Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification

A. Feder Cooper[1,2*], Katherine Lee[1,2], Madiha Zahrah Choksi[2], Solon Barocas[2, 3], Christopher De Sa[2], James Grimmelmann[1,2], Jon Kleinberg[2], Siddhartha Sen[3], Baobao Zhang[4]

[1]The Center for Generative AI, Law, and Policy Research
[2]Cornell University
[3]Microsoft Research
[4]Syracuse University

Abstract

Variance in predictions across different trained models is a significant, under-explored source of error in fair binary classification. In practice, the variance on some data examples is so large that decisions can be effectively *arbitrary*. To investigate this problem, we take an experimental approach and make four overarching contributions. We: 1) Define a metric called *self-consistency*, derived from variance, which we use as a proxy for measuring and reducing arbitrariness; 2) Develop an ensembling algorithm that abstains from classification when a prediction would be arbitrary; 3) Conduct the largest-to-date empirical study of the role of variance (*vis-a-vis* self-consistency and arbitrariness) in fair binary classification; and, 4) Release a toolkit that makes the US Home Mortgage Disclosure Act (HMDA) datasets easily usable for future research. Altogether, our experiments reveal shocking insights about the reliability of conclusions on benchmark datasets. Most fair binary classification benchmarks are *close-to-fair* when taking into account the amount of arbitrariness present in predictions — *before* we even try to apply *any* fairness interventions. This finding calls into question the practical utility of common algorithmic fairness methods, and in turn suggests that we should reconsider how we choose to measure fairness in binary classification.

1 Introduction

A goal of algorithmic fairness is to develop techniques that measure and mitigate discrimination in automated decision-making. In fair binary classification, this often involves training a model to satisfy a chosen *fairness metric*, which typically defines fairness as parity between model error rates for different demographic groups in the dataset (Barocas et al. 2019). However, even if a model's classifications satisfy a particular fairness metric, it is not necessarily the case that the model is equally confident in each classification.

To provide an intuition for what we mean by confidence, consider the following experiment: We fit 100 logistic regression models using the same learning process, which draws different subsamples of the training set from the COMPAS prison recidivism dataset (Larson et al. 2016; Friedler et al. 2019), and we compare the resulting classifications for two individuals in the test set. Figure 1 shows a difference in the consistency of predictions for both individuals: the 100 models agree completely to classify Individual 1 as "will recidivate" and disagree completely on whether to classify Individual 2 as "will" or "will not recidivate."

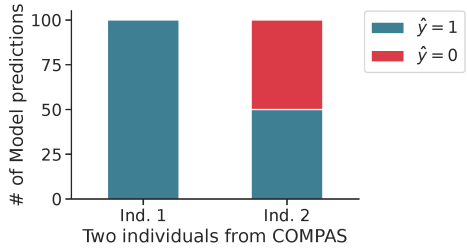


Figure 1: 100 bootstrapped logistic regression models show models can be very consistent in predictions $\hat{y}$ for some individuals (Ind. 1) and arbitrary for others (Ind. 2).

If we were to pick one model at random to use in practice, there would be no effect on how Individual 1 is classified; yet, for Individual 2, the prediction is effectively random. We can interpret this disagreement to mean that the learning process that produced these predictions is not sufficiently confident to justify assigning Individual 2 *either decision outcome*. In practice, instances like Individual 2 exhibit so little confidence that their classification is effectively *arbitrary* (Cooper et al. 2022a,b; Creel and Hellman 2022). Further, this arbitrariness can also bring about discrimination if classification decisions are *systematically more arbitrary* for individuals in certain demographic groups.

A key aspect of this example is that we use only one model to make predictions. This is the typical setup in fair binary classification: Popular metrics are commonly applied to evaluate the fairness of a *single model* (Hardt et al. 2016; Pleiss et al. 2017; Kleinberg et al. 2017). However, as is clear from the example learning process in Fig. 1, using only a single model can mask the arbitrariness of predictions. Instead, to reveal arbitrariness, we must examine *distributions over possible models for a given learning process*. With this shift in frame, we ask: *What is the empirical role of arbitrariness in fair binary classification tasks?*

To study this question, we make four contributions:

*Full, authoritative paper at https://arxiv.org/abs/2301.11562.

1. **Quantify arbitrariness.** We formalize a metric called *self-consistency*, derived from statistical variance, which we use as a quantitative proxy for arbitrariness of model outputs. Self-consistency is a simple yet powerful tool for empirical analyses of fair classification (§3).
2. **Ensemble to improve self-consistency.** We extend Breiman's classic bagging to allow for abstaining from classifying instances for which self-consistency is low. This improves overall self-consistency (i.e., reduces variance), and improves accuracy (§4).
3. **Perform a comprehensive experimental study of variance in fair binary classification.** We conduct the largest-to-date such study, through the lens of self-consistency and its relationship to arbitrariness. Surprisingly, we find that most benchmarks are *close-to-fair* when taking into account the amount of arbitrariness present in predictions — *before* we even try to apply *any* fairness interventions (§5). This shocking finding has huge implications for the field: it casts doubt on the reliability of prior work that claims there is baseline unfairness in these benchmarks, in order to demonstrate that methods to improve fairness work in practice. We instead find that such methods are often empirically unnecessary to improve fairness (§6).
4. **Release a large-scale fairness dataset package.** We observe that variance, particularly in small datasets, can undermine the reliability of conclusions about fairness. We therefore open-source a package that makes the large-scale US Home Mortgage Disclosure Act datasets (`HMDA`) easily usable for future research.

