the group-1 population in x^* is $\frac{2}{N}$, which is half the minimum difference between any two densities of group 2, so a LDA with disparity 0 could never include data x^*. The second slack parameter will never be selected as long as the utility of selecting it is negative, with greater magnitude than the sum of all other utilities in the entire search, since in that case, including it would always yield a utility less than $0 < U(h^0) = \frac{1}{N}$. Thus we require: $U(x^{**}) < -\left|\sum_{x \neq x^{**}} U(x)\right|$. We do some arithmetic on this condition and simplify to a requirement on the value α:

$$
\begin{aligned}
U(x^{**}) &< -\left|\sum_{x \neq x^{**}} U(x)\right| \\
-\lambda\frac{1}{N}\left(\left|2\sum_i w_i + 1\right| + 2\alpha\right) &< -\frac{2}{N}\sum_i |w_i| - \frac{1}{N} \\
\lambda\left|2\sum_i w_i + 1\right| + 2\lambda\alpha &> 2\sum_i |w_i| + 1 \\
2\lambda\alpha &> 2\sum_i |w_i| + 1 - \lambda\left|2\sum_i w_i + 1\right| \\
\alpha &> \frac{1}{\lambda}\sum_i |w_i| + \frac{1}{2\lambda} - \left|\sum_i w_i + \frac{1}{2}\right| \\
\alpha &> \frac{1}{\lambda}\left(\sum_i |w_i| + \frac{1}{2}\right)
\end{aligned}
$$

The point here is, our hardness result does not depend on any particular value of λ. We can complete the reduction for any feasible value $\lambda > 0$, as long as we set α to be sufficiently large. Now we're in a position to state the following Lemma, which represents the rest of what's needed for our proof:

Lemma 1. *Consider the LDA problem $\langle \mathcal{X}, \rho_g, \sigma, h^0 \rangle$ specified above, for any given value $\lambda > 0$. If an LDA does not exist, there is no solution to subset sum problem $\langle \mathcal{W} \rangle$. If an LDA h^* does exist, the indices of the data for which $h^*(x_i) = 1$ are the subset $\{w_i\} \subseteq \mathcal{W}$ which sums to 0.*

We prove this Lemma with the following sequence of claims:

Claim 3. *The LDA classifier will never select x^{**}.*

Proof. The utility from labeling x^{**} positively is a negative value that is strictly less than the sum of all other utilities. The utility of our starting classifier is $\frac{1}{N} > 0$. So, no classifier can positively label x^{**} and have at least as good utility as the starting classifier. □

Claim 4. *The LDA classifier will never select x^*.*

Proof. We've already established the LDA will never positively label x^{**}. If there exists an LDA, the disparity must be (strictly) less than the original classifier. But notice that every other data point has disparity value divisible by $\frac{4}{N}$. As the sole data point with disparity value $\frac{2}{N}$, including x^* must inevitably yield disparity at least $\frac{2}{N}$ rendering an LDA impossible. □

Claim 5. *An LDA exists if and only if the disparity values of positively-labeled data sum to 0.*

Proof. The total disparity of any classifier can by calculated by summing the densities $\rho_1(x) - \rho_2(x)$ of all positively-classified data and then taking an absolute value. Notice that aside from our slack variables $\{x^*, x^{**}\}$, all other data points' disparities are divisible by $\frac{4}{N}$. Thus, any sum of disparities from this group of data must also be divisible by $\frac{4}{N}$, meaning the least discriminatory grouping must have one of the following disparities: $0, \frac{4}{N}, \frac{8}{N}, \frac{12}{N}, ...$ Now, notice the disparity of our starting classifier h^0 is $\frac{1}{N}$. This value is strictly between 0 and the minimum non-zero disparity given our setup, $\frac{4}{N}$. Thus, any less-discriminatory classifier that is attainable from the non-slack variables ($x \in \{1, 2, ..., |\mathcal{W}|\}$) must have disparity values that sum to 0. Finally, we've already shown that no LDA solution will include a positive label for the slack variables x^*, x^{**}, so this concludes the proof. □

Claim 6. *The subset sum problem is invariant to a scaling factor k.*

Proof. This is a property of any summation. $a + b = c \iff ka + kb = kc$. In our case, the target is 0 so $k * 0 = 0$. The scaling factor we use is $k = \frac{2}{N}$ but this is easy to verify for any real-valued k. □

This concludes the reduction.

