The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice

Andrew Bell
New York University
New York, NY, USA
alb9742@nyu.edu

Lucius Bynum
New York University
New York, NY, USA
lucius@nyu.edu

Nazarii Drushchak
Ukrainian Catholic University
Lviv, Ukraine
naz2001r@gmail.com

Lucas Rosenblatt
New York University
New York, NY, USA
lucas.rosenblatt@nyu.edu

Tetiana Zakharchenko
Ukrainian Catholic University
Lviv, Ukraine
t.herasymova@ucu.edu.ua

Julia Stoyanovich*
New York University
New York, NY, USA
stoyanovich@nyu.edu

ABSTRACT

The "impossibility theorem" — which is considered foundational in algorithmic fairness literature — asserts that there must be trade-offs between common notions of fairness and performance when fitting statistical models, except in two special cases: when the prevalence of the outcome being predicted is equal across groups, or when a perfectly accurate predictor is used. However, theory does not always translate to practice. In this work, we challenge the implications of the impossibility theorem in practical settings. First, we show analytically that, by slightly relaxing the impossibility theorem (to accommodate a *practitioner's* perspective of fairness), it becomes possible to identify abundant sets of models that satisfy seemingly incompatible fairness constraints. Second, we demonstrate the existence of these models through extensive experiments on five real-world datasets. We conclude by offering tools and guidance for practitioners to understand when — and to what degree — fairness along multiple criteria can be achieved. This work has an important implication for the community: achieving fairness along multiple metrics for multiple groups (and their intersections) is much more possible than was previously believed.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Social and professional topics** → **Socio-technical systems**;

KEYWORDS

machine learning, fairness, public policy, responsible AI

ACM Reference Format:
Andrew Bell, Lucius Bynum, Nazarii Drushchak, Lucas Rosenblatt, Tetiana Zakharchenko, and Julia Stoyanovich. 2023. The Possibility of Fairness: Revisiting the Impossibility Theorem in Practice. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3593013.3594007

*Bell is the first author, Stoyanovich is the senior author, other authors are listed alphabetically.

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594007

1 INTRODUCTION

Increasingly, artificial intelligence (AI) and machine learning (ML) systems are being implemented in domains like employment, healthcare, and education to improve the efficiency of existing processes [39, 57, 65]. In tandem with this uptick in adoption, there are growing concerns about the potential for ML systems to cause significant harm to members of already marginalized groups. For example, it has been found that some lending algorithms discriminate against Latinx and African-American borrowers [4, 27], some prevalent medical algorithms discriminate against Black patients [48], and some educational risk-assessment algorithms perform worse for minority students [32, 48, 56].

The risk of discriminatory ML systems has led to significant interest in methods for measuring and ensuring "algorithmic fairness." Researchers have created robust processes and tools for auditing algorithmic systems for bias based on various definitions of fairness, such as *Demographic Parity, Equalized Odds Ratios*, and *Predictive Parity* [13, 18, 41, 54]. Choosing a context-specific fairness definition (also called a *fairness metric*) depends on value judgments, and often several metrics may be situationally relevant [3]. For instance, in contexts where the output of an algorithmic system is assistive, disparities in the *False Negative Rate* between groups can be used as a measure of discrimination with respect to group need [54].

In contexts where *more than one metric* is applicable, practitioners, stakeholders, and the wider public may engage in a debate about which metric to choose [60]. Debates of this nature have yielded a number of notable results in the algorithmic fairness literature, including a fundamental result known colloquially as the "impossibility theorem" simultaneously reported on by Chouldechova [16] and Kleinberg et al. [37]. The impossibility theorem asserts that, for binary classification, equalizing some specific set of multiple common performance metrics between protected classes is impossible, except in two special cases. The first special case is when an algorithm is a *perfect predictor*, and the second is when the *prevalence* of the outcome being predicted (i.e. the percentage of individuals in a group with the positive outcome, also called *base rate*) is equal across groups. As a consequence of this theorem, researchers and practitioners have focused on understanding trade-offs between fairness and predictive accuracy in an algorithmic

system, often designing bias audits and mitigation techniques that center on a single chosen fairness metric [18].

Though important and strong, the implicit assumption of the impossibility result (namely, that a practitioner might think about fairness as *exactly* equalizing metrics) may not actually apply to a wide array of real-world problems. In fact, a growing body of research suggests that the limitations to fairness derived by Kleinberg et al. [37] and Chouldechova [16] may not be particularly relevant in many practical settings [14, 31, 50, 61].

Note. Throughout this paper, we will often refer to metrics like FPR (False Positive Rate), FNR (False Negative Rate), PPV (Precision) and ACC (Accuracy). Though these metrics are common, we seek to make our work accessible across levels of technical expertise by providing equations and descriptions of these metrics in Appendix Section C. On a related note, throughout this work we concern ourselves with *binary classification*, a standard machine learning task where one attempts to assign the correct binary label (positive/negative) to each individual in a population. In fairness literature, it is common to consider at least two groups within that population, and then to compare the performance of a binary classifier on each sub-population. We also provide an in-depth definition of binary classification, and consideration of groups in the population, in Appendix Section C.

Summary of contributions. Variations on the "impossibility theorem" specific to binary classification (with a protected class) state that, when equalizing certain metrics (like FPR, FNR, PPV or ACC) between two groups, we should hesitate to consider multiple metrics at once. Why? The fairness constraints (equalizing three of these metrics between groups) will only be *exactly* satisfiable if we have a perfect predictor or outcome prevalence parity [16, 37].

We suggest that this setting is unrealistic. Our paper's driving insight is that practitioners are often more than comfortable with *approximate* fairness guarantees, as opposed to enforcing *exact* equality between metrics. Therefore, we focus on a set of more realistic fairness constraints, where we are allowed to slightly relax between-group metric equalities for FNR, FPR, PPV, and ACC. To our knowledge, we are the first to study at length this relaxed setting from a practitioner's point of view.[1] This framing yields a straightforward research question: under what type of setting and relaxation is it possible to find classifiers that are "fair" along seemingly incompatible fairness constraints? And how can practitioners determine if a "fair" classifier exists for *their* predictive context?

For example, it turns out that if I, as a practitioner, say "I have prevelances p_1 and p_2 for Groups 1 and 2 in dataset X, and I am willing to tolerate a difference of $Y\%$ when equalizing metrics," then I have all of the information I need to determine if finding such a model is truly impossible (or not!) *before even attempting the problem.* Encouragingly, and perhaps counter-intuitively, our analysis suggests that in practical settings the answer is often "yes, it's possible to find that fair model." **Our major finding is that if one allows only a small margin-of-error between metrics, sets of models satisfying three fairness constraints simultaneously are abundant, rather than rare.** In our corresponding experiments on real datasets, we find empirically that the resource constraint k (i.e., having k loans to give out or k job interview slots to fill) also plays a significant role in feasibility, where a smaller k can result in more feasible models.

Paper organization. We begin with background and related work in Section 2. After, we approach the problem analytically in Section 3. We state a formula balancing FPR, FNR and ACC between groups with fairness relaxations for each metric. In this setting, we are able to derive a powerful tool in the form of a simple formula relating the feasibility of fairness to a specific relaxation strength, given a classification scenario. However, in many resource-constrained settings, practitioners care more about PPV than ACC. So, we then turn our attention to the problem in terms of FPR, FNR and PPV, but find that it is difficult to analyze in closed-form. Instead, through principled approximations, we are able to provide much the same guidance to practitioners as a direct analytical solution would, and leave deriving a closed-form result to future work. Additionally, to demonstrate the utility of our fairness relaxation insights, we conduct extensive experimental evaluations on five real-world datasets. The results of these experiments, discussed in Section 4, corroborate our insights and compellingly demonstrate the *possibility* of fairness. We discuss our insights and offer guidance to practitioners in Section 5 and conclude in Section 6. **For those who wish to go directly to our margin-of-error based fairness feasibility recommendations, skip to Section 5**.

2 BACKGROUND AND RELATED WORK

Algorithmic fairness. Significant progress has been made in understanding algorithmic fairness [45]. Broadly, this literature concludes that fairness is not a monolith: there are *many* different ways to think about algorithmic fairness, and defining what is "fair" is a matter of philosophy, incorporating one's worldview, mitigation objectives, and an algorithm's context-of-use [3, 24]. In response to the complex and nuanced nature of fairness, researchers have defined dozens of *fairness metrics*, or mathematical assessments of an algorithm's prejudice, that address different aspects of fairness [7, 9, 11, 16, 25, 38, 44, 54, 59, 64]. These metrics can be divided into two categories: those that consider the output of an algorithm, and those that consider errors made by the algorithm. As an example of the former, *Disparate Impact* (or *Proportional Parity*) measures the proportion of a group receiving the positive classification outcome relative to the proportion of the group in the input. As an example of the latter, the difference in *False Negative Rates* between groups can be used to assess whether one group is erroneously "passed over" for a positive outcome relative to another. Importantly, there is no one-size-fits-all metric for evaluating the fairness of algorithms. Some tools (like the *Fairness Tree* [55]) have been developed to help navigate the challenge of selecting an appropriate fairness metric, but ultimately, it is necessary for researchers and practitioners to have meaningful conversations with those impacted by algorithms to select fairness metric(s) specific to the context-of-use [53, 54].

Typically, metrics judge the fairness of a predictor by considering the *imbalance* between group-specific metrics. We can calculate imbalance as a difference — mean, squared, absolute, etc. —

[1] Previous work showed that a version of the impossibility result exists on the boundary of the relaxed setting [37], but the authors did not fully explore the space of relaxed solutions, nor did they position it from the viewpoint of practitioners. This is further discussed in Section 2.

or as a *disparity* — the ratio of a metric of one group, g_j, to that of a *reference group*, g_{ref}, usually chosen as the majority group: $disparity_{g_j} = \frac{\text{metric}_{g_j}}{\text{metric}_{g_{ref}}}$. Often, the goal of algorithmic fairness is to achieve *parity*, that is, to *eliminate* the imbalance between fairness metrics entirely. Importantly, the tolerable level of difference/disparity for a given fairness metric is highly dependent on the algorithm's context-of-use. Perhaps counter-intuitively, there are cases where we want to enforce a large disparity (see Rodolfa et al. [51], who discuss intentionally over-representing a marginalized group for an assistive intervention).

While some fairness metrics are incompatible with one another[2], and others are approximately mathematically equivalent [52], many are compatible and distinct. Consider the following scenario: a high school is using an algorithm to predict which students are at risk of failing ninth grade, so that high-risk students can be offered a special tutoring intervention. School administrators may want an algorithm that selects an equal number of privileged and underprivileged students, and also does not unfairly pass over students who are badly in need of tutoring. This would imply a need for both *Demographic Parity* and *False Negative Rate Parity* between groups. Yet, as we argued in the introduction, multiple fairness metrics are rarely considered in practice, and most existing bias mitigation methods enforce a single metric (or at most two metrics) at a time [8, 23, 30, 34–36, 46, 49, 51, 64]. In part, this is due to the *impossibility theorem*, a foundational result, presented simultaneously by Chouldechova and by Kleinberg et al..

The impossibility theorem. As stated by Kleinberg et al. [37], this theorem shows that three common metrics — equalizing calibration within groups, and enforcing balance for the negative class and for the positive class — cannot be simultaneously satisfied for multiple groups, outside of two special cases [37]. These cases are (1) when the algorithm is a perfect predictor and (2) when there is no prevalence difference between groups. Chouldechova [16] states an equivalent impossibility, presented as the relationship between the *Predictive Positive Value* (PPV), *False Positive Rate* (FPR), *False Negative Rate* (FNR), and *prevalence* (p):

$$\text{FPR} = \frac{\text{p}}{1-\text{p}} \frac{1-\text{PPV}}{\text{PPV}} (1-\text{FNR}) \tag{1}$$

Exploring implications of the impossibility theorem. Importantly, the impossibility results imply an upper bound on how many fairness metrics can be satisfied simultaneously without a perfect predictor. Kleinberg et al. addressed a related question on approximate conditions of the impossibility result, showing that approximate fairness definitions can simultaneously hold, but only under ϵ-approximate prevelances or ϵ-approximate perfect prediction [37]. *Significantly, Kleinberg et al. did not explore the space of solutions under ϵ-approximate relaxations of constraints, nor did they detail the implications that these relaxations might have for practitioners.*

A thought experiment. To motivate our exploration of this space, consider the following thought experiment. The *achievement gap* is one of the most pervasive examples of racial disparities in education, in which Black and Brown students graduate from high school at a rate roughly 10% lower than that of White students [40]. How should practitioners think about a prevalence difference of 10% when designing algorithms that predict student performance? Is it a small or large difference, and what implication does it have for the abundance (or rarity) of models satisfying ϵ-approximate fairness constraints or ϵ-approximate perfect prediction?

Practice versus theory. Understanding the space of feasible models under a relaxation of the impossibility theorem is particularly salient in light of recent work showing that theoretical trade-offs do not always apply to real-world settings [14, 50, 61]. Rodolfa et al. introduced a method for finding models that were fair with respect to FNR without sacrificing a model's PPV, and demonstrated the effectiveness of their approach in four separate ML-for-public-policy problems [50].[3] It was hypothesized by the authors that the negligible trade-off is the result of the resource-constrained nature of applied problems, where fairness and model performance are measured with respect to the top-k, rather than at an arbitrary threshold. In our work, we begin to formalize this intuition in Theorem 3.5.