2 Preliminaries on Fair Binary Classification

To analyze arbitrariness in the context of fair binary classification, we first need to establish our background definitions. This material is likely familiar to most readers. Nevertheless, we highlight particular details that are important for understanding the experimental methods that enable our contributions. We present the fair-binary-classification problem formulation and associated empirical approximations, with an emphasis on the *distribution over possible models* that could be produced from training on different subsets of data drawn from the same data distribution.

Problem formulation. Consider a distribution $q(\cdot)$ from which we can sample *examples* $(\boldsymbol{x}, \boldsymbol{g}, o)$. The $\boldsymbol{x} \in \mathbb{X} \subseteq \mathbb{R}^m$ are feature *instances* and $\boldsymbol{g} \in \mathbb{G}$ is a *protected attribute* that we do not use for learning (e.g., race, gender). The $o \in \mathbb{O}$ are the associated *observed labels*, and $\mathbb{O} \subseteq \mathbb{Y}$, where $\mathbb{Y} = \{0, 1\}$ is the label space. From $q(\cdot)$ we can sample training datasets $\{(\boldsymbol{x}, \boldsymbol{g}, o)\}_{i=1}^n$, with $\mathbb{D}$ representing the set of all n-sized datasets. To reason about the possible models of a hypothesis class $\mathbb{H}$ that could be learned from the different subsampled datasets $\boldsymbol{D}_k \in \mathbb{D}$, we define a *learning process*:

Definition 2.1. A **learning process** is a randomized function that runs instances of a **training procedure** $\mathcal{A}$ on each $\boldsymbol{D}_k \in \mathbb{D}$ and a model specification, in order to produce **classifiers** $h_{\boldsymbol{D}_k} \in \mathbb{H}$. A particular run $\mathcal{A}(\boldsymbol{D}_k) \to h_{\boldsymbol{D}_k}$, where $h_{\boldsymbol{D}_k} : \mathbb{X} \to \mathbb{Y}$, which is deterministic mapping from the instance space $\mathbb{X}$ to the label space $\mathbb{Y}$. All such runs over $\mathbb{D}$ produce a distribution over possible trained models, μ.

Reasoning about μ, rather than individual models $h_{\boldsymbol{D}_k}$, enables us to contextualize arbitrariness in the data, which, in turn, is captured by learned models (§3). Each particular model $h_{\boldsymbol{D}_k} \sim \mu$ deterministically produces classifications $\hat{y} = h_{\boldsymbol{D}_k}(\boldsymbol{x})$. The classification rule is $h_{\boldsymbol{D}_k}(\boldsymbol{x}) = \mathbf{1}[r_{\boldsymbol{D}_k}(\boldsymbol{x}) \geq \tau]$, for some threshold τ, where regressor $r_{\boldsymbol{D}_k} : \mathbb{X} \to [0, 1]$ computes the probability of positive classification. Executing $\mathcal{A}(\boldsymbol{D}_k)$ produces $h_{\boldsymbol{D}_k} \sim \mu$ by minimizing the *loss* of predictions $\hat{y}$ with respect to their associated observed labels o in $\boldsymbol{D}_k$. This loss is computed by a chosen *loss function* $\ell : \mathbb{Y} \times \mathbb{Y} \mapsto \mathbb{R}$. We compute predictions for a *test set* of fresh examples and calculate their loss. The loss is an estimate of the *error* of $h_{\boldsymbol{D}_k}$, which is dependent on the specific dataset $\boldsymbol{D}_k$ used for training. To generalize to the error of all possible models produced by a specific learning process (Def. 2.1), we consider the *expected error*, $\texttt{Err}(\mathcal{A}, \mathbb{D}, (\boldsymbol{x},\ \boldsymbol{g},\ o)) = \mathbb{E}_{\mathbf{D}}[\ell(o,\ \hat{y})|\mathbf{x} = \boldsymbol{x}]$.

In fair classification, it is common to use *0-1 loss* $\triangleq \mathbf{1}[\hat{y} \neq o]$ or *cost-sensitive loss*, which assigns asymmetric costs C_{01} for false positives `FP` and C_{10} for false negatives `FN`. These costs are related to the classifier threshold $\tau = \frac{C_{01}}{C_{01}+C_{10}}$, with $C_{01}, C_{10} \in \mathbb{R}^+$ (§A.3). Common fairness metrics, such as Equality of Opportunity (Hardt et al. 2016), further analyze error by computing disparities across group-specific error rates $\texttt{FPR}_{\boldsymbol{g}}$ and $\texttt{FNR}_{\boldsymbol{g}}$. For example, $\texttt{FPR}_{\boldsymbol{g}} \triangleq p_\mu[r_{\mathbf{D}}(\mathbf{x}) \geq \tau | o = 0, \mathbf{g} = \boldsymbol{g}] = p_\mu[\hat{y} = 1 | o = 0, \mathbf{g} = \boldsymbol{g}]$. Model-specific $\texttt{FPR}_{\boldsymbol{g}}$ and $\texttt{FNR}_{\boldsymbol{g}}$ are further-conditioned on the dataset used in training, i.e., $\mathbf{D} = \boldsymbol{D}_k$.