We've thus shown the full-information LDA problem is at least as hard as Subset Sum. Showing that the full-information LDA problem is NP-complete further requires that the problem is in NP. We know this is true because, given a candidate solution and a problem statement, we can simply check whether the solution has lower demographic disparity and greater or equal utility (which both require only taking a sum over the population).

□

C.3 A $(1 + \epsilon)$ Approximation

We've shown that the task of finding a less discriminatory alternative classifier is NP-hard, even in cases where we can access the true probability of a positive outcome for every data point. However, hardness does not imply impossibility, and there remain strategies for identifying LDAs. Here, we ask, what if we're okay with identifying a less discriminatory alternative so long as the starting classifier h^0 is at least some minimum distance from the *least* discriminatory alternative classifier?

Hypothetically, consider a case where the difference in disparity between a classifier h^0 and the least discriminatory alternative (which we can call h'') is 0.01. We may decide that such small differences are legally and ethically *de minimus*, i.e., too trivial or small to merit consideration. Equipped with some minimal value of this sort, denoted ϵ, perhaps we'd be satisfied if we can reliably and efficiently identify an alternative classifier whose disparity is within a factor of at least $(1 + \epsilon)$ of the least discriminatory possible alternative. If we have this approximation, then the *only case* where we might fail to find an LDA where one exists is when the starting classifier h^0 is exceedingly close to the least discriminatory alternative h''. In cases where the starting classifier is outside of a factor of $(1 + \epsilon)$ of h'', we can guarantee that we will find a less discriminatory alternative.

Definition 3 (Full-information approximate LDA)**.** *Given a set of data $\mathcal{X}$, a Bayes-optimal predictor $\sigma(x)$, a group-specific probability density function $\rho_g(x)$, a utility function $\mathbf{U}(h; \lambda)$, a baseline classifier $h^0(x)$, and $\epsilon > 0$, a **full-information ϵ-approximate LDA** is a classifier h' with utility at least $\mathbf{U}(h^0)$ and absolute disparity less than $|\mathbf{\Delta}(h^0)|(\frac{1}{1+\epsilon})$.*

Claim 7. *The full-information ϵ-approximate LDA can be identified in polynomial running time bounded by $O(n^3\epsilon^{-1})$.*

This claim's proof can be thought of as an exercise — and the proof is deferred — given its resemblance to the subset sum algorithm and approximation scheme.

C.4 Empirical Demonstration of Approximation Algorithm

To demonstrate the effectiveness of the approximation derived above empirically, we implement both exact and approximation algorithms to solve any LDA problem, specified by $\rho_g(x), \sigma(x), h^0(x)$, for some discrete number of values $x \in \mathcal{X}$. Equipped with these algorithms, we can evaluate their performance on instances of the LDA problem. These findings corroborate the formal results put forward in this section: that although the worst-case runtime explodes for a general instances of the LDA, there are efficient (i.e., polynomial-time) approximation algorithms that identify LDAs with high accuracy.

Simulating LDA problems. To simulate instances of the LDA problem, we first specify the number of discrete data values, $\texttt{n}_{\texttt{options}} := |\mathcal{X}|$, and the maximum number of digits per density specification, `maxdigits`. Given these specifications, a simulator generates two sets of uniform random numbers, each of size $\texttt{n}_{\texttt{options}}$, and then performs a manipulation so that each set adds to 1 and contains float values with at most `maxdigits` digits after the decimal place. These sets are used as values ρ_1 and ρ_2, the group-specific probability densities over the data. Finally, using the identical procedure, a simulator establishes the true probability values for each data option, $\sigma(x)$, this time using at most 2 digits after the decimal place. Finally, to create variation in the total size of the input, we simulate instances with varying values for $\texttt{n}_{\texttt{options}} \in \{4, 5\}$ and $\texttt{maxdigits} \in \{1, 2, 3, 4\}$. In total, we generate 202 instances of LDA problems, of which 164 contain an identifiable less discriminatory alternative and 86 contain no LDA. In this simulation, the input sizes range from 14 to 47 total decimal digits.