Other works also challenge the idea that accuracy and fairness are in tension. Celis et al. [14] developed a meta-algorithm for a large family of classification problems with convex constraints, and demonstrated that one can achieve near-perfect fairness while sacrificing negligible accuracy. Similarly, Wick et al. [61] propose a semi-supervised learning approach that improves both fairness and accuracy. A third recent example is the *MFOpt* framework proposed by Hsu et al. [31] that simultaneously optimizes *Demographic Parity*, *Equalized Odds*, and *Predictive Rate Parity*—those fairness notions that are mathematically incompatible according to the impossibility theorem. Similar to our work, Hsu et al. were motivated by doubts about the strength of the impossibility theorem in practical settings. Notably, these works have focused on methods for mitigating disparity for multiple fairness metrics while maintaining high model accuracy, but have *not* provided much analysis of their implicit relaxing of fairness metric parity.

In our analysis, we center two considerations common to practical settings. First, in practice, one generally does not require fairness metrics to be *exactly equal* across groups to achieve fairness. For example, depending on the context of use, a classifier that has an FPR difference between groups of 2%, 5% or even 10% may be satisfactory. Second, we consider the presence of a resource constraint. For example, a commonly used performance metric in applied ML problems is PPV-at-k, where k represents a real-world resource constraint [1, 12, 62].

3 FINDING FEASIBLE MODELS

To encode that practitioners are generally okay with approximate fairness constraints as opposed to strict constraints, we begin by directly re-parameterizing the impossibility theorem with relaxations for each parameter. The relaxed constraints afford us a *space of solutions*, where each solution represents a potential classifier that balances all three metrics within our desired tolerance. We call this space of solutions the **fairness region**. Exploring how this region changes across different contexts/relaxations/metric settings can tell us when satisfying all constraints is feasible and, further, when

[2]For example, one cannot simultaneously satisfy equal selection and proportional parity unless the prevalence of outcomes is the same in both groups

[3]Rodolfa et al. refer to *Recall Parity* in their work, but state that it is mathematically equivalent to FNR Parity for small population sizes.

we can expect greater flexibility when searching for a model across multiple metrics.

The impossibility theorem can be stated in terms of different metrics, and our choice of metric impacts the ease or difficulty of characterizing the fairness region in closed-form. We start with a choice of metrics for which we have a closed-form characterization of the fairness region: FNR, FPR, and ACC (Section 3.1). Motivated by the fact that practitioners in resource constrained settings often consider PPV instead of ACC, we then consider a fairness region for FNR, FPR, and PPV (Section 3.2). Deriving a closed-form solution for the fairness region in the second case is much more difficult, requiring us to approach our analysis computationally.

3.1 Characterizing the fairness region using FPR, FNR, and ACC

We begin by defining an alternative expression for the impossibility result, this time in terms of FPR, FNR, and ACC. Proofs of all results in this section (Corollary 3.1, Proposition 3.2, and Theorem 3.3) can be found in Appendix D.

Corollary 3.1 (Impossibility Result Variation [16] [37]). *In a binary classification setting (see Appendix Section C), the relationship between ACC, FNR, FPR and p can be characterized by:*

$$ACC = (1 - FNR)p + (1 - FPR)(1 - p).$$

Next, we add a relaxation term for each parameter in Corollary 3.1. In the case of two groups, we let $\text{FPR}_2 = \text{FPR}_1 + \epsilon_{\text{FPR}}$, where ϵ_{FPR} is a tolerable difference between the metric for the two groups. Similarly, let $\text{FNR}_2 = \text{FNR}_1 + \epsilon_{\text{FNR}}$ and $\text{ACC}_2 = \text{ACC}_1 + \epsilon_{\text{ACC}}$. Using these relaxations, we can express a "governing equation" for the fairness region as follows.

Proposition 3.2 (Describing The Fairness Region). *Consider Corollary 3.1. Assume that $p_2 = p_1 + \epsilon_p$, $ACC_2 = ACC_1 + \epsilon_{ACC}$, $FPR_2 = FPR_1 + \epsilon_{FPR}$, and $FNR_2 = FNR_1 + \epsilon_{FNR}$, where each $\epsilon_{FPR}, \epsilon_{FNR}, \epsilon_{ACC}, \epsilon_p \in (-1, 1)$ term captures the difference between two groups for p, ACC, FPR, and FNR, respectively. Then, the following is equal to FNR_1:*

$$\frac{-\epsilon_{FPR} + \epsilon_{ACC} + \epsilon_{FPR} \cdot p_1 - \epsilon_{FNR} \cdot p_1 + FPR_1 \cdot \epsilon_p + \epsilon_{FPR} \cdot \epsilon_p - \epsilon_{FNR} \cdot \epsilon_p}{\epsilon_p} \tag{2}$$

While Equation 2 may look complex, an important insight is that it shows FNR can be expressed as a function of mostly fixed and known terms. **Observe that p and ϵ_{p} are known *a priori*, as they can be calculated directly from the dataset.** By deciding on bounds for the acceptable tolerance between fairness metrics (i.e., maximum allowable values for $\epsilon_{\text{FPR}}, \epsilon_{\text{FNR}}, \epsilon_{\text{ACC}}$), we can then create plots of FNR vs. FPR, as seen in Figure 1 (a). Each point in these plots represents an FPR, FNR (and, implicitly, an ACC value) of a feasible model. In other words, these points correspond to the existence of feasible models satisfying fairness constraints for FPR, FNR, and ACC within an ϵ-margin-of-error. Similarly, the absence of a point corresponds to the emptiness of a set of models (i.e., the infeasibility of *finding* a model). In general, we can say that plotting FNR vs. FPR according to Proposition 3.2 gives us a projection of the fairness region, where the size of the area provides a measure (out of the entire FPR,FNR $\in [0, 1]$ region) of the proportion of feasible models that are fair across all three metrics of interest (out of all possible metric values). Significantly, we can use Equation 2 to find a closed-form expression for the size of the fairness region over the unit square FNR, FPR $\in [0, 1]$:

Theorem 3.3 (Size of the Fairness Region). *Assume $\epsilon_p < 1 - p$. Allow $\pm\gamma$ to be the symmetric acceptable error (our "fair" relaxation) between groups for metrics FPR, FNR, and ACC. Consider the size of the space of possible $\epsilon_{FPR}, \epsilon_{FNR}, \epsilon_{ACC}$ assignments, given ϵ_p and p that satisfy the constraints from Proposition 3.2. We will denote the size of that space as $|A_f|$ (as shorthand, we will call this the "fairness region"). For a set of fairness constraints $-\gamma \leq \epsilon_{FPR}, \epsilon_{FNR}, \epsilon_{ACC} \leq \gamma$, where $|\gamma| \leq 1$ and $\gamma \neq 0$, we have that $|A_f|$ is simply:*

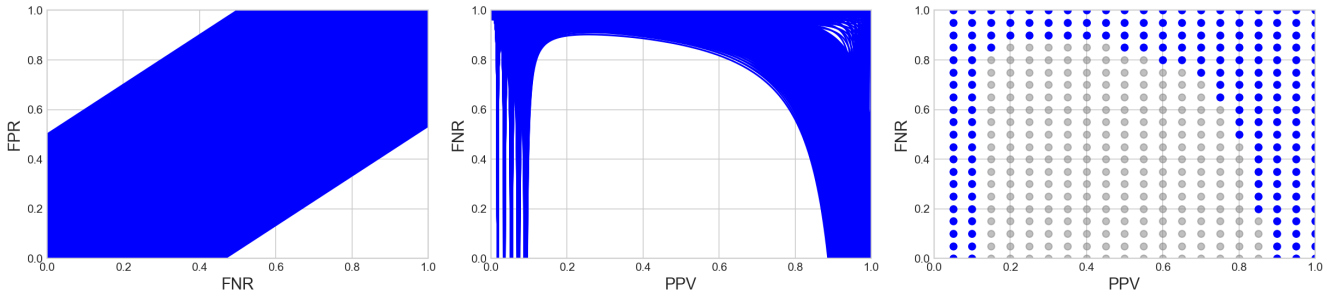
$$|A_f| = \frac{4\gamma}{\epsilon_p} - \frac{4\gamma^2}{{\epsilon_p}^2} \tag{3}$$

The practical implication of Theorem 3.3 is simple: for a practitioner with a target tolerance for FPR, FNR, and ACC across groups, we can show them whether their fairness relaxation values will work or not given their context (i.e., p and ϵ_{p}), and furthermore, how relaxing (or tightening) their ϵ-margin-of-error affects the fairness region.

3.2 Characterizing the fairness region using FPR, FNR, PPV

Theorem 3.3 provides a clean and convenient result for FPR, FNR and ACC, but it does not allow us to meaningfully analyze the type of resource-constrained settings generally faced by practitioners. In many contexts, a classifier's ACC has less meaning than its PPV [1, 5, 12, 50, 62]. To this end, we attempted to recreate the analysis in Section 3.1 instead using FPR, FNR, and PPV, which is found in Appendix E. This analysis with PPV instead of ACC leads us to an analogous expression for FNR as a function of the other parameters (see 8). However, the expression in the PPV case is ripe with non-linearities and possible discontinuities, making it more difficult to find a closed-form expression for the size of the fairness region (in the same way we did for ACC in Theorem 3.3). A corresponding plot of the fairness region projected onto two dimensions (FNR and PPV) is shown in Figure 1 (b). The figure is a discretized set of solutions created by sweeping out a range of parameter values and plotting feasible lines following the equation for FNR (see 8).

With no closed-form expression, we take two computational approaches to understanding the size of this fairness region. The first approach is to directly estimate the fraction of the unit square (FNR, PPV $\in [0, 1]$) taken up by a discretized feasible region by using a *dot planimeter*, which is a well-studied method for estimating complex two-dimensional areas [10, 26]. Intuitively, dot planimeter estimates the fraction of the unit square taken up by the set of solutions by overlaying a regular grid of points. For each point (also known as a *detector*), we check whether or not any feasible lines pass within a specific distance tolerance, which is a function of the the grid's granularity. An example of this procedure is shown in the corresponding Figure 1 (c). Unfortunately, the process of dot-planimeter-style estimation introduces additional approximation error on top of discretizing the fairness region. Our analysis of



(a) Fairness region when relating FPR, FNR, ACC

(b) Fairness region when relating FPR, FNR, PPV

(c) Estimating the size of (b) with a dot planimeter

Figure 1: $p_1 = 0.3, p_2 = 0.5$; $\epsilon_{\text{FPR}}, \epsilon_{\text{FNR}}, \epsilon_{\text{ACC}}, \epsilon_{\text{PPV}} \in [-0.05, 0.05]$

upper-bounding this error (under some assumptions) can be found in Section F of the Appendix.

To avoid this additional approximation error, for our second approach we re-frame our description of the fairness region using a Constraint Program (CP). A constraint program provides an alternative means of measuring how large the space of feasible solutions is for a given setting of tolerances. Rather than measuring the area taken up by a projection on two dimensions (PPV and FNR), we can describe the fairness region directly as the set of *feasible solutions* to a constraint program. Using the CP-SAT solver in Google ORTools, we express our problem's governing equations as a set of integer variables and constraints. Our quantities of interest (FPR, FNR, PPV, p, $\epsilon_{\text{FPR}}, \epsilon_{\text{FNR}}, \epsilon_{\text{PPV}}, \epsilon_{\text{p}}$) are all real numbers rather than integers, with infinitely many possible values in their respective ranges. To characterize the size of the solution space for different tolerances, we discretize the interval $[0, 1]$ into $N + 1$ bins (and, correspondingly, the interval $[-1, 1]$ into $2N + 1$ bins). For example, an FPR $= 0.91$ corresponds to an integer value of 91 when $N = 100$. With this discretization, we represent the fairness region using the following constraint program:

$$
\begin{array}{rll}
\alpha_i, \beta_i, p_i, v_i & \in [0, N] \text{ integer} & \forall i \in \{1, 2\} \\
\epsilon_j & \in [-N, N] \text{ integer} & \forall j \in \{\alpha, \beta, p, v\} \\
m_i, d_i & \in [0, N^2] \text{ integer} & \forall i \in \{1, 2\} \\
n_i & \in [0, N^3] \text{ integer} & \forall i \in \{1, 2\} \\
p_i & = b_i \cdot N & \forall i \in \{1, 2\} \\
j_1 & = j_2 + \epsilon_j & \forall j \in \{\alpha, \beta, p, v\} \\
\epsilon_j & \geq -\epsilon_{\max} \cdot N & \forall j \in \{\alpha, \beta, v\} \\
\epsilon_j & \leq \epsilon_{\max} \cdot N & \forall j \in \{\alpha, \beta, v\} \\
m_i & = p_i \cdot (N - v_i) & \forall i \in \{1, 2\} \\
n_i & = m_i \cdot (N - \beta_i) & \forall i \in \{1, 2\} \\
d_i & = v_i \cdot (N - p_i) & \forall i \in \{1, 2\} \\
n_i & = \alpha_i \cdot d_i & \forall i \in \{1, 2\}
\end{array}
$$

Here, N is the number of integers; n, m, d are intermediate variables used to represent multiplicative constraints for the CP-SAT solver; $\epsilon_{\max}$ represents the maximum allowable value of $|\epsilon_\alpha|, |\epsilon_\beta|, |\epsilon_v|$; b_i represent the observed prevalences in the real-valued range $[0, 1]$; and , α, β, v represent FPR, FNR, PPV, respectively. The CP-SAT solver allows us to enumerate all possible solutions to a constraint program. With N^6 possible values for the set $\{\text{FPR}_1, \text{FPR}_2, \text{FNR}_1, \text{FNR}_2, \text{PPV}_1, \text{PPV}_2\}$, for any fixed N, we can characterize the size of the discretized solution space as a function of changes to the other inputs simply as the number of feasible solutions.