Empirical approximation. We typically only have access to one dataset, not the data distribution $q(\cdot)$. In fair binary classification experiments, it is common to estimate expected error by performing *cross validation* (CV) on this dataset to produce a small handful of models (Chen et al. 2018; Corbett-Davies et al. 2017). CV can be unreliable when there is high variance; it can produce error estimates that are themselves high variance, and does not reliably estimate expected error with respect to possible models μ (§5). For more details, see Efron and Tibshirani (1997, 1993) and Wager (2020).

To get around these reliability issues, one can *bootstrap*. Bootstrapping splits the available data into train and test sets, and simulates drawing different training datasets from a distribution by resampling the train set $\hat{\boldsymbol{D}}$, generating replicates $\hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2, \ldots, \hat{\boldsymbol{D}}_B := \hat{\mathbb{D}}$. We use these replicates $\hat{\mathbb{D}}$ to approximate the learning process on $\mathbb{D}$ (Def. 2.1). We treat the resulting $\hat{h}_{\hat{\boldsymbol{D}}_1}, \hat{h}_{\hat{\boldsymbol{D}}_2}, \ldots, \hat{h}_{\hat{\boldsymbol{D}}_B}$ as our empirical estimate for the distribution $\hat{\mu}$, and evaluate their predictions for the *same reserved test set*. This enables us to produce comparisons of classifications across test instances like in Fig. 1 (§A.4).

3 Variance, Self-Consistency & Arbitrariness

We develop a quantitative proxy for measuring arbitrariness, called *self-consistency* (§3), which is derived from a definition of statistical *variance* between different model predictions (§3). We then illustrate how self-consistency is a simple-yet-powerful tool for revealing the role of arbitrariness in fair classification (§3). Next, we will introduce an algorithm to improve self-consistency (§4) and compute self-consistency on popular fair binary classification benchmarks (§5).

Arbitrariness Resembles Statistical Variance

In Section 2, we discussed how common fairness definitions analyze error by computing false positive rate (FPR) and false negative rate (FNR). Another common way to formalize error is as a decomposition of different statistical sources: *noise*-, *bias*-, and *variance*-induced error (Abu-Mostafa et al. 2012; Geman et al. 1992). To understand our metric for self-consistency (§3), we first describe how the arbitrariness in Figure 1 (almost, but not quite) resembles variance.

Informally, variance-induced error quantifies fluctuations in individual example predictions for different models $h_{\boldsymbol{D}_k} \sim \mu$. Variance is the error in the learning process that comes from training on different datasets $\boldsymbol{D}_k \in \mathbb{D}$. In theory, we measure variance by imagining training all possible $h_{\boldsymbol{D}_k} \sim \mu$, testing them all on the same test instance $(\boldsymbol{x}, \boldsymbol{g})$, and then quantifying how much the resulting classifications for $(\boldsymbol{x}, \boldsymbol{g})$ deviate *from each other*. More formally,

Definition 3.1. For all pairs of possible models $h_{\boldsymbol{D}_i}, h_{\boldsymbol{D}_j} \sim \mu$ $(i \neq j)$, the **variance** for a test $(\boldsymbol{x}, \boldsymbol{g})$ is

$$\texttt{var}\big(\mathcal{A}, \mathbb{D}, (\boldsymbol{x}, \boldsymbol{g})\big) \triangleq \mathbb{E}_{h_{\boldsymbol{D}_i} \sim \mu, h_{\boldsymbol{D}_j} \sim \mu}\Big[\ell\Big(h_{\boldsymbol{D}_i}(\boldsymbol{x}), h_{\boldsymbol{D}_j}(\boldsymbol{x})\Big)\Big].$$

We can approximate variance directly by using the bootstrap method (§2, §B.1). For 0-1 and cost-sensitive loss with costs $C_{01}, C_{10} \in \mathbb{R}^+$ (§2), we can generate B replicates to train B concrete models that serve as our approximation for the distribution $\hat{\mu}$. For $B = B_0 + B_1 > 1$, where B_0 and B_1 denote the number of 0- and 1-class predictions for $(\boldsymbol{x}, \boldsymbol{g})$,

$$\begin{aligned}\hat{\texttt{var}}\big(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\big) &\coloneqq \frac{1}{B(B-1)} \sum_{i \neq j} \ell\Big(\hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x}), \hat{h}_{\hat{\boldsymbol{D}}_j}(\boldsymbol{x})\Big) \\ &= \frac{(C_{01} + C_{10}) B_0 B_1}{B(B-1)}. \end{aligned} \tag{1}$$

We derive (1) in Appendix B.2 and show that, for increasingly large B, $\hat{\texttt{var}}$ is defined on $[0, \frac{C_{01}+C_{10}}{4} + \epsilon]$.