The runtime and accuracy performances are displayed in Figure 2.

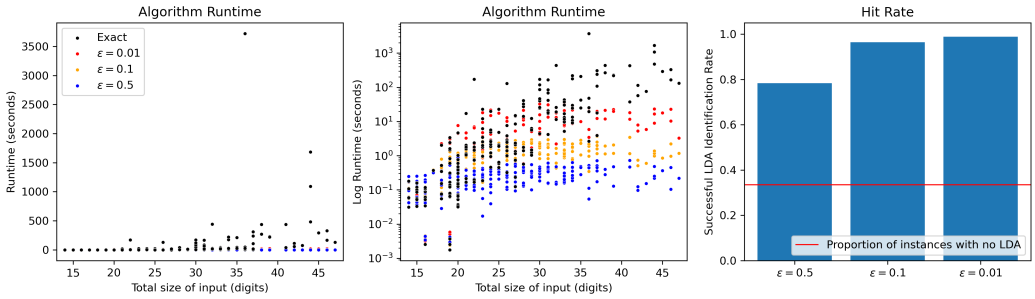


Figure 2: Runtime (left, center) and accuracy performance (right) of exact and approximate algorithms for computing the least discriminatory alternative classifier, given oracle access to $\mathcal{X}$, $\mathcal{G}$, and the Bayes-optimal predictor $\sigma(x)$. A total of 250 instances of the LDA problem were randomly generated for varying numbers of data values, starting classifiers, and the number of digits used to specify probability densities. For every generated instance, all four algorithms were run and runtime and accuracy were recorded. As the size of the input increases (measured as the total number of digits beyond the decimal places used to specify $\sigma(x)$ and $\rho_g(x)$), the worst-case runtime complexity for the exact algorithm explodes. However, polynomial-time approximations achieve high hit rate.

D Supplementary Materials on Consumer Welfare

Here we provide supplementary materials for the claims and proofs in Section 5.

D.1 Firm and Consumer Utility Differ

The LDA is a less discriminatory policy with some welfare guarantee. Part of what makes it attractive, conceptually, is that it seems to be a way to achieve a desirable goal (lowering disparity) without harming the decision maker (the firm). However, the firm is not the only stakeholder whose interests matter. The decision subject's welfare is also important as part of the broader set of societal con-

siderations [31], and indeed protecting the interests of disadvantaged consumers is the underlying motivation for fair lending interventions and regulations, including those that use the LDA.

Why might consumer welfare be left out of discussions of LDA? There are a number of reasons why this consideration isn't made explicit, or is paid only lip service. First, it can be enticing to make the simplifying assumption that firm and consumer welfare are aligned. Firms don't want to loan to consumers if it'll bury them in debt they aren't able to repay, and similarly, consumers are harmed by these sorts of predatory practices. At the same time, firms want to provide loans to applicants who are likely to repay a loan, because they profit from interest payments, and these applicants benefit from access to capital. Second, consumer welfare is hard to directly observe, compared to selection rates. Consumer welfare — the actual preferences of consumers over loan granting decision outcomes — is heterogeneous across the consumer population, depending on a host of factors including the true probability of default and circumstantial and contextual factors motivating the present need for cash, none of which can be observed by a regulator or a firm. Selection rates, on the other hand, are more easily and immediately observed.

Using lending rates and disparities as a proxy for the underlying goal (consumer welfare) begs the question: how good of a proxy is it? Though the answer to this question depends largely on context, we can make some strides towards answering the question with the help of modeling. In controlled environments where we can either observe ground truth labels or simulate them, perhaps we can begin to answer the question of when LDA searches based on selection rates align with consumer utility, and when they do not.