3.3 Revisiting the impossibility theorem

Recall that there are two known exceptions to the impossibility theorem: when the two groups' prevalence values are the same, and under perfect prediction [16, 37]. However, given our formalization for the *relaxed* case of fairness constraints, perhaps we should ask: to what *degree* do the exceptions from the impossibility result apply? Specifically:

(1) How large can prevalence differences be ($\epsilon_{\text{p}} \in \{1\%, 10\%, 50\%\}$) and still imply a large fairness region?
(2) How far can a model depart from perfect prediction (PPV $\in \{99\%, 75\%\}$) and still imply a large fairness region?

3.3.1 Varying prevalence difference. First we explore the impact of varying the prevalence difference between two groups on the size of the fairness region, using the CP described in Section 3.2. The results of these experiments are in Figure 2, which shows heatmaps plotting the number of feasible models for any pair of prevalence values p_1, p_2 over a range of values from 0.01 to 0.99. Figures 2 (a), (b), (c), and (d) correspond to settings where the allowable difference between metrics is $\epsilon \leq 0.0, 0.02, 0.05$, and 0.1, respectively. Note that in each setting we fix performance such that FNR, PPV $\in [0, 0.99]$ to avoid the pathological cases covered by Equation 1.

Several important insights can be gleaned from Figure 2. As expected, in the case where the ϵ-margin-of-error is 0, feasible models are only found on the diagonal, when prevalences are equal (implied by [16, 37]). Interestingly, we observe for all settings of ϵ that the number of feasible models is densest around $\text{p}_1 = \text{p}_2 = 0.5$. For example, the fairness region is larger when $\text{p}_1 = 0.4, \text{p}_2 = 0.5$ than when $\text{p}_1 = 0.1, \text{p}_2 = 0.2$, even though $\epsilon_{\text{p}} = 0.1$ in both cases.

As the ϵ-margin-of-error increases from 0.0 to 0.1, the total number of feasible models increases dramatically from 3,640 to 199,314. While the specific values of these numbers are a function of our discretization and the value of N used in the constraint program, they still enable us to make relative comparisons about the size of the fairness region. For example, Figure 2 (c), where $\epsilon \leq 0.05$ (i.e., the maximum allowable difference between group metrics is 5%),

provides a valuable insight: if the prevalence difference between groups is less than 0.2 (or 20%), the fairness region is quite dense, especially relative to plot 2 (a), where $\epsilon = 0.0$. *This is good news for practitioners* because: (1) prevalence differences between 10% and 15% are commonly observed, and (2) setting $\epsilon \leq 0.05$ is reasonable in many contexts.

3.3.2 Varying performance. Next, we test the effect of "imperfect prediction" on the size of the fairness region. As a reference point, we focus on the case where $\epsilon \leq 0.05$ (Figure 2 (c)), and create bins that corresponded to four ranges of PPV: $[0.00, 0.24]$, $[0.25, 0.49]$, $[0.50, 0.74]$, and $[0.75, 0.99]$. As expected, the closer the setting is to "perfect prediction" (i.e., the higher the PPV), the larger the size of the fairness region. The number of feasible models increases from $7,554$ in the lowest PPV bin to $10,007$ in the highest bin. Notably, there are still many feasible models available in all bins even when the prevalence difference between groups is as high as 20%.

There is another key insight implicit in Figure 3: it not only shows how many feasible models there are under different PPV settings, but that many of those models are *high-performing*. In each figure, it can be seen that the number of feasible models is most dense when the group prevalences are below the maximum PPV value. Recall that any model with a PPV greater than the overall prevalence of the dataset on which it is being used offers value over random chance, suggesting not only many possible models, but many *useful* models in these settings.

3.3.3 Considering intersectional groups. The ultimate goal of algorithmic fairness should not just be ensuring fairness according to multiple metrics, but also for multiple groups (e.g., defined both based on sex, and based on race), including intersections of these groups (e.g., defined by a combination of sex and race) [29]. Intersectional discrimination [20, 43] states that individuals who belong to several protected groups simultaneously (e.g., Black women) experience stronger discrimination compared to individuals who belong to a single protected group (e.g., White women or Black men), and that this disadvantage compounds more than additively. This effect has been demonstrated by numerous case studies, and by theoretical and empirical work [17, 21, 47, 58].

Intersectionaliy is an analytical framework for understanding human beings that considers the outcome of intersections of different social locations, power relations and experiences [29]. For example, an intersectional approach to fairness could be thinking beyond an individual's sex or race, and instead accounting for a set of important characteristics about that individual like their sex, race, ethnicity, and social class. In this paper, we consider a limited interpretation of intersectionality, and investigate how stating fairness constraints with respect to intersections of several sensitive attributes impacts the existence of feasible models. In the following proposition, we show that the maximum prevalence difference across groups defined by an intersection of sensitive attributes (e.g., on sex and race) is at least as high as when groups are defined based on each sensitive attribute independently (e.g., on sex or race).

Proposition 3.4 (Intersectional Prevalence Differences). *Given a dataset that is subdivided into two groups, let $0 < p_1 < n_1$ and $0 < p_2 < n_2$, where p_i is the number of positive class members of group i, and n_i is the total number of members in group i. Suppose $\frac{p_1}{n_1} \leq \frac{p_2}{n_2}$. Then the following holds: $\frac{p_1}{n_1} \leq \frac{p_1+p_2}{n_1+n_2} \leq \frac{p_2}{n_2}$. (Proof deferred to Appendix G).* □

Consider a toy example where there are two binary sensitive attributes: *sex* coded as male and female, and *race* coded as majority and minority. Under the mild assumption of Proposition 3.4, the prevalence of the minority group as a whole must be between the prevalence of the intersectional minority male and minority female groups. The same is true for the prevalence of the majority group. As a result, the prevalence difference between the four intersectional groups (majority male, majority female, minority male, minority female) *must be greater than or equal* to the prevalence difference between only the majority and the minority group. The same reasoning can be used to understand prevalence differences for sex and intersectional sex. The overall implication of Proposition 3.4 is that considering intersectional groups leads to at least equal, but more commonly greater, prevalence differences between groups, which suggests there will be fewer feasible models that are fair with respect to FNR, FPR, and PPV.

3.3.4 Varying the resource constraint k. We now investigate the impact of k on our ability to identify a feasible solution.

Proposition 3.5 (Reducing k increases PPV). *Given a* well-calibrated classifier *being used under a resource constraint k, reducing the size of k will monotonically increase the PPV of the classifier. (Proof deferred to Appendix G)* □

The insight of Proposition 3.5 is that reducing k causes a chain reaction: first, it increases the PPV of the classifier, and second, an increase in PPV results in a more dense space of feasible solutions (as observed in Section 3.3.2). Taken together, this suggests that reducing k can result in a denser space of feasible models on FPR, FNR and PPV.

4 EXPERIMENTS

The analysis in Section 3 shows that, by slightly relaxing fairness constraints between metrics, there are a large number of models satisfying approximate fairness constraints across multiple metrics. In this section, we design an experiment to demonstrate the existence of those models *on real data*.

Our insights suggest that the possibility of finding fair models is influenced by (1) the group prevalences' proximity to 50%, (2) the differences between group prevalences, and (3) the performance of the classifier. To better understand how these parameters impact one's ability to find fair models on real-world data, we developed an experiment to answer the following question: Given a dataset X, a resource constraint k, a set of fairness constraints, and a classifier with a given PPV, does there exist a set of k observations for which (1) fairness constraints for FPR, FNR and PPV are satisfied, and (2) those fairness constraints do not reduce the PPV? If there does exist a set of observations in X that satisfies these requirements, then there also exists a model that could select those observations. Trivially, one can think of a function that uses the index of each element to map to an outcome. In other words, one could use such a set as the labels Y for creating a function $i : X \rightarrow Y$.

To implement the experiment we created a Mixed Integer Linear Program as follows: The objective function is to maximize the PPV of a selection of k observations, subject to 5 constraints: (1)

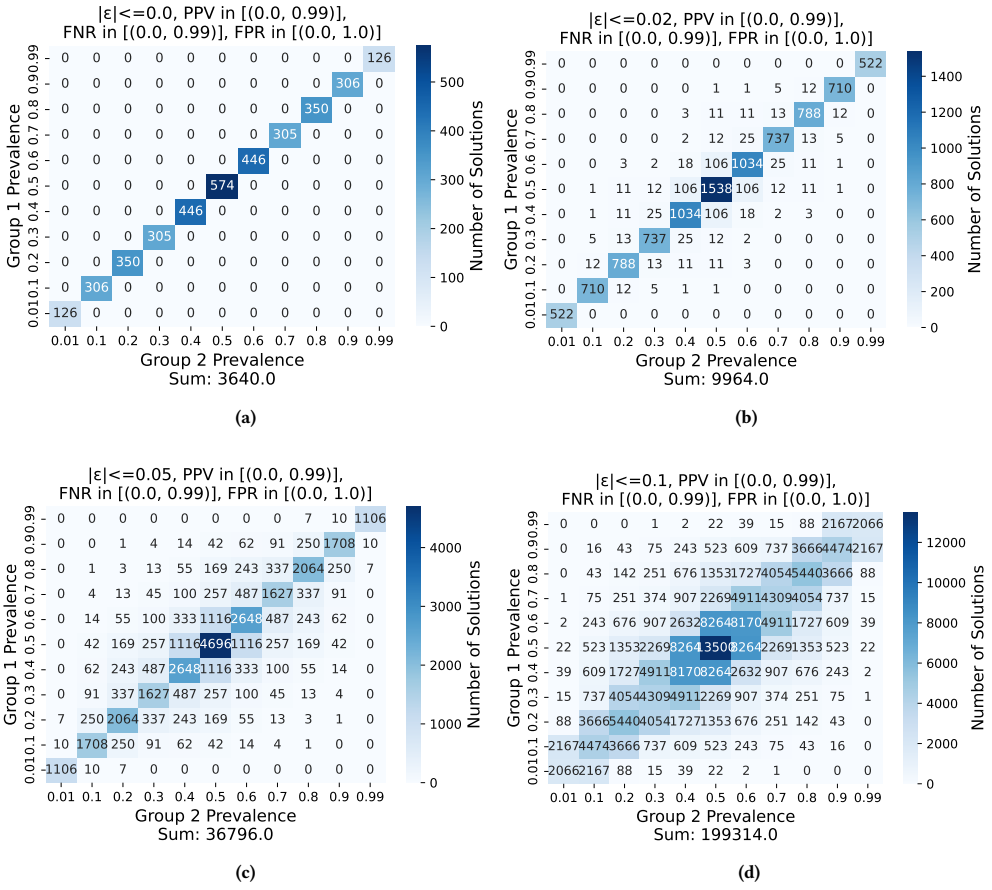


Figure 2: Effect of varying group prevalence values $\mathbf{p}_1, \mathbf{p}_2$ on the number of feasible models for different values of ϵ, where PPV, FNR $\in [0, 0.99]$, FPR $\in [0, 1.0]$, $N = 100$.

k observations must be selected, (2) the classifier has *at most* a pre-defined PPV, (3-5) fairness constraints for FNR, FPR, and PPV are met. The full program details can be found in Appendix A.[4]

One strength of this experiment is that it allows us to work with datasets that may have non-binary sensitive attributes or multiple sensitive attributes. This means we can also explore the feasibility of finding fair sets *when there are more than two groups and consider intersectionality* (see Section 3.3.3).

Note that there are two complications in our problem: First, because PPV cannot be written as a linear constraint [31], we used an existing approach to refactor our problem into an approximately equivalent Quadratic Linear Program that includes constraints for PPV. Second, as a result of the transformations and the inherent complexity, running times become intractable for large datasets. To circumvent this problem, we conducted our experiments on a sample of each dataset stratified on the outcome and sensitive attributes. We ran several sensitivity analyses and did not find any meaningful difference in the experimental results due to down-sampling, but we acknowledge this as a limitation.

4.1 Datasets

We worked with 5 real-world datasets, all with varying outcome prevalences, and representing a variety of sensitive attributes, including sex, race, and education level. We used these datasets in scope of 16 tasks (i.e., outcomes): 8 for *Ukrainian EIE*, 5 for *folktables*, 1 for each of the other 3 datasets).

Ukranian External Independent Evaluation (EIE). [5] EIE data contains standardized tests for secondary school graduates in Ukraine. The 2021 data, used in our experiments, contains 389,322 records. Sensitive attributes include the students' sex and whether they live in an urban or a rural area. Outcomes are students' performance on 5 tests (e.g., *history*, *German*).