Defining Self-Consistency from Variance

It is clear from above that, in general, variance (1) is unbounded. We can always increase the maximum possible $\hat{\texttt{var}}$ by increasing the magnitudes of our chosen C_{01} and C_{10} (§2. However, as we can see from our intuition for arbitrariness in Figure 1, the most important takeaway is the amount of (dis)agreement, reflected in the counts B_0 and B_1. Here, there is no notion of the cost of misclassifications. So, variance (1) does not exactly measure what we want to capture. Instead, we want to focus unambiguously on the (dis)agreement part of variance, which we call *self-consistency of the learning process*:

Definition 3.2. For all pairs of possible models $h_{\boldsymbol{D}_i}, h_{\boldsymbol{D}_j} \sim \mu$ $(i \neq j)$, the **self-consistency of the learning process** for a test $(\boldsymbol{x}, \boldsymbol{g})$ is

$$\begin{aligned}\texttt{SC}\big(\mathcal{A}, \mathbb{D}, (\boldsymbol{x}, \boldsymbol{g})\big) &\triangleq \mathbb{E}_{h_{\boldsymbol{D}_i} \sim \mu, h_{\boldsymbol{D}_j} \sim \mu}\Big[h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})\Big] \\ &= p_{h_{\boldsymbol{D}_i} \sim \mu, h_{\boldsymbol{D}_j} \sim \mu}\big(h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})\big). \end{aligned} \tag{2}$$

In words, (2) models the probability that two models produced by the same learning process on different n-sized training datasets agree on their predictions for the same test instance. Like variance, we can derive an empirical approximation of SC. Using the bootstrap method with $B = B_0 + B_1 > 1$,

$$\begin{aligned}\hat{\texttt{SC}}\big(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\big) &\coloneqq \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{1}\Big[\hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x}) = \hat{h}_{\hat{\boldsymbol{D}}_j}(\boldsymbol{x})\Big] \\ &= 1 - \frac{2 B_0 B_1}{B(B-1)}. \end{aligned} \tag{3}$$

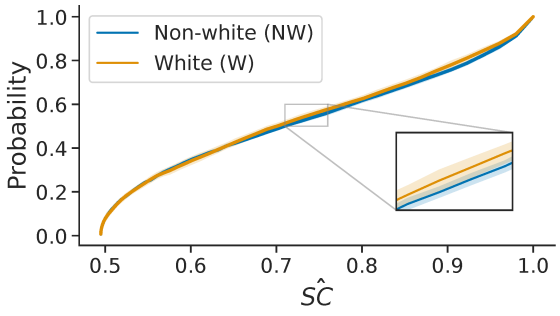
For increasingly large B, $\hat{\texttt{SC}}$ is defined on $[0.5 - \epsilon, 1]$ (§B.3). Throughout, we use the shorthand *self-consistency*, but it is important to note that Definition 3.2 is a property of the distribution over possible models μ produced by the learning process, not of individual models. We summarize other important takeaways below:

Terminology. In naming our metric, we intentionally evoke related notions of "consistency" in logic and the law (Fuller (1965); Stalnaker (2006); §B.3).

Interpretation. Definition 3.2 is defined on $[0.5, 1]$, which coheres with the intuition in Figure 1: 0.5 and 1 respectively reflect minimal (Individual 2) and maximal (Individual 1) possible SC. SC, unlike FPR and FNR (§2), does *not* depend on the observed label o. It captures the learning process's confidence in a classification $\hat{y}$, but says nothing directly about $\hat{y}$'s accuracy. By construction, low self-consistency indicates high variance, and vice versa. We derive empirical $\hat{\texttt{SC}}$ (3) from $\hat{\texttt{var}}$ (1) by leveraging observations about the definition of $\hat{\texttt{var}}$ for 0-1 loss (§B.3). While there are no costs C_{01}, C_{10} in computing (3), they still affect empirical measurements of $\hat{\texttt{SC}}$. Because C_{01} and C_{10} affect τ (§2), they control the concrete number of B_0 and B_1, and thus the $\hat{\texttt{SC}}$ we measure in experiments.

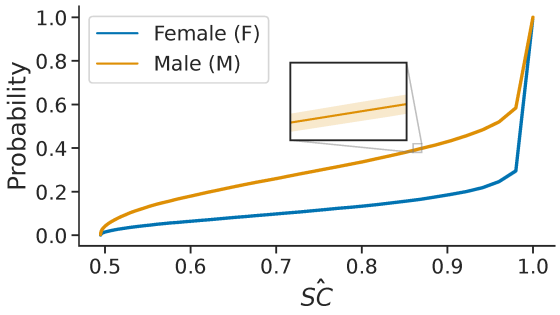
Empirical focus. Since self-consistency depends on the particular data subsets used in training, conclusions about its relevance vary according to task. This is why we take a practical approach for our main results — of running a large-scale experimental study on many different datasets to extract general observations about $\hat{\texttt{SC}}$'s practical effects (§5). In our experiments, we typically use $B = 101$, which yields a $\hat{\texttt{SC}}$ range of $[\approx 0.495, 1]$ in practice.

Relationship to other fairness concepts. Self-consistency is qualitatively different from traditional fairness metrics. Unlike FPR and FNR, SC does not depend on observed label o. This has two important implications. First, while calibration also measures a notion of confidence, it is different: calibration reflects confidence with respect to *a model* predicting o, but says nothing about the relative confidence in predictions $\hat{y}$ produced by the *possible models* μ that result from the learning process (Pleiss et al. 2017). Second, a common assumption in algorithmic fairness is that there is *label bias* — that unfairness is due in part to discrimination reflected in recorded, observed decisions o (Friedler et al. 2016; Cooper and Abrams 2021). As a result, it is arguably a nice side effect that self-consistency does not depend on o. However, it is also possible to be perfectly self-consistent and inaccurate (e.g., $\forall k, \hat{y}_k \neq o$; §6).