Formal Utility Model. Recall that we define utility to be $\mathbf{U} = \texttt{TPR} - \lambda\texttt{FPR}$. This definition of utility might be applicable for certain contexts. For example, in lending, there are benefits associated with being offered a loan, and costs associated when credit is awarded even though the borrower is likely to default. We note, however, that there are other functions (outside this function family) that might better represent welfare interests, especially in other contexts such as hiring, where a consumer might benefit from being offered a job whether or not they are 'deserving.' For the sake of this section's analysis, we say that instead of a single real value, λ can be a *vector* of real values, denoted $\vec{\lambda} \in \mathbb{R}^d$. We will most often consider the case where $d = 2$, and refer to the two utility values of interest as λ_f and λ_c, or the utility of the firm and consumer, respectively. Later in this section, we will motivate why we only consider *two* welfare values, because any vector of welfare considerations of length greater than 2 can be reduced to an LDA search with only two welfare considerations.

D.2 Proof of Claim 1

Proof. We provide an example to prove the above claim.

$\mathcal{X}$	$\mathcal{G}$	ρ	σ	h^0	h'
1	1	0.0	0.0	0	0
	2	0.7	0.0	0	0
2	1	0.8	0.5	1	0
	2	0.1	0.5	1	0
3	1	0.1	0.7	1	1
	2	0.1	0.7	1	1
4	1	0.1	1.0	1	1
	2	0.1	1.0	1	1

In the above case, h' represents an LDA when $\lambda_f = 0.5$ but is strictly consumer-harming when $\lambda_c = 2$. □

D.3 Proof of Claim 2

Proof. Given a classifier h' that satisfies $\mathbf{U}(h'; \lambda_1) > \mathbf{U}(h^0; \lambda_1)$ and $\mathbf{U}(h'; \lambda_2) > \mathbf{U}(h^0; \lambda_2)$, assume for contradiction that there exists $\lambda_{1.5} \in (\lambda_1, \lambda_2)$ with $\mathbf{U}(h'; \lambda_{1.5}) < \mathbf{U}(h^0; \lambda_{1.5})$. Then we have three inequalities:

- $\texttt{TPR}' - \lambda_1\texttt{FPR}' > \texttt{TPR}^0 - \lambda_1\texttt{FPR}^0 \rightarrow \lambda_1(\texttt{FPR}^0 - \texttt{FPR}') > (\texttt{TPR}^0 - \texttt{TPR}')$

- $\texttt{TPR}' - \lambda_2 \texttt{FPR}' > \texttt{TPR}^0 - \lambda_2 \texttt{FPR}^0 \rightarrow \lambda_2(\texttt{FPR}^0 - \texttt{FPR}') > (\texttt{TPR}^0 - \texttt{TPR}')$
- $\texttt{TPR}' - \lambda_{1.5} \texttt{FPR}' < \texttt{TPR}^0 - \lambda_{1.5} \texttt{FPR}^0 \rightarrow \lambda_{1.5}(\texttt{FPR}^0 - \texttt{FPR}') < (\texttt{TPR}^0 - \texttt{TPR}')$

The above inequalities are impossible to satisfy unless $\lambda_{1.5} < \lambda_1$ or $\lambda_{1.5} > \lambda_2$, which is false. □

E Empirical Results

Here we describe a set of empirical experiments. The aim of these tests is to better understand the effectiveness and characteristics of certain simple heuristic search methods for finding an alternative classifier with lower disparity and higher utility performance on new data. We do not expect these methods to identify a Pareto-optimal, fair-and-accurate classifier on unseen data — an algorithm designer would need to be unfathomably lucky to arrive at a classifier of that sort, even on relatively small datasets. Instead, the methods we test search for alternative classifiers that are similar to the starting classifier, by changing the random seed or randomly sampling and re-training. After producing, say, 100 similar models, the methods we test evaluate the level of disparity using available data (i.e., an evaluation set) and select the minimally-disparate alternative.