Portuguese Student Performance. [6] [19]. This dataset contains the performance of Portuguese high school students in two subjects, mathematics and Portuguese language arts. The dataset contains 1,044 records from students at two high schools, and was collected in 2005 and 2006. Features includes administrative records

[4] All data, code, and experimental results are available in the following GitHub repository: https://github.com/DataResponsibly/the-possibility-of-fairness

[5] https://zno.testportal.com.ua/opendata

[6] https://archive.ics.uci.edu/ml/datasets/student+performance

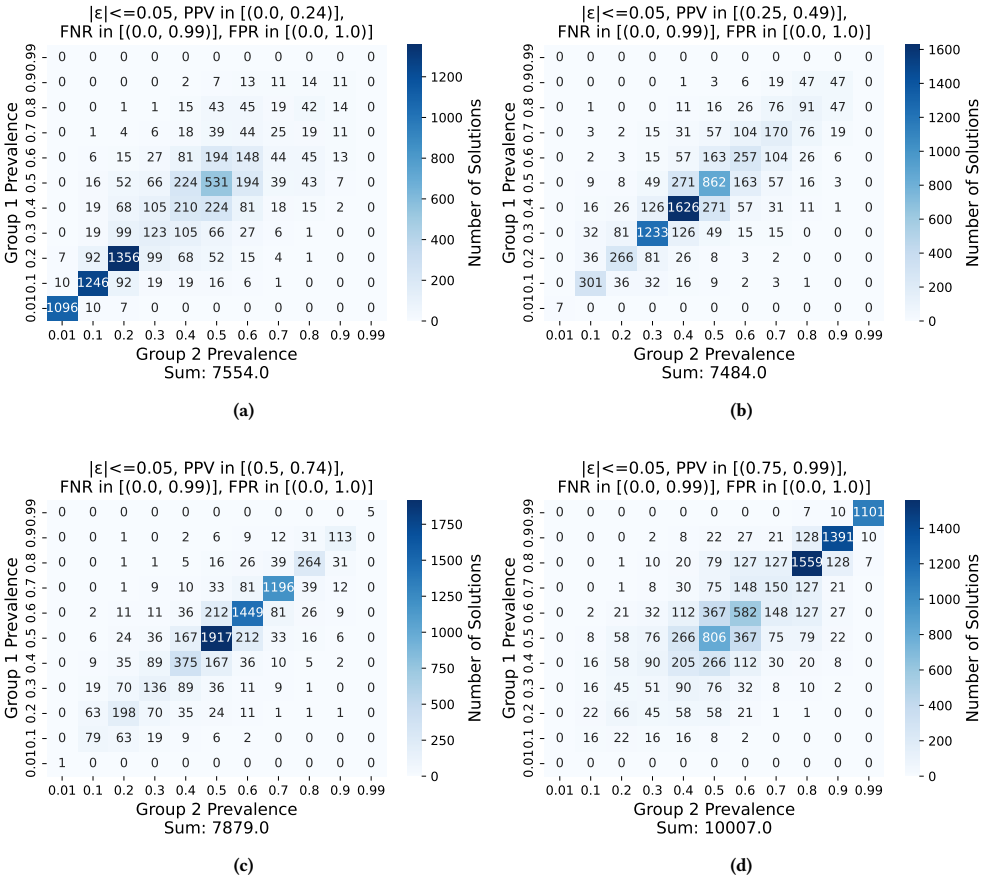


Figure 3: Effect of varying PPV on the fairness region, where $|\epsilon_\alpha|, |\epsilon_\beta|, |\epsilon_v| \leq 0.05$, $N = 100$.

from schools (e.g., grades, number of absences), a lifestyle questionnaire completed by each student (e.g., how many hours per week they study), and sensitive attributes like the student's sex, their parents' education levels, and whether the students live in an urban or rural location. The associated prediction task is identifying students at risk of failure to provide additional school resources.

Taiwanese Loan Assessment. [7] [63]. This dataset contains customer loan data from a bank (and cash issuer) in Taiwan. The data was collected in 2005, has 30,000 records, and includes sensitive attributes like sex and education level. The associated task is to identify customers at risk of defaulting on their loan payments.

Bangladeshi Diabetes Risk Assessment. [8][33]. This dataset, published in 2020, has 520 patient records with information on diabetes-related symptoms, obtained through a questionnaire by the Sylhet Diabetes Hospital in Bangladesh. The task is to identify individuals at risk of early-stage diabetes. Sensitive attributes include age and sex.

Folktables. [9] [22]. This data is from the American Community Survey Public Use Microdata Samples (ACM PUMS), and contains individual-level and household-level data related to income, employment, health, transportation, and housing in the United States. The data is updated yearly, is available at both the national or state level, and contains millions of records. Sensitive attributes include race and sex. We look at 5 separate pre-defined prediction tasks, using data from New York state in 2018.

4.2 Results

Full experimental results are reported in Table 2 in Appendix B, and truncated results are in Table 1. These tables show the "Optimal k Range" for each *(dataset (outcome), sensitive attribute) pair*, where k is expressed as a percentage of the observations. The optimal range shows for which values of k there is a set of observations that satisfy fairness constraints for FPR, FNR, and PPV, without sacrificing any additional classifier PPV. Note that in Table 2, each k list must contain 30% errors (i.e., false positives). This means that the precision of the list k is at-most 70%. This was an arbitrary choice; however, we did conduct extensive sensitivity analysis (see Table 5

[7] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[8] https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.
[9] https://github.com/zykls/folktables

Table 1: Truncated experimental results.

Dataset (Outcome)	Overall Prevalence (%)	Sensitive Attribute	Group Prevalence (%)	Maximum Prevalence Difference (%)	Optimal k Range (% Samples)
EIE (Ukranian)	7.34	Sex	Female: 4.26; Male: 10.66	6.4	None
EIE (Ukranian)	7.34	Territory	Rural: 10.38; Urban: 6.3	4.08	[5,20]
EIE (Math)	31.1	Sex	Female: 30.92; Male: 31.26	0.34	All
EIE (Math)	31.1	Territory	Rural: 39.94; Urban: 28.12	11.82	[5,90]
EIE (Geography)	5.3	Sex	Female: 4.21; Male: 6.31	2.1	All
EIE (Geography)	5.3	Territory	Rural: 6.43; Urban: 4.87	1.56	All
EIE (German)	11.35	Sex	Female: 10.03; Male: 13.87	3.84	[5,90]
EIE (German)	**11.35**	**Territory**	Rural: 26.97; Urban: 8.57	**18.4**	**None**
Folktables (Employment)	46.45	Sex	Female: 44.3; Male: 48.74	4.44	All
Folktables (Employment)	46.45	Race	Asian alone: 50.0; Black or African American alone: 42.08; Other: 41.71; White: 47.36	8.29	All
Folktables (Travel Time)	53.78	Sex	Female: 51.72; Male: 55.8	4.08	All
Folktables (Travel Time)	**53.78**	**Race**	Asian alone: 66.86; Black or African American alone: 69.15; Other: 64.63; White: 48.51	**20.64**	**[5,55]**
Loan Assessment	22.17	Sex	Female: 20.86; Male: 24.16	3.3	All
Diabetes Risk Assessment	61.54	Sex	Female: 90.1; Male: 44.82	45.28	[5,10]
Student Performance	22.03	Sex	Female: 21.13; Male: 23.22	2.09	All
Student Performance	22.03	Parent's education level	High school: 23.55; Not high school or university or greater: 25.85; University or greater: 16.52	9.33	[5,55]

in the Appendix) and found that increasing the PPV generally increases the Optimal k Range, as suggested in Section 3.3.2. Note also that "Maximum Group Difference" refers to the maximum pairwise difference between group prevalence values.

Recall our discussion about fairness constraints over intersectional groups in Section 3.3.3. In Table 3, we have included results for *Folktables* for a new "race-sex" attribute that partitions the data based on a combination of values of these two attributes. It can be seen that the same observations made in Section 4.2 hold here, but with one key difference: in general, the maximum prevalence difference in groups defined by an intersection of attributes (e.g., on sex and race) is at least as high as when groups are defined based on each sensitive attribute independently (e.g., on sex or race). This is consistent with the insights of Proposition 3.4 in Section 3.3.3.

The results in Table 2 support the conceptual findings presented in Section 3. The size of the optimal k range mirrors some previous findings, with evidence of larger k ranges for prevalences closer to 50% and for smaller group prevalence differences. As evidence for the former, for any row in Table 2 for which the optimal k range is *All*, the group prevalence differences are small (less than 10%). For the latter, consider this interesting observation from the table: for the *(EIE (German), Territory)* pair, the maximum group prevalence difference is 18.4% — yet there is no value of k where it is possible to simultaneously satisfy all three fairness constraints. Notice that the overall prevalence is $\approx 10.35\%$. In contrast, the *(Folktables (Travel Time), Race)* pair has a maximum prevalence difference that is even higher at 20.64%, but in this case the group prevalences are closer to 50%, and the optimal k range spans over half the dataset ($[5, 55]$).

Another salient result from our experiments is that out of all 32 combinations of *(dataset (outcome), sensitive attribute)*, only 3 have no k value for which it was possible to find a set of observations satisfying every fairness constraint. This is a promising result: **across five separate and diverse real-world datasets, we demonstrated that there is nearly always at least *some* chance of finding a model that is simultaneously fair with respect to FPR, FNR, and PPV with a small margin-of-error.**

5 DISCUSSION

This paper sought to revisit the impossibility theorem in practical settings. Taken together, our analytical and experimental results, offer a promising perspective on the feasibility of finding models that are fair with respect to approximate constraints for FNR, FPR, and

PPV: **even under slight relaxations, fair models are abundant rather than rare**.

In this section, we present our findings as guidance for practitioners on when it will be feasible to find those fair models. There are several considerations: (1) group prevalence values, (2) prevalence difference between groups, (3) classifier performance, and (4) resource constraint k.

In the first two considerations, our findings suggest that if one allows a small margin-of-error difference between metrics, then there exist many models that simultaneously satisfy parity across FNR, FPR, PPV even when there is a moderate prevalence difference between groups. Further exploration is needed to understand what exactly constitutes *small* and *moderate*, but in our analysis we observed that cases with a 5% margin-of-error and prevalence differences up to 10% (and in some cases up to 20%) afforded feasible solutions. We are unsure how well these particular settings will generalize, but the larger implication is hopeful. For example, revisiting the thought experiment from Section 2: when predicting student graduation in the US where the prevalence difference between minority and majority students is roughly 10%, we expect that *it is possible to build good models that are fair with respect to multiple metrics.* To allow practitioners to answer questions like these for their own datasets, we offer an open-source tool they can use to assess the feasibility of finding models that are fair across multiple constraints, given an input dataset. (See https://github.com/DataResponsibly/the-possibility-of-fairness.)

Regarding considerations (3) and (4), our analytical work suggests that a higher PPV yields a larger number of feasible models. This furthers claims by other researchers that increasing the performance of a model *actually improves* the possibility of finding a fair model [61]. This is in-line with a paradigm shift away from thinking one must choose between high performance *or* fairness — from a fairness perspective, it can be worthwhile to improve the performance of your classifier to further enable fairness across multiple constraints. Connecting this insight to the resource-constrained setting with k, it follows that **resource constraints can, perhaps counter-intuitively, lead to higher chances of finding fair models** (see Proposition 3.5). This is particularly impactful for practitioners working in ML for public policy, where resource constraints can be as small as 1% (k = 0.01) or 5% (k = 0.05) [6, 50].

We also offer two other meta-considerations. The first is the ϵ-margin-of-error allowed between fairness metrics. In practice, ϵ should be decided *a priori*, and by consulting stakeholders and subject area experts [53, 54]—but generally, the guidance here is unsurprising: the larger the tolerable difference between metrics, the larger the feasible region of fair models.

The second meta-consideration is the number of groups of sensitive attributes. We find that adding intersectional groups will increase prevalence differences (see Proposition 3.4), which reduces the number of possibilities for fair models. However, this is by no means an argument against considering intersectionality. On the contrary, we frame this finding as follows: you can continue to add sensitive attributes and intersectional groups and still have a chance of finding models that are fair across multiple metrics.

Future work and limitations. Our work leaves open an important next step in ensuring fairness across multiple metrics and for multiple groups: **once we know there is a large set of feasible models, how do we find such a model? Further, does having a large number of feasible models make it easier to find one of those models?** Answering these questions is beyond the scope of this current paper, but the authors hope to answer them in follow-up work. Notably, as of the time this work was published, there has been at least one effort made to develop an algorithm to find classifiers that are fair with respect to FPR, FNR, and PPV [31]. Significant additional study of this problem should be done with closed-form expressions describing model feasibility in terms of FPR, FNR, and PPV.

6 CONCLUSIONS AND SOCIAL IMPACT

This paper provides evidence that challenges commonly held assumptions about the impossibility theorem in practical settings, suggesting that practitioners can strive for *more* fairness in the algorithms they implement. This is an important part of the social impact of this paper: it exists as part of a growing body of literature showing that strong limits to fairness, like trade-offs with performance, with other metrics, or between groups, may be overstated or even self-imposed. The impossibility theorem is not a rigid barrier to equitable machine learning.

This work also further demonstrates the importance of reducing societal biases, which are ultimately what cause prevalence differences between groups to appear in data. There is a similar implication for designing better models — by knowing that high model performance and fairness are interrelated, we can shift away from a paradigm of wanting to build algorithms that are either better performing *or* more fair, and towards one where we build algorithms that are better performing *and* more fair.

Our efforts, along with related work that challenges commonly-held beliefs about the fairness-accuracy trade-off, may represent an inflection point in the fair-ML community: fewer and fewer researchers and practitioners conform to the idea that we must choose between (a single notion of) fairness and accuracy [14, 31, 50, 61]. Our main take-away is that **achieving fairness along multiple metrics, for multiple groups, and without sacrificing accuracy is much more attainable than previously believed.**

ACKNOWLEDGMENTS

This research was supported in part by NSF Awards No. 1922658 and 1916505, by the NSF Graduate Research Fellowship under Award No. DGE-1839302, and by UL Research Institutes through the Center for Advancing Safety of Machine Intelligence.

REFERENCES

[1] Everaldo Aguiar, Himabindu Lakkaraju, Nasir Bhanpuri, David Miller, Ben Yuhas, and Kecia L Addison. 2015. Who, when, and why: A machine learning approach to prioritizing students at risk of not graduating high school on time. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. 93–102.

[2] Tiago Andrade, Fabricio Oliveira, Silvio Hamacher, and Andrew Eberhard. 2019. Enhancing the normalized multi-parametric disaggregation technique for mixed-integer quadratic programming. *J. of Global Optimization,* 73(4) (2019), 701–722. https://doi.org/10.1007/s10898-018-0728-9

[3] Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22).* Association for Computing Machinery, New York, NY, USA, Article 18, 10 pages. https://doi.org/10.1145/3551624.3555303

[4] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2021. Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics* (2021). https://doi.org/10.1016/j.jfineco.2021.05.047

[5] Andrew Bell, Alexander Rich, Melisande Teng, Tin Orešković, Nuno B Bras, Lénia Mestrinho, Srdan Golubovic, Ivan Pristas, and Leid Zejnilovic. 2019. Proactive advising: a machine learning driven approach to vaccine hesitancy. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–6.