	$\Delta\hat{\texttt{Err}}$	$\Delta\hat{\texttt{FPR}}$	$\Delta\hat{\texttt{FNR}}$
	$1.0 \pm 1.4\%$	$2.0 \pm 1.4\%$	$0.9 \pm 1.4\%$
	$\hat{\texttt{Err}}$	$\hat{\texttt{FPR}}$	$\hat{\texttt{FNR}}$
Total	$36.6 \pm 0.5\%$	$17.3 \pm 0.8\%$	$19.3 \pm 0.7\%$
NW	$36.9 \pm 0.5\%$	$18.0 \pm 0.7\%$	$19.0 \pm 0.8\%$
W	$35.9 \pm 1.3\%$	$16.0 \pm 1.2\%$	$19.9 \pm 1.1\%$

(a) `COMPAS` split by `race`; random forests (RFs)



	$\Delta\hat{\texttt{Err}}$	$\Delta\hat{\texttt{FPR}}$	$\Delta\hat{\texttt{FNR}}$
	$12.2 \pm 0.4\%$	$6.0 \pm 0.3\%$	$6.3 \pm 0.3\%$
	$\hat{\texttt{Err}}$	$\hat{\texttt{FPR}}$	$\hat{\texttt{FNR}}$
Total	$17.3 \pm 0.3\%$	$7.7 \pm 0.3\%$	$9.6 \pm 0.1\%$
F	$9.0 \pm 0.3\%$	$3.7 \pm 0.1\%$	$5.3 \pm 0.3\%$
M	$21.2 \pm 0.3\%$	$9.7 \pm 0.3\%$	$11.6 \pm 0.1\%$

(b) `Old Adult` split by `sex`; random forests (RFs)

Figure 2: $\hat{\texttt{SC}}$ CDFs for `COMPAS` (2a) and `Old Adult` (2b). We train random forests ($B = 101$ replicates), and repeat with 10 train/test splits to produce (very tight) confidence intervals. $\hat{\texttt{SC}}$ is effectively identical across subgroups $\boldsymbol{g}$ in `COMPAS`; `Old Adult` exhibits systematic differences in arbitrariness across $\boldsymbol{g}$. T ables show mean $\pm$ STD of the relative disparities, e.g., $\Delta\hat{\texttt{Err}} = |\hat{\texttt{Err}}_0 - \hat{\texttt{Err}}_1|$ (top); and, the absolute $\hat{\texttt{Err}}, \hat{\texttt{FPR}}, \hat{\texttt{FNR}}$, and $\hat{\texttt{SC}}$, also broken down by $\boldsymbol{g}$ (bottom) (§E).

Illustrating Self-Consistency in Practice

$\hat{\texttt{SC}}$ enables us to evaluate arbitrariness in classification experiments. It is straightforward to compute $\hat{\texttt{SC}}$ (3) with respect to multiple test instances $(\boldsymbol{x}, \boldsymbol{g})$ — for all instances in a test set or for all instances conditioned on membership in $\boldsymbol{g}$. Therefore, beyond visualizing $\hat{\texttt{SC}}$ for individuals (Fig. 1), we can also do so across sets of individuals. We plot the cumulative distribution (CDF) of $\hat{\texttt{SC}}$ for the groups $\boldsymbol{g}$ in the test set (i.e., the x-axis shows the range of $\hat{\texttt{SC}}$ for $B = 101$, $[\approx 0.495, 1]$). In Figure 2, we provide illustrative examples from two of the most common fair classification benchmarks (Fabris et al. 2022), `COMPAS` and `Old Adult` using random forests (RFs). We split the available data into train and test sets, and bootstrap the train set $B = 101$ times to train models $\hat{h}_1, \hat{h}_2, \ldots, \hat{h}_{101}$ (§2). We repeat this process on 10 train/test splits, and the resulting confidence intervals (shown in the inset) indicate that our $\hat{\texttt{SC}}$ estimates are stable. We group observations into two categories:

Individual arbitrariness. Both CDFs show that $\hat{\texttt{SC}}$ varies drastically across test instances. For random forests on the `COMPAS` dataset, about one-half of instances are under .7 self-consistent. Nearly one-quarter of test instances are effectively .5 self-consistent; they resemble Individual 2 in Figure 1, meaning that their predictions are essentially arbitrary. These differences in $\hat{\texttt{SC}}$ across the test set persist even though the 101 models exhibit relatively small average disparities $\Delta\hat{\texttt{Err}}$, $\Delta\hat{\texttt{FPR}}$, and $\Delta\hat{\texttt{FNR}}$ (Fig. 2a, bottom; §5). This supports our motivating claim: it is possible to come close to satisfying fairness metrics, while the learning process exhibits very different levels of confidence for the underlying classifications that inform those metrics (§1).