Although these methods only search a narrow slice of possible classifiers, they have some desirable properties that make them easy to work with and analyze. First of all, they do not explicitly model disparity and they produce a selection policy that does not, explicitly, use the protected attribute. In regulated industries like lending, discrimination based on these features is illegal. Second, the methods produce a set of alternative models that are in the same model class as the original classifier. This provides a convincing argument for meeting the business need for generalizable performance: if a model falls within the same class of models as the starting classifier, it does not introduce new complexity or expressiveness that could exhibit good performance on training data that does not translate, in reality, to new data. Third and finally, because the methods we use produce a set of similar models from the same model class, their disparity and utility performance on held-out evaluation data will represent identical and independent draws from a single distribution. This gives the algorithm designer good reason to select the *minimum disparity* classifier on evaluation data as a best-guess for the best-performing model in general. So-called disparate learning processes — attempts to produce group-independent classifiers using a group-aware learning process — have been analyzed at length, and there is reason to believe other methods may more effectively find less discriminatory models [33].

Data and models. For this test, we use two datasets, `Adult` and `German Credit`. Both datasets are collected in financial settings. We use the `Adult` dataset to perform a classification where the task is to predict which individuals make over $50,000 in income per year. We use the `German Credit` dataset to perform a classification where the task is to predict credit-worthiness (as defined in the dataset's meta-data). We test three starting classifiers: a Logistic Regression, a Random Forest, and a Decision Tree classifier. We implement these classifiers in Python using `sklearn`, typically with pre-set default hyper-parameters (though we do set `max_depth=5` for the random forest classifier). Though the datasets do not have balanced class weights, we train our models using balanced weights. A more detailed description of the datasets and models is provided in Appendix F.

Search processes. We conduct two search types to identify proximate models from the same model class. The first is sampling and re-training — we take a random sample, with replacement, the size of the training data and re-train the same model on this new sample of data. The second method is testing different random seeds. We only use this second method for the Random Forest model because it is an ensemble model that explicitly uses pseudorandomness and the model changes significantly when we perturb the `random_seed` parameter.

Experiment procedure. First, we split the data into train, evaluation, and test sets. Using the training data, for each search process and model type, we train 10,000 models (generated at random) and record their performance and qualities in a dataset. The utility, disparity, and selection rate is measured on training, evaluation and testing data for every model, and recorded in a dataset. To test the procedure of searching over n alternative models, we draw from our dataset n times for $n \in \{1, ..., 100\}$. Equipped with a subset of n models, we select the model with lowest disparity on the evaluation data, and record its utility and disparity performance on test data. The average 'lift' in utility or reduction in disparity is attained by comparing this selected model to the average performance over the n models. For each value of n, we perform this sub-sampling procedure 2000

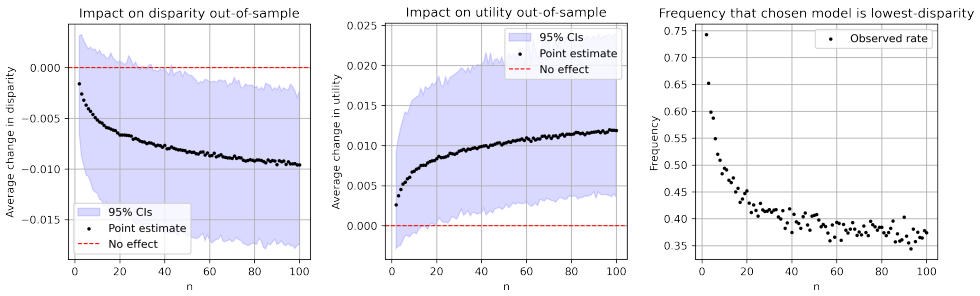


Figure 3: Results from a simple randomized search for a less discriminatory alternative algorithm on the `Adult` dataset. We search by randomly sampling with replacement and re-training a Random Forest classifier n times for $n \in \{2, ..., 100\}$ on a training dataset, and selecting the minimum-disparity classifier on an evaluation set. As n increases within this range, we find disparity decreases, on average, on a held out (out-of-sample) test data (a; left). In this case, we also find that utility does not diminish from this procedure (b; center). However, as the number of random draws increases, the probability of having selected the optimally-performing model out-of-sample decreases (c; right) suggesting that conducting an effective search might require passing over models that end up having lower disparity.