[6] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 248–266.

[7] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. https://arxiv.org/abs/1810.01943

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[9] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[10] Michail Kuzmic Bocarov. 1957. *Matematiko-statisticeskie metody v kartografii.*

[11] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery* 21, 2 (2010), 277–292.

[12] Samuel Carton, Jennifer Helsby, Kenneth Joseph, Ayesha Mahmud, Youngsoo Park, Joe Walsh, Crystal Cody, CPT Estella Patterson, Lauren Haynes, and Rayid Ghani. 2016. Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 67–76.

[13] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[14] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*. 319–328.

[15] A. Charnes and W.W. Cooper. 1962. Programming with linear fractional functionals. *Naval Research Logistics Quarterly* 9(3-4) (1962), 181–186. https://doi.org/10.1002/nav.3800090303

[16] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[17] Patricia Hill Collins. 2002. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

[18] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[19] P. Cortez and A. M. G. Silva. 2008. Using data mining to predict secondary school student performance.

[20] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990), 1241.

[21] Catherine D'Ignazio and Lauren F Klein. 2020. *Data feminism.* MIT Press.

[22] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021), 6478–6490.

[23] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[24] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[25] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[26] YS Frolov and DH Maling. 1969. The accuracy of area measurement by point counting techniques. *The Cartographic Journal* 6, 1 (1969), 21–35.

[27] Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2020. Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (October 1, 2020)* (2020).

[28] LLC Gurobi Optimization. 2022. Gurobi Optimizer Reference Manual. (2022). https://www.gurobi.com/

[29] Olena Hankivsky. 2022. INTERSECTIONALITY 101. (2022).

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

[31] Brian Hsu, Rahul Mazumder, Preetam Nandy, and Kinjal Basu. 2022. Pushing the limits of fairness impossibility: Who's the fairest of them all? *arXiv preprint arXiv:2208.12606* (2022).

[32] Qian Hu and Huzefa Rangwala. 2020. Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. https://eric.ed.gov/?id=ED608050

[33] MM Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. 2020. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer, 113–125.

[34] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, 1–6.

[35] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, Vol. 1. Citeseer.

[36] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[37] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[38] Nikita Kozodoi and Tibor V. Varga. 2021. *Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness R package version 1.2.2.

[39] Danijel Kučak, Vedran Juričić, and Goran Đambić. 2018. MACHINE LEARNING IN EDUCATION-A SURVEY OF CURRENT RESEARCH TRENDS. *Annals of DAAAM & Proceedings* 29 (2018).

[40] Barbara A Langham. 2009. The achievement gap: What early childhood educators need to know. *Texas Child Care Quarterly* (2009), 14–16.

[41] Nicol Turner Lee. 2018. Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society* (2018).

[42] David Lowry-Duda. 2017. On some variants of the Gauss circle problem. *arXiv preprint arXiv:1704.02376* (2017).

[43] Timo Makkonen. 2002. Multiple, compound and intersectional discrimination: Bringing the experiences of the most marginalized to the fore. *Institute for Human Rights, Åbo Akademi University* (2002).

[44] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[45] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163.

[46] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2022. Achieving Fairness via Post-Processing in Web-Scale Recommender Systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 715–725.

[47] Safiya Umoja Noble. 2018. *Algorithms of oppression: How search engines reinforce racism*. nyu Press.

[48] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

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

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

[51] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 142–153.

[52] Lucas Rosenblatt and R. Teal Witter. 2022. Counterfactual Fairness Is Basically Demographic Parity. https://doi.org/10.48550/ARXIV.2208.03843

[53] Boris Ruf and Marcin Detyniecki. 2021. Towards the right kind of fairness in AI. *arXiv preprint arXiv:2102.08453* (2021).

[54] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[55] Pedro Saleiro, Kit T Rodolfa, and Rayid Ghani. 2020. Dealing with bias and fairness in data science systems: A practical hands-on tutorial. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 3513–3514.

[56] Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. 2017. Academic performance prediction in a gender-imbalanced environment.

[57] K Shailaja, B Seetharamulu, and MA Jabbar. 2018. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 910–914.

[58] Stephanie A Shields. 2008. Gender: An intersectionality perspective. *Sex roles* 59, 5-6 (2008), 301–311.

[59] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[60] Anne L. Washington. 2019. How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate.

[61] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8780–8789. https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html

[62] Harrison Wilde, Lucia L. Chen, Austin Nguyen, Zoe Kimpel, Joshua Sidgwick, Adolfo De Unanue, Davide Veronese, Bilal Mateen, Rayid Ghani, Sebastian Vollmer, and et al. 2021. A recommendation and risk classification system for connecting rough sleepers to essential outreach services. *Data and Policy* 3 (2021), e2. https://doi.org/10.1017/dap.2020.23

[63] I-Cheng Yeh and Che-hui Lien. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications* 36, 2 (2009), 2473–2480.

[64] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[65] Leid Zejnilovic, Susana Lavado, Carlos Soares, Íñigo Martínez De Rituerto De Troya, Andrew Bell, and Rayid Ghani. 2021. Machine Learning Informed Decision-Making with Interpreted Model's Outputs: A Field Intervention. In *Academy of Management Proceedings*, Vol. 2021. Academy of Management Briarcliff Manor, NY 10510, 15424.

A QUADRATIC MIXED INTEGER LINEAR PROGRAM DETAILS

In this section, we present the Mixed Integer Linear Program used in our experiment in Section 4. The objective function is to maximize the PPV of a selection of k observations, subject to 5 constraints: (1) k observations must be selected, (2) the classifier has *at most* a predefined PPV, (3-5) fairness constraints for FNR, FPR, and PPV are met. All data, code, and experimental results are available in a GitHub repository at https://github.com/DataResponsibly/the-possibility-of-fairness. The repository will be made public upon publication.

Note that the fairness constraints are enforced using *disparity ratios* where $0.8 \leq disparity \leq 1.2$, rather than a $\pm\epsilon$ distance between group metrics. We made this decision because disparity ratios are more robust to small real values. For example, consider a model that has $\text{FPR}_1 = 0.002$ for one group and $\text{FPR}_2 = 0.04$ for the other. If $\epsilon = 0.05$, technically these values would satisfy a fairness constraint where $|\text{FPR}_1 - \text{FPR}_2| <= \epsilon$, but this is likely not desirable to practitioners. However, using disparity ratios, this scenario would be considered unfair since $0.8 \nleq \frac{\text{FPR}_1}{\text{FPR}_2} \leq 1.2$.

We frame the Mixed Integer Linear Program in Figure 4.

where N is the set of entities, $n = |N|$; x is a binary array of length n where entry x_i indicates whether or not entity $i \in N$ is included in the final list; l is a binary array of length n such that $l_i = 1$ if the outcome entity $x_i, i \in N$ is 1 and 0 otherwise; G is a set of protected groups; g_j is a binary array of length n where entry g_{j_i} indicates whether or not entry entity $i \in N$ is in the group j (note that group g_{ref} is the reference group for disparity calculations); k is the final list size; ub and lb are the upper and lower bounds for the disparity ratios, respectively.

KLS is the "k-list-size" constraint, FPRU and FPRL are the upper and lower bounds for the False Positive Rate, respectively, FNRU and FNRL are the upper and lower bounds for the False Negative Rate, respectively, and PPVU and PPVL are the upper and lower bounds for the PPV respectively.

The PPV constraints (PPVU and PPVL) halt the problem from being the Mixed Integer Linear Programming problem (MILP). We were inspired by an approach that was used when faced with a similar obstacle in creating MFOpt (Multiple Fairness Optimization Framework) [31], and propose a reformulation of the MIP in a way where we can apply the normalized multiparametric disaggregation technique (NMDT [2]). We go through the following four steps:

Step 1. Make the following substitution:

$$P_{g_j} = \sum_{i=1}^{n} x_i \cdot g_{j_i} \quad \text{and} \quad T_{g_j} = \frac{\sum_{i=1}^{n} x_i \cdot l_i \cdot g_{j_i}}{\sum_{i=1}^{n} x_i \cdot g_{j_i}}$$

Note that P_{g_j} is the number of entities from the group g_j that are included in the final list, and T_{g_j} is the PPV for the group g_j.

Step 2. Find the upper PU_{g_j} and the lower PL_{g_j} bounds for each P_{g_j} by solving the following MILP:

$$\begin{array}{lll} \text{maximize}_{x_i} & P_{g_j} & \\ \text{subject to} & \text{(KLS)} & \\ & \text{(FPRU),} & \text{(FPRL)} \\ & \text{(FNRU),} & \text{(FNRL)} \end{array}$$

Step 3. Find the upper TU_{g_j} and the lower TL_{g_j} bounds for each T_{g_j} by solving the following MIP with linear constraints but with the fractional objective:

$$\begin{array}{lll} \text{maximize}_{x_i} & T_{g_j} & \\ \text{subject to} & \text{(KLS)} & \\ & \text{(FPRU),} & \text{(FPRL)} \\ & \text{(FNRU),} & \text{(FNRL)} \end{array}$$

To solve we use the Charnes-Cooper transformation [15].

Step 4. Reformulate the initial optimization problem in terms of P_{g_j} and T_{g_j} with corresponding lower and upper bounds from steps 2 and 3 in order to use NMDT transformation, so that it can be easily handled by MIP solver [28]. Also, note that all denominators there are constant for the given dataset. This reformulation is shown in Figure 5.

Note that P_{g_j} are integer variables, and T_{g_j} are continuous variables, but with bounds found in step 3 ($T_{g_j} \in [TL_{g_j}, TU_{g_j}]$), and precision factor p as a negative integer, we can represent this continuous variable exactly as

$$T_{g_j} = (TU_{g_j} - TL_{g_j}) \cdot \lambda + TL_{g_j}$$

where

$$\lambda = \sum_{m \in \{-p, \ldots, -1\}} 2^m \cdot z_m$$

and $z_m \in \{0, 1\}$ are binary optimization variables.

B EXPERIMENTAL RESULTS

Our full experimental results can be found in the Python notebooks in the "experiments" folder of our Github repository at https://github.com/DataResponsibly/the-possibility-of-fairness. In this section, we wanted to include examples of the output of our experiment for each *Dataset (Outcome), sensitive attribute* pair. Here we highlight two such pairs from the EIE dataset, which can be seen in Figure 6. Plot (a) shows results for the *EIE (Geography), territory* pair, and plot (b) shows the results for the *EIE (Ukrainian), territory* pair. Each of those plots contains two subplots. On top, it shows the PPV (Precision) and Recall (dotted lines) of an unconstrained linear program that has the specified PPV. The solid lines show the PPV and Recall of the selected sets. The bottom plot shows the disparity of each metric, where the dashed lines show the limits of 1.2 and 0.8. We can tell when a model is no longer optimal when the constrained PPV and Recall meaningfully deviate from the unconstrained PPV and Recall. Note that in some experiments, the PPV, FPR, and FNR disparities may be outside of the disparity window ($[0.8, 1.2]$) — but these instances are either pathological or due to a rounding error. The pathological cases occur when there is only one or two False Positives or False Negatives in a group.

- Full experimental results: Table 2
- Intersectional results: Table 3
- Sample size sensitivity analysis: Table 4
- PPV sensitivity analysis: Table 5

$$
\begin{aligned}
\text{maximize}_{x_i} \quad & \sum_{i=1}^{n} x_i \cdot l_i \\
\text{subject to} \quad & \sum_{i=1}^{n} x_i = k, && && \text{(KLS)} \\
& \frac{\sum_{i=1}^{n} x_i \cdot (1-l_i) \cdot g_{j_i}}{\sum_{i=1}^{n} (1-l_i) \cdot g_{j_i}} \leq ub \cdot \frac{\sum_{i=1}^{n} x_i \cdot (1-l_i) \cdot g_{ref_i}}{\sum_{i=1}^{n} (1-l_i) \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FPRU)} \\
& \frac{\sum_{i=1}^{n} x_i \cdot (1-l_i) \cdot g_{j_i}}{\sum_{i=1}^{n} (1-l_i) \cdot g_{j_i}} \geq lb \cdot \frac{\sum_{i=1}^{n} x_i \cdot (1-l_i) \cdot g_{ref_i}}{\sum_{i=1}^{n} (1-l_i) \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FPRL)} \\
& \frac{\sum_{i=1}^{n} (1-x_i) \cdot l_i \cdot g_{j_i}}{\sum_{i=1}^{n} l_i \cdot g_{j_i}} \leq ub \cdot \frac{\sum_{i=1}^{n} (1-x_i) \cdot l_i \cdot g_{ref_i}}{\sum_{i=1}^{n} l_i \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FNRU)} \\
& \frac{\sum_{i=1}^{n} (1-x_i) \cdot l_i \cdot g_{j_i}}{\sum_{i=1}^{n} l_i \cdot g_{j_i}} \geq lb \cdot \frac{\sum_{i=1}^{n} (1-x_i) \cdot l_i \cdot g_{ref_i}}{\sum_{i=1}^{n} l_i \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FNRL)} \\
& \frac{\sum_{i=1}^{n} x_i \cdot l_i \cdot g_{j_i}}{\sum_{i=1}^{n} x_i \cdot g_{j_i}} \leq ub \cdot \frac{\sum_{i=1}^{n} x_i \cdot l_i \cdot g_{ref_i}}{\sum_{i=1}^{n} x_i \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(PPVU)} \\
& \frac{\sum_{i=1}^{n} x_i \cdot l_i \cdot g_{j_i}}{\sum_{i=1}^{n} x_i \cdot g_{j_i}} \geq lb \cdot \frac{\sum_{i=1}^{n} x_i \cdot l_i \cdot g_{ref_i}}{\sum_{i=1}^{n} x_i \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(PPVL)} \\
& && x_i, l_i \in \{0,1\}, && i = 1, \ldots, n \\
& && g_{j_i} \in \{0,1\}, && j \in G \\
& && && \text{(4)}
\end{aligned}
$$