Systematic arbitrariness. We can also highlight $\hat{\texttt{SC}}$ according to groups $\boldsymbol{g}$. The $\hat{\texttt{SC}}$ plot for `Old Adult` shows that it is possible for the degree of arbitrariness to be *systematically worse* for a particular demographic $\boldsymbol{g}$ (Fig. 2b). While the lack of $\hat{\texttt{SC}}$ is not as extreme as it is for `COMPAS` (Fig. 2a) — the majority of test instances exhibit over .9 $\hat{\texttt{SC}}$ — there is more arbitrariness in the `Male` subgroup. We can quantify such *systematic arbitrariness* using a measure of distance between probability distributions. We use the Wasserstein-1 distance ($\mathcal{W}_1$), which has a closed form for CDFs (Ramdas et al. 2015). The $\mathcal{W}_1$ distance has an intuitive interpretation for measuring systematic arbitrariness: it computes the total disparity in `SC` by examining all possible `SC` levels κ at once (§B.3). For two groups $\boldsymbol{g} = 0$ and $\boldsymbol{g} = 1$ with respective `SC` CDFs F_0 and F_1, $\mathcal{W}_1 \triangleq \int_{\mathbb{R}} |F_0(\kappa) - F_1(\kappa)| \, d\kappa$. For `Old Adult`, $\hat{\mathcal{W}}_1 = 0.127$; for `COMPAS`, which does not show systematic arbitrariness, $\hat{\mathcal{W}}_1 = 0.007$.

4 Accounting for Self-Consistency

By definition, low $\hat{\texttt{SC}}$ signals that there is high `var` (§3). It is therefore a natural idea to use variance reduction techniques to improve $\hat{\texttt{SC}}$ (and thus reduce arbitrariness).

As a starting point for improving $\hat{\texttt{SC}}$, we perform variance reduction with Breiman (1996)'s *bootstrap aggregation*, or *bagging*, ensembling algorithm. Bagging involves bootstrapping to produce a set of B models (§2), and then, for each test instance, producing an aggregated prediction $\hat{y}_A$, which takes the majority vote of the $\hat{y}_1, \ldots, \hat{y}_B$ classifications. This procedure is practically effective for classifiers with high vari-

Algorithm 1: $\widehat{\texttt{SC}}$ Ensembling with Abstention

Input: training dataset $(\boldsymbol{X}, \boldsymbol{o})$, $\mathcal{A}$, B, $\widehat{\texttt{SC}}$ $\kappa \in [0.5, 1]$, $\boldsymbol{x}_{\text{test}}$
Output: $\hat{y}$ with $\widehat{\texttt{SC}} \geq \kappa$ or `Abstain`

1: $\hat{y}_A \coloneqq \mathsf{list}()$ ▷ To store ensemble predictions
2: **for** $1 \dots B$ **do**
3: $\quad \boldsymbol{D}_B \leftarrow \mathsf{Bootstrap}\big((\boldsymbol{X}, \boldsymbol{o})\big)$
4: $\quad \triangleright$ $\hat{h}_{\boldsymbol{D}_B}$ can itself be a bagged model, with $\mathcal{A}$ bagging on
5: $\quad\quad \boldsymbol{D}_B$ as the dataset to bootstrap
6: $\quad \hat{h}_{\boldsymbol{D}_B} \leftarrow \mathcal{A}(\boldsymbol{D}_B)$
7: $\quad \hat{y}_A.\mathsf{append}\big(\hat{h}_{\boldsymbol{D}_B}(\boldsymbol{x}_{\text{test}})\big)$ ▷ $\hat{y}_A = [\hat{y}_1, \dots, \hat{y}_B]$
8: **end for**
9: **return** $\mathsf{Aggregate}(\hat{y}_A, \kappa)$
10: ▷ Returns κ-majority prediction or abstains
11: **function** $\mathsf{Aggregate}\big(\hat{y}_1, \dots, \hat{y}_B, \kappa\big)$
12: $\quad$ **if** $\mathsf{SelfConsistency}(\hat{y}_1, \dots, \hat{y}_B) \geq \kappa$ ▷ Compute $\widehat{\texttt{SC}}$ (3)
13: $\quad\quad$ **return** $\arg\max_{y' \in \mathbb{Y}} \left[\sum_{i=1}^{B} \mathbf{1}[y' = \hat{y}_i] \right]$
14: $\quad$ **end if**
15: $\quad$ **return** `Abstain`
16: **end function**

ance (Breiman 1996, 1998). However, by taking the majority vote, bagging embeds the idea that having slightly-better-than-random classifiers is sufficient for improving ensembled predictions, $\hat{y}_A$. Unfortunately, there exist instances like Individual 2 (Fig. 1), where the classifiers in the ensemble are evenly split between classes. This means that bagging alone cannot overcome arbitrariness (§D.1).

To remedy this, we add the option to abstain from prediction if $\widehat{\texttt{SC}}$ is low (Alg. 1). A minor adjustment to (3) accounts for abstentions, and a simple proof follows that Algorithm 1 improves $\widehat{\texttt{SC}}$ (§D). We bootstrap as usual, but produce a prediction $\hat{y} \in [0, 1]$ for instance $\boldsymbol{x}$ only if $\boldsymbol{x}$ surpasses a user-specified minimum level κ of $\widehat{\texttt{SC}}$; otherwise, if an instance fails to achieve $\widehat{\texttt{SC}}$ of at least κ, we `Abstain` from predicting. For evaluation, we divide the test set into two subsets: we group together the instances we `Abstain` on in an *abstention set* and those we predict on in a *prediction set*. This method improves self-consistency through two complementary mechanisms: 1) variance reduction (due to bagging, see §D) and 2) abstaining from instances that exhibit low $\widehat{\texttt{SC}}$ (thereby raising the overall amount of $\widehat{\texttt{SC}}$ for the prediction set, see §D).