times, and record the mean, 2.5th and 97.5th percentiles to produce point estimates, lower-bounds, and upper-bounds, respectively.

Finally, we also track the rate at which the procedure tested produces a *perfect guess* of the out-of-sample disparity-minimizing model, over the set of models tested. That is, if we conduct our procedure with $n = 100$, we measure the frequency that the lowest-disparity model according to evaluation data is also the lowest-disparity model on the test dataset.

We note that we do not claim to have the *best* or *optimal* procedure for searching, nor do we claim that the method we put forward is *reasonable* in the legal interpretation as advocated by Black et al. [6]. Instead, we aim to explore whether, in certain instances, simple methods can attain generalizable reductions in discrimination at no cost to accuracy (utility). We also wish to test whether these methods consistently arrive at the *least* discriminatory classifiers out-of-sample, or whether they open up the possibility that firms test and reject alternatives that end up with lower disparity in hindsight.

Table 1: Out-of-sample performance of various LDA search procedures

Dataset	Model	Search	Disparity	Utility	Freq. min-disp.
`Adult`	Decision Tree	Sample	-0.0453*	-0.0110	0.7310
	Logistic Reg.	Sample	-0.0170*	-0.0189*	0.6390
	Random Forest	Seed	-0.0058*	0.0130*	0.2980
		Sample	-0.0096*	0.0119*	0.3740
`GC`	Decision Tree	Sample	-0.0019	-0.0008	0.0105
	Logistic Reg.	Sample	-0.0136	-0.0061	0.0175
	Random Forest	Seed	0.0015	-0.0032	0.0040
		Sample	-0.0061	0.0034	0.0125

Results. The main results are reported in Table 1. The tests provide evidence that certain search methods can reveal the existence of lower-disparity alternative classifiers whose performance generalizes out-of-sample. The asterisk * signifies that the results are statistically significant according to the boostrapped 95% confidence intervals, attained by repeating the search procedure 2000 times. In other words, the asterisk tests whether the directional change in disparity or utility was observed in at least 95% of the 2000 times the procedure was tested. Disparity reductions are statistically significant according to this procedure on the `Adult` dataset for all models and search methods tested. The disparity effects on the `German Credit` dataset are not significant, which is explained both by the much smaller sample size and the minuscule starting disparity level.

The Random Forest model has the highest utility measure on both datasets compared to the other models tested. It also had the highest absolute disparity on both datasets. These results are depicted