Figure 4: Mixed Integer Linear Program.

$$
\begin{aligned}
\text{maximize} \quad & \sum_{j=1}^{n} T_{g_j} \cdot P_{g_j} \\
\text{subject to} \quad & \sum_{j=1}^{n} P_{g_j} = k, && && \text{(KLS)} \\
& \frac{P_{g_j} \cdot (1 - T_{g_j})}{\sum_{i=1}^{n} (1 - l_i) \cdot g_{j_i}} \leq ub \cdot \frac{P_{g_{ref}} \cdot (1 - T_{g_{ref}})}{\sum_{i=1}^{n} (1 - l_i) \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FPRU)} \\
& \frac{P_{g_j} \cdot (1 - T_{g_j})}{\sum_{i=1}^{n} (1 - l_i) \cdot g_{j_i}} \geq lb \cdot \frac{P_{g_{ref}} \cdot (1 - T_{g_{ref}})}{\sum_{i=1}^{n} (1 - l_i) \cdot g_{ref_i}}, && \forall j \in G, j \neq ref && \text{(FPRL)} \\
& 1 - \frac{P_{g_j} \cdot T_{g_j}}{\sum_{i=1}^{n} l_i \cdot g_{j_i}} \leq ub \cdot \left(1 - \frac{P_{g_{ref}} \cdot T_{g_{ref}}}{\sum_{i=1}^{n} l_i \cdot g_{ref_i}} \right), && \forall j \in G, j \neq ref && \text{(FNRU)} \\
& 1 - \frac{P_{g_j} \cdot T_{g_j}}{\sum_{i=1}^{n} l_i \cdot g_{j_i}} \geq lb \cdot \left(1 - \frac{P_{g_{ref}} \cdot T_{g_{ref}}}{\sum_{i=1}^{n} l_i \cdot g_{ref_i}} \right), && \forall j \in G, j \neq ref && \text{(FNRL)} \\
& T_{g_j} \leq ub \cdot T_{g_{ref}} && \forall j \in G, j \neq ref && \text{(PPVU)} \\
& T_{g_j} \leq lb \cdot T_{g_{ref}} && \forall j \in G, j \neq ref && \text{(PPVL)} \\
& P_j \in \{PL_{g_j}, \ldots, PU_{g_j}\}, && j \in G \\
& T_j \in [TL_{g_j}, TU_{g_j}], && j \in G \\
& l_i \in \{0, 1\}, && i = 1, \ldots, n \\
& g_{j_i} \in \{0, 1\}, && j \in G
\end{aligned}
$$

Figure 5: Reformulated initial optimization problem in terms of P_{g_j} and T_{g_j} with corresponding lower and upper bounds from steps 2 and 3 in order to use NMDT transformation, so that it can be easily handled by MIP solver [28].

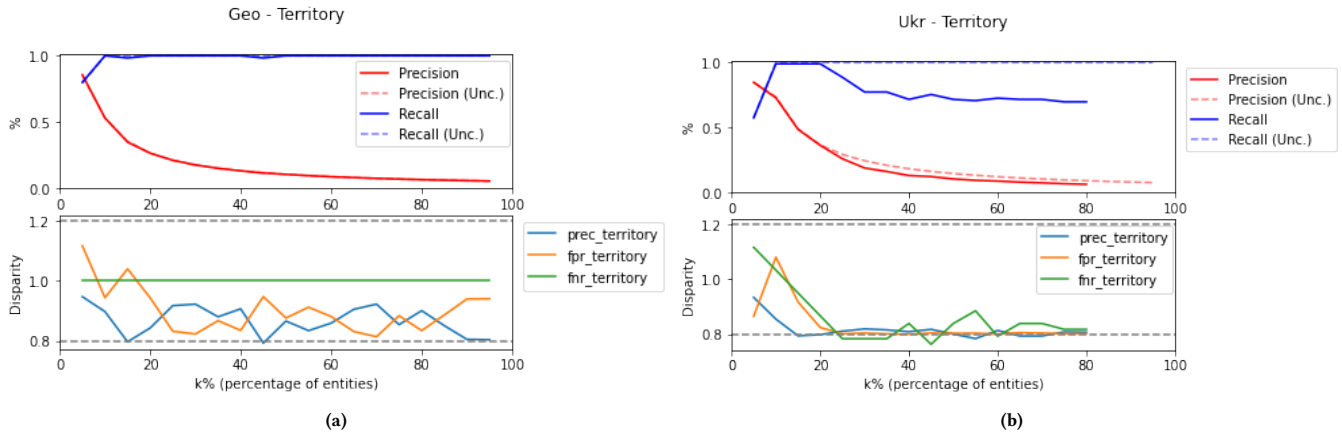


Figure 6: Plot (a) shows that overall values of k, PPV, and Recall of the selected set do not substantially deviate from an unconstrained model, and that PPV, FPR, and FNR remain within the bounds of the disparity window for all values. Plot (b) shows that over the k range of $[5, 20]$, PPV and Recall of the identified sets are inline with unconstrained Precision and Recall.

Table 2: Full empirical results

Dataset (Outcome)	Overall Prevalence (%)	Sensitive Attribute	Group Prevalence (%)	Maximum Prevalence Difference (%)	Optimal k Range (% Samples)
EIE (Ukranian)	7.34	Sex	Female: 4.26; Male: 10.66	6.4	None
EIE (Ukranian)	7.34	Territory	Rural: 10.38; Urban: 6.3	4.08	[5,20]
EIE (History)	18	Sex	Female: 14.48; Male: 22.2	7.72	[5,60]
EIE (History)	18	Territory	Rural: 20.27; Urban: 17.05	3.22	All
EIE (Math)	31.1	Sex	Female: 30.92; Male: 31.26	0.34	All
EIE (Math)	31.1	Territory	Rural: 39.94; Urban: 28.12	11.82	[5,90]
EIE (Physics)	8.33	Sex	Female: 10.46; Male: 7.96	2.5	All
EIE (Physics)	8.33	Territory	Rural: 12.16; Urban: 7.09	5.07	[5,20]
EIE (Chemistry)	10.68	Sex	Female: 11.18; Male: 9.81	1.37	All
EIE (Chemistry)	10.68	Territory	Rural: 15.72; Urban: 9.24	6.48	[5,25]
EIE (Geography)	5.3	Sex	Female: 4.21; Male: 6.31	2.1	All
EIE (Geography)	5.3	Territory	Rural: 6.43; Urban: 4.87	1.56	All
EIE (English)	10.64	Sex	Female: 9.55; Male: 11.76	2.21	All
EIE (English)	10.64	Territory	Rural: 17.35; Urban: 9.43	7.92	[5,20]
EIE (German)	11.35	Sex	Female: 10.03; Male: 13.87	3.84	[5,90]
EIE (German)	11.35	Territory	Rural: 26.97; Urban: 8.57	18.4	None
Folktables (Employment)	46.45	Sex	Female: 44.3; Male: 48.74	4.44	All
Folktables (Employment)	46.45	Race	Asian alone: 50.0; Black or African American alone: 42.08; Other: 41.71; White: 47.36	8.29	All
Folktables (Income)	41.51	Sex	Female: 35.76; Male: 47.13	11.37	All
Folktables (Income)	41.51	Race	Asian alone: 41.05; Black or African American alone: 31.9; Other: 25.3; White: 44.91	19.61	[5,50]
Folktables (Medical Cover)	40.09	Sex	Female: 38.87; Male: 41.65	2.78	All
Folktables (Medical Cover)	40.09	Race	Asian alone: 41.13; Black or African American alone: 52.43; Other: 49.06; White: 35.22	17.21	[5,60]
Folktables (Mobility)	78.17	Sex	Female: 77.29; Male: 79.07	1.78	[5,70]
Folktables (Mobility)	78.17	Race	Asian alone: 75.57; Black or African American alone: 81.33; Other: 81.68; White: 77.35	6.11	[5,60]
Folktables (Travel Time)	53.78	Sex	Female: 51.72; Male: 55.8	4.08	All
Folktables (Travel Time)	53.78	Race	Asian alone: 66.86; Black or African American alone: 69.15; Other: 64.63; White: 48.51	20.64	[5,55]
Loan Assessment	22.17	Education level	High school: 25.38; Not high school or university or greater: 10.53; University or greater: 21.75	14.85	None
Loan Assessment	22.17	Sex	Female: 20.86; Male: 24.16	3.3	All
Diabetes Risk Assessment	61.54	Sex	Female: 90.1; Male: 44.82	45.28	[5,10]
Student Performance	22.03	Sex	Female: 21.13; Male: 23.22	2.09	All
Student Performance	22.03	Address	Rural: 27.27; Urban: 20.05	7.22	All
Student Performance	22.03	Parent's education level	High school: 23.55; Not high school or university or greater: 25.85; University or greater: 16.52	9.33	[5,55]

Table 3: Experimental results for Folktables with intersectionality

Dataset (Outcome)	Number of Samples	Overall Prevalence (%)	Sensitive Attribute	Group Prevalence (%)	Maximum Group Prevalence Difference (%)	Optimal k Range (% Samples)
Folktables (Employment)	1970	46.45	Race and sex	Asian alone, Female: 46.07; Asian alone, Male: 54.32; Black or African American alone, Female: 44.19; Black or African American alone, Male: 39.64; Other; Female: 39.56; Other; Male: 44.05; White, Female: 44.71; White, Male: 50.15	14.76	All
Folktables (Income)	1031	41.51	Race and sex	Asian alone, Female: 39.13; Asian alone, Male: 42.86; Black or African American alone, Female: 30.77; Black or African American alone, Male: 33.33; Other; Female: 21.95; Other; Male: 28.57; White, Female: 37.82; White, Male: 51.58	29.63	[5,35]
Folktables (Public Medical Coverage)	1352	40.09	Race and sex	Asian alone, Female: 40.24; Asian alone, Male: 42.37; Black or African American alone, Female: 53.77; Black or African American alone, Male: 51.0; Other; Female: 51.14; Other; Male: 46.48; White, Female: 33.13; White, Male: 38.02	20.64	All
Folktables (Mobility)	1214	78.17	Race and sex	Asian alone, Female: 74.63; Asian alone, Male: 76.56; Black or African American alone, Female: 81.18; Black or African American alone, Male: 81.48; Other; Female: 81.82; Other; Male: 81.54; White, Female: 76.14; White, Male: 78.57	7.19	[5,60]
Folktables (Travel Time)	1824	53.78	Race and sex	Asian alone, Female: 65.85; Asian alone, Male: 67.82; Black or African American alone, Female: 69.3; Black or African American alone, Male: 68.97; Other; Female: 63.01; Other; Male: 66.22; White, Female: 45.41; White, Male: 51.41	23.89	[5,80]

Table 4: Sample size sensitivity analysis

Dataset (Outcome)	Number of Samples	Overall Prevalence (%)	Sensitive Attribute	Group Prevalence (%)	Maximum Group Prevalence Difference (%)	Optimal k Range (% Samples)
EIE (Ukranian)	1445	7.34	Sex	Female: 4.26; Male: 10.66	6.4	None
EIE (Ukranian)	1445	7.34	Territory	Rural: 10.38; Urban: 6.3	4.08	[5,20]
EIE (History)	1994	18	Sex	Female: 14.48; Male: 22.2	7.72	[5,60]
EIE (History)	1994	18	Territory	Rural: 20.27; Urban: 17.05	3.22	All
EIE (Math)	1222	31.1	Sex	Female: 30.92; Male: 31.26	0.34	All
EIE (Math)	1222	31.1	Territory	Rural: 39.94; Urban: 28.12	11.82	[5,90]
EIE (Ukranian)	5777	7.34	Sex	Female: 4.26; Male: 10.67	6.41	None
EIE (Ukranian)	5777	7.34	Territory	Rural: 10.38; Urban: 6.31	4.07	[5,20]
EIE (History)	5982	17.99	Sex	Female: 14.43; Male: 22.23	7.8	[5,60]
EIE (History)	5982	17.99	Territory	Rural: 20.21; Urban: 17.05	3.16	All
EIE (Math)	7327	31.05	Sex	Female: 30.91; Male: 31.18	0.27	All
EIE (Math)	7327	31.05	Territory	Rural: 39.74; Urban: 28.13	11.61	[5,90]
EIE (Ukranian)	51991	7.34	Sex	Female: 4.26; Male: 10.68	6.42	None
EIE (Ukranian)	51991	7.34	Territory	Rural: 10.37; Urban: 6.31	4.06	[5,20]
EIE (History)	51837	17.98	Sex	Female: 14.43; Male: 22.2	7.77	[5,60]
EIE (History)	51837	17.98	Territory	Rural: 20.21; Urban: 17.03	3.18	All
EIE (Math)	51283	31.05	Sex	Female: 30.92; Male: 31.18	0.26	All
EIE (Math)	51283	31.05	Territory	Rural: 39.75; Urban: 28.13	11.62	[5,90]

C LIST OF ALL METRICS AND THEIR EQUATIONS

Metrics. A traditional confusion matrix is a standard tool for understanding binary classification tasks, and details potential model outcomes, giving names to the relation between what the model predicts and what the ground truth labels actually are - see Table 6. From the confusion matrix comes a set of standard metrics that capture relationships in the outcomes of a binary classification model. Here we detail potentially relevant metrics to this paper.