Further, since variance is a component of error (§3), variance reduction also tends to improve accuracy (Breiman 1996). This leads to an important observation: the abstention set, by definition, exhibits high variance; we can therefore expect it to exhibit higher error than the prediction set (§5, §E). So, while at first glance it may seem odd that our solution for arbitrariness is to *not predict*, it is worth noting that we often would have predicted incorrectly on a large portion of the abstention set, anyway (§D). In practice, we test two versions of our method:

Simple ensembling. We run Algorithm 1 to build ensembles of typical hypothesis classes in algorithmic fairness. For example, running with $B = 101$ decision trees and $\kappa = 0.75$ produces a bagged classifier that contains 101 underlying decision trees, for which the bagged classifier abstains from predicting on test instances that exhibit less than 0.75 $\widehat{\texttt{SC}}$. If overall $\widehat{\texttt{SC}}$ is low, then simple ensembling will lead to a large number of abstentions. For example, almost half of all test instances in `COMPAS` using random forests would fail to surpass the threshold $\kappa = 0.75$ (Fig. 2a). The potential for large abstention sets informs our second approach.

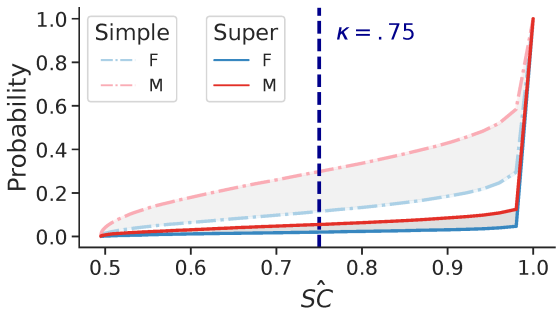
Super ensembling. We run Algorithm 1 on *bagged* models $\hat{h}$. When there is low $\widehat{\texttt{SC}}$ (i.e., high $\widehat{\texttt{var}}$) it can be beneficial to do an initial pass of variance reduction. We produce bagged classifiers using traditional bagging, but without abstaining (at Alg. 1, lines 4-5); *then* we $\mathsf{Aggregate}$ using those bagged classifiers as the underlying models $\hat{h}$. The first round of bagging raises the overall $\widehat{\texttt{SC}}$ before the second round, which is when we decide whether to `Abstain` or not. We therefore expect this approach to abstain less; however, it may potentially incur higher error, if, by happenstance, simple-majority-vote bagging chooses $\hat{y} \neq o$ for instances with very low $\widehat{\texttt{SC}}$ (§D). We also experiment with an $\mathsf{Aggregate}$ rule that averages the output probabilities of the underlying regressors $r_{\boldsymbol{D}_k}$, and then applies threshold τ to produce ensembled predictions. We do not observe major differences in results.

5 Experiments

We release an extensible package of different $\mathsf{Aggregate}$ methods, with which we trained and compared several million different models (all told, taking on the order of 10 hours of compute). We include results covering common datasets and models: `COMPAS`, `Old Adult`, `German` and `Taiwan Credit`, and 3 large-scale `New Adult - CA` tasks on logistic regression (LR), decision trees (DTs), random forests (RFs), MLPs, and SVMs (§E). Our results are shocking: by using Algorithm 1, we happened to observe close-to-fairness in nearly every task. Mitigating arbitrariness leads to fairness, *without* applying common fairness-improving interventions (§5, §E).

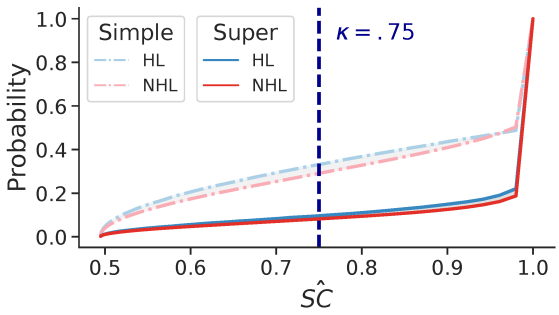
Releasing an `HMDA` toolkit. A possible explanation is that most fairness benchmarks are small ($< 25,000$ examples) and therefore exhibit high variance. We therefore clean a larger, more diverse, and newer dataset for investigating fair binary classification — the Home Mortgage Disclosure Act (`HMDA`) 2007-2017 datasets (FFIEC 2017) — and release them with a standalone, easy-to-use software package. In this paper, we examine the `NY` and `TX 2017` subsets of `HMDA`, which have $244,107$ and $576,978$ examples, respectively, and we still find close-to-fairness (§5, §E).