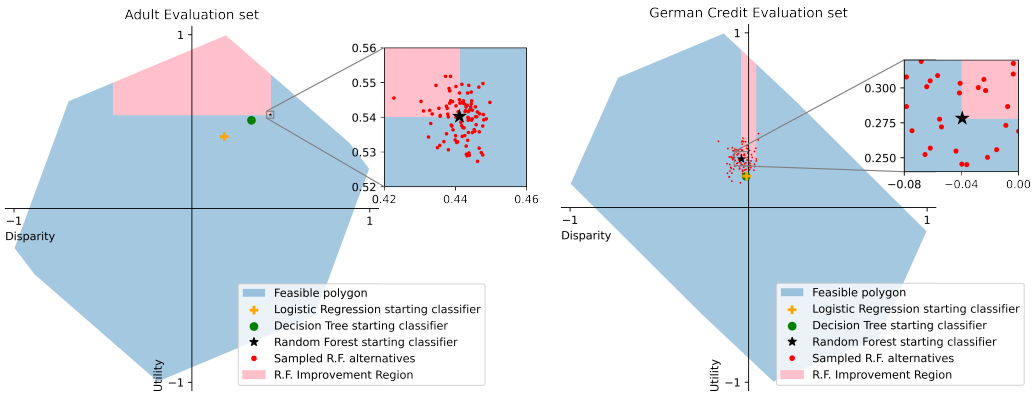


Figure 4: Empirically observed achievable polygon using the evaluation data sets for the `Adult` and `German Credit`. In the case of `Adult` (left), the Random Forest starting classifier exhibits wide disparities and randomly sampled alternatives uncovers LDAs with significant (though small) gains on the training set. In the case of `German Credit` (right), the starting classifier does not exhibit wide disparities, leaving a narrow region for LDA improvement. No statistically significant disparity reduction is observed on training data.

in Figure 4. Performing an LDA search procedure on the RF *increases* utility in our `Adult` test, meaning in this case, utility increases and disparity decreases statistically— no trade-off is observed. The observed differences in these values is visualized in Figure 3. In other cases, with other models, however, the utility decreases. This makes sense, given nothing about our procedure guarantees or tries to maximize directly the utility.

One feature of the procedure we test is that it checks, and rejects, 99 models in favor of the model that has lowest disparity on an evaluation dataset. We find this method, even when it consistently reduces disparity out-of-sample, can increase the likelihood of *passing over* the actually disparity-optimal model out-of-sample. The frequency that the selected model is disparity-optimal of those considered is displayed in the rightmost plot in Figure 3.

The methods tested are not the only way to systematically search for less discriminatory algorithms. We could imagine altering the loss function of an algorithm directly so that the algorithm optimizes some weighted combination of disparity reduction and utility. We could similarly imagine encoding knowledge about the structure of disparity — if certain attributes are likely proxies for the protected group belonging, using more bespoke modeling assumptions could better target de-biasing interventions. Some may find the methods we test desirable, however, because they do not use pre-existing knowledge about groups or disparity, as they simply draw from the set of plausible good models given a pre-existing model class.

F Supplementary Materials on Empirics

Data and Models (further information). Every row in the `Adult` dataset represents features of an individual, and the outcome variable is an indicator representing whether they make over $50,000 in annual income. The data is on American adults from census information in 1994. We use a total of four variables – `maritalstatus` (a categorical variable representing whether the individual is single/divorced/married/etc), `hoursperweek` (a numerical variable representing the number of hours the individual works per week), `education` (a categorical variable specifying the level of education achieved, e.g. college degree), and `workclass` (a categorical variable representing the type of work). The protected attribute is the `gender` (given as a binary M/F). The dataset contains a total of 32,561 rows.

The `German Credit` dataset contains loan decisions in Germany, and was accessed via the UCI machine learning database. The dataset contains a total of 1,000 rows with categorical and numerical features related to individual's credit, financial status, employment, and loan application. The following five features were used to train the models: `credit_history_category` (categorical variable representing information about credit history), `credit_amount` (numerical variable representing amount of credit requested by loan applicant),

unemployment_category (categorical variable with information about whether the individual is unemployed), installment_rate_percentage_income (numerical feature representing the ratio between loan amount and income), and present_residence_duration (a numerical variable representing the amount of time the applicant has resided in their current residence). The protected attribute is the gender (given as a binary M/F).

G Acknowledgments

The authors would like to thank Emily Black, Pauline Kim, Logan Koepke, Mingwei Hsu, and the members of the Fairness, Accountability, Transparency, and Ethics group (FATE) at Microsoft Research, the AI, Policy and Practice group (AIPP) at Cornell University, and the Digital Life Initiative (DLI) at Cornell Tech for providing feedback on this work. The work is supported in part by a grant from the John D. and Catherine T. MacArthur Foundation. Much of the work was completed while Laufer was an intern at MSR. He is additionally supported by a LinkedIn-Bowers CIS PhD Fellowship, a doctoral fellowship from DLI, and a SaTC NSF grant CNS-1704527. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of NSF or other funding agencies.