- $TPR = \frac{TP}{TP+FN}$
- $TNR = \frac{TN}{TN+FP}$
- $FPR = \frac{FP}{FP+TN}$
- $FNR = \frac{FN}{FN+TP}$
- $PPV = \frac{TP}{TP+FP}$
- $\mathsf{ACC} = \frac{TP+TN}{TP+FP+TN+FN}$

Binary Classification. We define binary classification as follows. Consider dataset X consisting of observations for individuals $x_1, x_2, \ldots, x_n$. For individual $x \in X$ with a target value of interest (or "label") $y \in Y$, where $Y \in \{0,1\}^n$, we seek to *classify* x correctly (i.e. assign label y to x). More formally, we attempt to learn a function $\hat{Y}$ such that $\forall (x_i, y_i) \in X \cup Y$, $\hat{Y}(x_i) \rightarrow \hat{y_i}$ and $\hat{y_i} == y_i$.

Often, we utilize a statistical technique to find a function $\hat{Y}$ that produces a real valued score $\tilde{y} \in (0,1)$, and to binarize the outputs we apply a standard thresholding function $\tau(\tilde{y}) = \hat{y}$ to produce a binary label $\hat{y} \in \{0,1\}$. For example, if $\tilde{y}$ is interpreted as a *probability* of 1 being the correct label (a "positive" assignment), then τ thresholding at a greater than $\frac{1}{2}$ probability of positive class assignment is a way to convert from score or probability to concrete class.

$$\tau(\tilde{y}) = \left\{ \begin{array}{ll} 1, & \text{for } \frac{1}{2} \leq \tilde{y} \leq 1 \\ 0, & \text{otherwise} \end{array} \right\}$$

However, in situations with a resource constraint k that governs how many positive labels we are allowed to assign (say, in a college admissions scenario), we may be forced to adjust $\tau(\tilde{y})$ to accept a function of k, i.e., $f(k) = t$ where $t \in [0,1]$ such that $\tau(\tilde{y}, f(k))$ produces exactly k positive classifications. For example:

$$\tau(\tilde{y}, f(k)) = \left\{ \begin{array}{ll} 1, & \text{for } f(k) \leq \tilde{y} \leq 1 \\ 0, & \text{otherwise} \end{array} \right\} s.t. \sum_{i=1}^{n} \mathbf{1}(\hat{y}_i = 1) = k$$

Fairness Considerations: Binary Classification with Sensitive Features. Often, when considering the algorithmic fairness of a binary classifier, we consider a *sensitive* or *protected* attribute in the data that denotes *group membership*. For example, many datasets collected in social settings have information about the *race* or *gender* of individuals in the population. Both of these attributes are and should be "protected," morally and lawfully. Thus, when we evaluate our binary classifier (say, along FPR or FNR), we can evaluate each metric for the entire population, and we can also evaluate each metric *conditioned on group membership*. In the simplest case (which is our focus for much of this paper), our *sensitive* attribute is binary, and thus we consider FPR_1 and FPR_2, FNR_1 and FNR_2, etc. (metrics evaluated on the disjoint sets of outcomes based on group conditioning).

D ANALYTICAL APPROACH TO CHARACTERIZING THE FAIRNESS REGION USING FPR, FNR, ACC

Corollary D.1 (Impossibility Result Variation [16] [37]). *Consider a binary classification setting. The relationship between ACC, FNR, FPR and p can be characterized by:*

$$ACC = (1 - FNR)p + (1 - FPR)(1 - p) \tag{5}$$

Proof. Consider the following statements over accuracy, and note that they apply overall as well as for for some Group i. Thus, each of these quantities could be subscripted with i i.e. $ACC = ACC_i$, etc.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$FNR = 1 - \frac{TP}{TP + FN} \tag{7}$$

$$FPR = 1 - \frac{TN}{TN + FP} \tag{8}$$

$$p = \frac{TP + FN}{TP + TN + FP + FN} \tag{9}$$

$$1 - p = \frac{TN + FP}{TP + TN + FP + FN} \tag{10}$$

Split the numerator for ACC, and multiply by a clever 1:

$$= \frac{TP}{TP + TN + FP + FN} + \frac{TN}{TP + TN + FP + FN} \tag{11}$$

$$= \frac{TP + FN}{TP + FN} \times \frac{TP}{TP + TN + FP + FN} + \frac{TN + FP}{TN + FP} \times \frac{TN}{TP + TN + FP + FN} \tag{12}$$

$$= \frac{TP}{TP + FN} \times \frac{TP + FN}{TP + TN + FP + FN} + \frac{TN}{TN + FP} \times \frac{TN + FP}{TP + TN + FP + FN} \tag{13}$$

$$= (1 - FNR)p + (1 - FPR)(1 - p) \tag{14}$$

□

Lemma D.2 (Expressing Fairness Area Variation). *Consider Corrolary D.1. Assume that $p_2 = p_1 + \epsilon_p$, $ACC_2 = ACC_1 + \epsilon_{ACC}$, $FPR_2 = FPR_1 + \epsilon_{FPR}$, and $FNR_2 = FNR_1 + \epsilon_{FNR}$, where each $\epsilon_{FPR}, \epsilon_{ACC}, \epsilon_{FNR}, \epsilon_p \in (-1, 1)$ term captures the difference between two groups for FPR, ACC, FNR, p respectively. Then the following equality holds:*

$$FNR = \frac{-\epsilon_{FPR} + \epsilon_{ACC} + \epsilon_{FPR}\, p - \epsilon_{FNR}\, p + \alpha\epsilon_p + \epsilon_{FPR}\, \epsilon_p - \epsilon_{FNR}\, \epsilon_p}{\epsilon_P} \tag{15}$$

Proof. Consider *Corollary* D.1 in the setting where there are two groups. Suppose $ACC_1 = ACC_2$. Then:

$$(1 - FNR_1)p_1 + (1 - FPR_1)(1 - p_1) = (1 - FNR_2)p_2 + (1 - FPR_2)(1 - p_2) \tag{16}$$

Make the same substitutions as in Lemma D.2 to find:

$$\begin{aligned} (1 - \mathsf{FNR})p + (1 - \mathsf{FPR})(1 - p) = &(1 - (\mathsf{FNR} + \epsilon_{\mathsf{FNR}})(p + \epsilon_p)) + \\ &(1 - (\mathsf{FPR} + \epsilon_{\mathsf{FPR}}))(1 - (p + \epsilon_p)) \\ &+ \epsilon_{\mathsf{ACC}} \end{aligned} \tag{17}$$

Solving for FNR yields (15). □

Table 5: PPV sensitivity analysis

Dataset (Outcome)	Overall Prevalence (%)	Sensitive Attribute	Maximum Group Prevalence Difference (%)	Optimal k Range (PPV = 0.7)	Optimal k Range (PPV = 0.85)
EIE (Ukranian)	7.34	Sex	6.4	None	None
EIE (Ukranian)	7.34	Territory	4.08	[5,20]	[5,20]
EIE (History)	18	Sex	7.72	[5,60]	[5,60]
EIE (History)	18	Territory	3.22	All	All
EIE (Math)	31.1	Sex	0.34	All	All
EIE (Math)	31.1	Territory	11.82	[5,90]	[5,90]
EIE (Physics)	8.33	Sex	2.5	All	[5,90]
EIE (Physics)	8.33	Territory	5.07	[5,20]	[5,20]
EIE (Chemistry)	10.68	Sex	1.37	All	All
EIE (Chemistry)	10.68	Territory	6.48	[5,25]	[5,25]
EIE (Geography)	5.3	Sex	2.1	All	[5,90]
EIE (Geography)	5.3	Territory	1.56	All	All
EIE (English)	10.64	Sex	2.21	All	All
EIE (English)	10.64	Territory	7.92	[5,20]	[5,20]
EIE (German)	11.35	Sex	3.84	[5,90]	[5,90]
EIE (German)	11.35	Territory	18.4	None	[5,10]
Folktables (Employment)	46.45	Sex	4.44	All	All
Folktables (Employment)	46.45	Race	8.29	All	All
Folktables (Income)	41.51	Sex	11.37	All	All
Folktables (Income)	41.51	Race	19.61	[5,50]	[5,55]
Folktables (Public Medical Coverage)	40.09	Sex	2.78	All	All
Folktables (Public Medical Coverage)	40.09	Race	17.21	[5,60]	[5,60]
Folktables (Mobility)	78.17	Sex	1.78	[5,70]	All
Folktables (Mobility)	78.17	Race	6.11	[5,60]	All
Folktables (Travel Time)	53.78	Sex	4.08	All	All
Folktables (Travel Time)	53.78	Race	20.64	[5,55]	[5,70]
Loan Assessment	22.17	Education level	14.85	None	None
Loan Assessment	22.17	Sex	3.3	All	All
Diabetes Risk Assessment	61.54	Sex	45.28	[5,10]	[5,10]
Student Performance	22.03	Sex	2.09	All	All
Student Performance	22.03	Address	7.22	All	All
Student Performance	22.03	Parent's education level	9.33	[5,55]	[5,55]

Table 6: Standard confusion matrix.

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	TP	FP
	Negative (0)	FN	TN

Lemma D.3 (Closed-Form for Fairness Area Variation). *Assume $\epsilon_p < 1 - p$. Allow $\pm\gamma$ to be the symmetric acceptable error (our "fair" relaxation) between groups for metrics FPR, FNR and ACC. Consider the size of the space of possible $\epsilon_{FPR}, \epsilon_{FNR}, \epsilon_{ACC}$ assignments, given ϵ_p and p, that satisfy the constraints from Lemma D.2. We will denote the size of that space as $|A_f|$ (as a shorthand, we will call that the "fairness region", but the reality is more nuanced). For a set of fairness constraints $\epsilon_{FPR}, \epsilon_{FNR}, \epsilon_{ACC} \in (-\gamma, \gamma)$, where $|\gamma| \leq 1$ and $\gamma \neq 0$, we have that $|A_f|$ is simply:*

$$|A_f| = \frac{4\gamma}{\epsilon_p} - \frac{4\gamma^2}{{\epsilon_p}^2} \tag{18}$$

Before proving Lemma D.3, let's briefly motivate the three primary assumptions: first, we should expect $\epsilon_p << p$, as our relaxation constant should not really be on the same order of magnitude as our per-group prevalence (think $p \approx 0.5$ and $\epsilon_p \approx 0.05$). Thus, the assumption that $\epsilon_p < 1 - p$ is very reasonable.

Second, pre-specifying an acceptable γ relaxation term may seem odd, but it is very common among practitioners, who prefer small groups variations to large ones. Thus, think of γ as a small value, something like $\gamma \leq 0.05$.

Third, assuming that $|\gamma| > 0$ is necessary, as when $|\gamma| = 0$ we recover Corollary D.1. We also ignore the case where $\epsilon_p = 0$ because the implications of the impossibility theorem do not apply in the case of equal base rates.

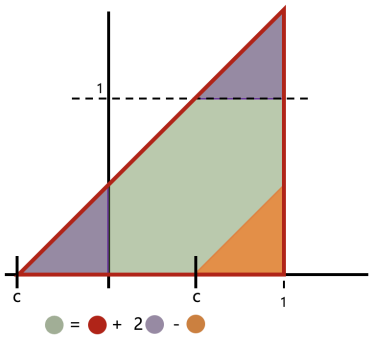


Figure 7: A sketch of the integral construction.

Proof. We begin with the result from Lemma D.2. Rearranging terms, we find the following expression for FPR:

$$\left(\frac{\epsilon_{\text{ACC}} + \epsilon_{\text{FPR}} - p(\epsilon_{\text{FPR}} - \epsilon_{\text{FNR}})}{\epsilon_p} + \epsilon_{\text{FNR}} - \epsilon_{\text{FPR}}\right) + \text{FNR} = \text{FPR} \tag{19}$$

Set $c = \text{FPR} - \text{FNR} = \left(\frac{\epsilon_{\text{ACC}} + \epsilon_{\text{FPR}} - p(\epsilon_{\text{FPR}} - \epsilon_{\text{FNR}})}{\epsilon_p} + \epsilon_{\text{FNR}} - \epsilon_{\text{FPR}}\right)$, which is fixed for prevalence p, prevalence difference ϵ_p, and a set relaxation factors $\epsilon_{\text{ACC}}, \epsilon_{\text{FNR}}, \epsilon_{\text{FPR}}$.

It's clear that the relationship between FPR and FNR is linear, and controlled by c, which can take on many possible values as we vary the relaxation parameters. We notate the set of values that c can take on as $C = \{c_1, c_2 ... c_m\}$. C is an infinite set.

However, C contains maximum and minimum values. From the linear relationship between FPR and FNR, we have $c_{max} = max(C)$ and $c_{min} = min(C)$. c_{max} and c_{min} will help us define the boundaries of the solution space A_f. It follows that $|A_f|$, the size of the solution space, when $|c| < 1$ and FPR, FNR < 1, is:

$$\begin{aligned} |A_f| = &\int_{-c_{max}}^{1} (c_{max} + \text{FNR})\, d\text{FNR} \\ &- \left(2\int_{-c_{max}}^{0} (c_{max} + \text{FNR})\, d\text{FNR}\right) \\ &- \left(\int_{-c_{min}}^{1} (c_{min} + \text{FNR})\, d\text{FNR}\right) \end{aligned} \tag{20}$$

For a sketch of the integral construction, see Figure 7. Using our construction over $\epsilon_{\text{ACC}}, \epsilon_{\text{FPR}}, \epsilon_{\text{FNR}} \in (-\gamma, \gamma)$, we can deduce c_{max} and c_{min}.

$$c = \frac{\epsilon_{\text{ACC}} + \epsilon_{\text{FPR}} - p(\epsilon_{\text{FPR}} - \epsilon_{\text{FNR}})}{\epsilon_p} + \epsilon_{\text{FNR}} - \epsilon_{\text{FPR}} \tag{21}$$

$$= \frac{\epsilon_{\text{ACC}}}{\epsilon_p} + \frac{\epsilon_{\text{FPR}}}{\epsilon_p} - \frac{p\epsilon_{\text{FPR}}}{\epsilon_p} + \frac{p\epsilon_{\text{FNR}}}{\epsilon_p} + \epsilon_{\text{FNR}} - \epsilon_{\text{FPR}} \tag{22}$$

$$= \frac{\epsilon_{\text{ACC}}}{\epsilon_p} + \frac{\epsilon_{\text{FPR}}}{\epsilon_p} - \frac{p\epsilon_{\text{FPR}}}{\epsilon_p} - \epsilon_{\text{FPR}} + \frac{p\epsilon_{\text{FNR}}}{\epsilon_p} + \epsilon_{\text{FNR}} \tag{23}$$

$$= \frac{\epsilon_{\text{ACC}}}{\epsilon_p} + \epsilon_{\text{FPR}}\left(\frac{1-p}{\epsilon_p} - 1\right) + \frac{p\epsilon_{\text{FNR}}}{\epsilon_p} + \epsilon_{\text{FNR}} \tag{24}$$

Note our assumption that $\epsilon_p < 1 - p$ gives us that $\frac{1-p}{\epsilon_p} > 1$ which yields: (25)

$$\leq \frac{\gamma}{\epsilon_p} + \gamma\left(\frac{1-p}{\epsilon_p} - 1\right) + \frac{p\gamma}{\epsilon_p} + \gamma \tag{26}$$

$$= \frac{\gamma + \gamma - \gamma p + \gamma p}{\epsilon_p} = \frac{2\gamma}{\epsilon_p} = c_{max} \tag{27}$$

A symmetric argument gives $c_{min} = -\frac{2\gamma}{\epsilon_p}$. To find $|A_f|$, we set $c = \frac{2\gamma}{\epsilon_p}$, and replace into the above integration:

$$\begin{aligned} |A_f| = &\int_{-c}^{1} c + \text{FNR}\, d\text{FNR} \\ &- \left(2\int_{-c}^{0} c + \text{FNR}\, d\text{FNR}\right) \\ &- \left(\int_{c}^{1} \text{FNR} - c\, d\text{FNR}\right) = 2c - c^2 = \frac{4\gamma}{\epsilon_p} - \frac{4\gamma^2}{{\epsilon_p}^2} \leq 1 \end{aligned} \tag{28}$$

This yields the result. □

E ANALYTICAL APPROACH TO CHARACTERIZING THE FAIRNESS REGION AREA USING FPR, FNR, PPV

See Figure 8.

Lemma E.1 (Expressing Fairness Area). *Consider groups g_1 with prevalence p_1 and g_2 with prevalence p_2. Without loss of generality, let us assume that $p_2 = p_1 + \epsilon_p$. Next, let us denote a predictor's performance as FPR_1, FPR_1 and PPV_1 for g_1, and FPR_1, FPR_1 and PPV_1 for g_2. Let ϵ_{FPR}, ϵ_{FNR} and ϵ_v denote acceptable differences in FPR, FNR and PPV between groups, respectively. That is, $|FPR_1 - FPR_2| \leq \epsilon_{FPR}$, $|FNR_1 - FNR_2| \leq \epsilon_{FNR}$ and $|PPV_1 - PPV_2| \leq \epsilon_v$. Then, the following equality holds (for the sake of space, let $FPR = \alpha$, $FNR = \beta$, and $PPV = v$):*

$$\beta = \frac{\epsilon_p(v^2(\epsilon_{FPR}(p-1)-1) + v\epsilon_v(\epsilon_{FPR}(p-1)-1) + p\epsilon_v + v) + (p-1)(\epsilon_{FPR}(p-1)v(v+\epsilon_v) + p\epsilon_v) - \epsilon_{FNR}(p-1)v(p+\epsilon_p)(v+\epsilon_v-1)}{(\epsilon_p(p\epsilon_v - v^2 - v\epsilon_v + v) + (p-1)p\epsilon_v)} \tag{29}$$

Proof. Consider (1) in the setting where there are two groups. Let the FPR of the two groups be equal, subject to the relaxation $\text{FPR}_2 = \text{FPR}_1 + \epsilon_{\text{FPR}}$. Make substitutions respective to the assumptions made in Lemma 8 to find:

$$\frac{p}{1-p}\frac{1-v}{v}(1-\beta) = \frac{p+\epsilon_p}{1-(p+\epsilon_p)}\frac{1-(v+\epsilon_v)}{v+\epsilon_v}(1-(\beta+\epsilon_{\text{FNR}})) + \epsilon_{\text{FPR}} \tag{30}$$

Solving for β yields (29). □

Figure 8: Analytical approach to characterizing the fairness region area with FPR, FNR, PPV.

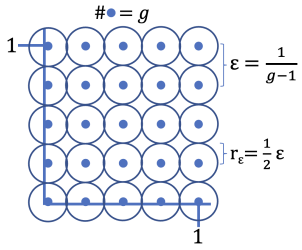


Figure 9: A full planimeter (g=25).

F INFORMAL ANALYSIS OF THE DOT PLANIMETER

One of the strategies we undertook in estimating the "fairness region" of 8 borrowed from previous work on "dot-planimetry," a two-dimensional area estimation technique that has been consistently reinvented over the past century, and has been studied extensively in relation to cartographic area estimation and pure math (see Gauss's circle problem) [10, 26, 42] (Perhaps sadly, GPS has contributed to the under-exploration of this subject in recent years). Here, we offer a brief informal analysis of our dot-planimetry strategy, and how we chose values based on approximated upper bounds on our over-estimation error for the area of the fairness region when considering FNR, FPR and PPV.

Dot-planimetry provides a simple way of estimating the area of complex enclosed shapes in a two dimensional space (which can be difficult to integrate directly). Intuitively, we create a regular grid over the space composed of points. We refer to each point as a "detector," who is responsible for a pre-specified radius r_ϵ in the 2-dimensional space of interest. We say that a "detector" i is "satisfied" if the edge of the shape of interest is anywhere within r_ϵ distance away from i. To compute the final area of the shape, we simply sum up the total number of satisfied detectors multiplied by each of their individual areas (i.e. the area of a bunch of circles defined by r_ϵ). In our case, we are interested in a very particular space: a 1×1 square with the bottom left corner at the origin (this is how we can visualize two metrics varying based on our relaxations). Our dot planimeter in this case has g^2 total detectors. They are distributed so that they are $\epsilon = \frac{1}{g-1}$ apart, and so that the bottom row and leftmost column each touch the x and y axis respectively, while the rightmost and top rows each have x and y values of 1 respectively. Thus, each detector has a radius of $r_\epsilon = \frac{1}{2}\epsilon$. Refer to Figure 9 for a sketch of this setup, and as we walk through the problem.

For our analysis, we will call two detectors i and j "neighboring" if $i_y = j_y$ and $|i_x - j_x| = \epsilon$ i.e. the values of their y-axis are equal and they are next to each other. Note that, for any of the following arguments, symmetric variations apply were we to switch the definition of "neighboring" to $i_x = j_x$. Neighboring detectors are shown in Figure 10.

We will also define our "detector function" to be:

$$f(i_{xy}, h(x)) = \begin{cases} 1, & |project(i_{xy}) \rightarrow h(x)| \leq r_\epsilon \\ 0, & \text{otherwise} \end{cases}$$

Intuitively, our detector function $f(i_{xy}, h(x))$ takes a detector i_{xy} and a boundary function of interest $h(x)$ and returns 1 (or true) if, for any x, $h(x)$ passes within r_ϵ of i_{xy}.

Our overall argument for a coarse upper bound on the over-estimation error when using a dot-planimeter is intuitive: we will reason about which function through our $[0,1]\times[0,1]$ metric region would lead to the highest number of detectors satisfied. We will then assume that this is the boundary function for our area (i.e. this function partitions our space, and is dense on one side). Our estimation error is then the proportion of total detectors satisfied by the boundary function, assuming that they all make minimal contact with the detector radius. **Again, this is a very coarse approach, and we might assume the actual overestimation error is much lower.** Due to the computational nature of the problem, this can provide us some measure of confidence in selecting a value for for g that gives reasonable error (like $< 5\%$).

Critical points all satisfying of neighboring detectors. Consider two sets of side by side (or "neighboring") detectors, i_1, j_1 and i_2, j_2 (one can refer to Figure 10). By the definition of a function,

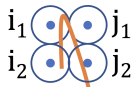


Figure 10: A set of two sets of side by side detectors, i_1, j_1 and i_2, j_2. Note that a critical point is necessary for $h(x)$ (in orange) to satisfy all 4.

$f(i_1, h(x)) + f(i_2, h(x)) + f(j_1, h(x)) + f(j_2, h(x)) \leq 3$ **unless** $g(x)$ has a critical point in the window $i_x \leq x \leq j_x$.

Extending the argument to columns. Consider two columns of neighboring detectors, which can represented by sets $\{i_1, i_2, ..., i_m\} \in I$ and $\{j_1, j_2, ..., j_m\} \in J$. Note that detectors in I share the same x value, as for J.

What is the maximum number of detectors satisfied by $h(x)$ between sets x? If we assume $|I|, |J| = g \geq 3$, then by our argument that critical points all satisfy neighboring detectors, if $h(x)$ has 0 critical points lie between I_x and J_x, then the answer is at most $g + 1$ (this can be seen through a geometric argument). However, we allow for any critical points, than the answer is $2g$ (or, the size of the entire union between the sets).

These arguments suggest that the determining factor in satisfying the highest number of detectors in a space is the number of critical points allowed for the function $h(x)$. These two column arguments can be extended to cover all detectors in the space, and provides a coarse upper bound on the over-estimation error.

Coarse upper bound on error of (ϵ)-dot-planimeter under assumptions of fairness region. Assume that $h(x)$ is the boundary function for our area $\in [0, 1] \times [0, 1]$ and has at most b critical points. Then, a coarse upper bound on the max detectors satisfied by $h(x)$ is $g * c$. This yields the simple *percent* error calculator: $\frac{c}{g}$ (as there are g^2 total detectors in our dot-planimeter).

Thus, for a 5% upper bound on error, assuming no more than 6 critical points for boundary function $h(x)$ for $x \in (0, 1)$, we have $\frac{6}{0.05} = g = 120$. Thus, with a granularity of $120^2 = 14400$ detector in our dot-planimeter, we have high confidence that the over-estimation error is no higher than 5%, so long as some assumptions we made about $h(x)$ hold (experimentally, we found this to be the case).

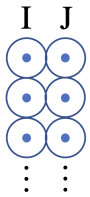


Figure 11: A set of two detector columns, I and J.

G ADDITIONAL PROOFS

Proposition G.1 (Intersectional Prevelance Differences). *Given a dataset that is subdivided into two groups, let $0 < p_1 < n_1$ and $0 < p_2 < n_2$, where p_i is the number of positive class members of group i, and n_i is the total number of members in group i. Suppose $\frac{p_1}{n_1} \leq \frac{p_2}{n_2}$.*

The following holds:

$$\frac{p_1}{n_1} \leq \frac{p_1 + p_2}{n_1 + n_2} \leq \frac{p_2}{n_2} \tag{31}$$

Proof. We know that $0 < p_1, n_1, p_2, n_2$. Note that the assumption $\frac{p_1}{n_1} \leq \frac{p_2}{n_2}$ implies $n_2 \leq \frac{n_1 p_2}{p_1}$.

Consider the left hand side of the equality proposed in the theorem statement:

$$\frac{p_1}{n_1} \leq \frac{p_1 + p_2}{n_1 + n_2} \tag{32}$$

$$p_1(n_1 + n_2) \leq n_1(p_1 + p_2) \tag{33}$$

$$p_1 n_2 \leq n_1 p_2 \tag{34}$$

$$n_2 \leq \frac{n_1 p_2}{p_1} \tag{35}$$

We have now derived an inequality specific to n_2. We can verify that exceeding this value invalidates the claim by plugging in $n_2 = (\frac{n_1 p_2}{p_1} + a)$, where $a > 0$ to find a contradiction:

$$p_1(n_1 + \frac{n_1 p_2}{p_1} + a) \leq n_1(p_1 + p_2) \tag{36}$$

$$p_1 n_1 + n_1 p_2 + a p_1 \leq p_1 n_1 + n_1 p_2 \tag{37}$$

$$a p_1 \leq 0 \tag{38}$$

which contradicts $a > 0$. An analogous argument exists for the right hand side of the equation. Thus, so long as $n_2 \leq \frac{n_1 p_2}{p_1}$, the main inequality holds.

□

Proposition G.2 (Reducing k increases PPV). *Given a well-calibrated classifier being used under a resource constraint k, reducing the size of k will monotonically increase the PPV of the classifier.*

Proof. Consider the top k outputs of a well-calibrated classifier. Since the classifier is calibrated, the list k is ordered in the following way: those elements closest to the first position have a higher probability of a positive outcome, and those elements closest to position k have a lower probability of a positive outcome. In other words, $p_1 \geq p_2 \geq, \ldots, \geq p_k$, where p_i is the probability that element i is a member of the positive class. The PPV of the classifier, in expectation, can then be expressed in the following way:

$$\frac{1}{k} \sum_{i=1}^{k} p_i = \frac{p_1 + p_2 + ... + p_k}{k}. \tag{39}$$

Notice that Equation 39 is simply the average of p_i over the k elements. Consider another set of size $k' < k$, that is constructed by removing the elements $p_{k'+1}, ...p_k$. Since by construction all the elements in the new set are *greater than or equal to* to the elements of the previous set, the average of p_i in the new set will be greater than or equal — or in other words, the PPV will be greater than or equal. □