Presentation. To visualize Algorithm 1, we plot the CDFs of the $\widehat{\texttt{SC}}$ of the underlying models used in each ensembling method. We simultaneously plot the results of simple ensembling (dotted curves) and super ensembling (solid curves). Instances to the left of the vertical line (the minimum $\widehat{\texttt{SC}}$ threshold κ) form the abstention set. We also provide corresponding mean $\pm$ STD fairness and accuracy metrics for individual models (our expected, but not-necessarily-practically-attainable baseline) and for both simple and super ensembling. For ensembling methods, we report these metrics on the prediction set, along with the abstention rate ($\widehat{\texttt{AR}}$).



	Baseline	Simple	Super
$\Delta\hat{\text{FNR}}$	$6.3 \pm .3\%$	$4.1 \pm .3\%$	$5.8 \pm .4\%$
$\hat{\text{FNR}}_{\text{F}}$	$5.3 \pm .3\%$	$3.5 \pm .1\%$	$4.9 \pm .2\%$
$\hat{\text{FNR}}_{\text{M}}$	$11.6 \pm .1\%$	$7.6 \pm .3\%$	$10.7 \pm .3\%$

(a) `Old Adult` split by `sex`; random forests (RFs)



	Baseline	Simple	Super
$\Delta\hat{\text{FNR}}$	$0.7 \pm .2\%$	$1.1 \pm .3\%$	$2.2 \pm .3\%$
$\hat{\text{FNR}}_{\text{HL}}$	$10.1 \pm .2\%$	$3.3 \pm .3\%$	$8.0 \pm .3\%$
$\hat{\text{FNR}}_{\text{NHL}}$	$9.4 \pm .1\%$	$2.2 \pm .1\%$	$5.8 \pm .1\%$

(b) `HMDA-NY-2017` split by `ethnicity`; random forests (RFs)

Figure 3: Algorithm 1: simple and super ensembling RFs for `Old Adult` (3a) and `HMDA-NY-2017` (3b). Tables show $\hat{\text{FNR}}$ (mean $\pm$ STD) for individual models (Baseline) and each ensembling method's prediction set; $B = 101$, 10 train/test splits (§E). To highlight systematic arbitrariness (§3), we shade in gray the area between group-specific $\hat{SC}$ CDFs for each method. An initial pass of variance reduction in super significantly decreases the systematic arbitrariness in `Old Adult`.

We necessarily defer most of our results to the §E. Here, we exemplify two overarching themes: the effectiveness of both ensembling variants (§5), and how our results reveal shocking insights about reliability in fair binary classification research (§5). For all experiments, we illustrate Algorithm 1 with $\kappa = 0.75$, but note that κ is task-dependent in practice.

Validating Algorithm 1

We highlight results for two illustrative examples: `Old Adult` and `HMDA-NY-2017`, for `ethnicity` (Hispanic or Latino (HL), Non-Hispanic or Latino (NHL)). We plot $\hat{SC}$ CDFs and show $\hat{\text{FNR}}$ metrics using random forests (RFs). For `Old Adult`, the expected disparity of the RF baseline is $\Delta\hat{\text{FNR}} = 6.3\%$. The dashed set of curves plots the underlying $\hat{SC}$ for these RFs (Fig. 3a). When we apply simple to these RFs, overall $\hat{\text{Err}}$ decreases (§E), shown in part by the decrease in $\hat{\text{FNR}}_{\text{F}}$ and $\hat{\text{FNR}}_{\text{M}}$. Fairness also improves: $\Delta\hat{\text{FNR}}$ decreases to 4.1%. However, the corresponding $\hat{\text{AR}}$ is quite high, especially for the `Male` subgroup ($\boldsymbol{g} = \text{M}$, Fig. 4).

As expected, super improves overall $\hat{SC}$ through a first pass of variance reduction (§4). The $\hat{SC}$ CDF curves are brought down, indicating a lower proportion of the test set exhibits low $\hat{SC}$. Abstention rate $\hat{\text{AR}}$ is lower and more equal (Fig. 4); however, error, while still lower than the baseline RFs, has gone up for all metrics. There is also a decrease in systematic arbitrariness (§3): the dark gray area for super ($\hat{\mathcal{W}}_1 = .014$) is smaller than the light gray area for simple ($\hat{\mathcal{W}}_1 = .063$) (§B.3, E.4).

For `HMDA` (Fig. 3b), simple similarly improves $\hat{\text{FNR}}$, but has a less beneficial effect on fairness ($\Delta\hat{\text{FNR}}$). However, note that since the baseline is the empirical expected error over thousands of RF models, the specific $\Delta\hat{\text{FNR}}$ is not necessarily attainable by any individual model. In this respect, simple has the benefit of actually obtaining a specific (ensemble) model that yields this disparity reliably in practice: $\Delta\hat{\text{FNR}} = 1.1\%$ is the mean over 10 simple ensembles. Notably, this is extremely low, even without applying traditional fairness techniques. Similar to `Old Adult`, simple exhibits high $\hat{\text{AR}}$, which decreases with super at the cost of higher error. $\hat{\text{FNR}}$ still improves for both $\boldsymbol{g}$ in comparison to the baseline, but the benefits are unequally applied: $\hat{\text{FNR}}_{\text{W}}$ has a larger benefit, so $\Delta\hat{\text{FNR}}$ increases slightly.

Abstention set error. As an example, the average $\hat{\text{Err}}$ in the `Old Adult` simple abstention set is close to 40% — compared to 17% for the RF baseline, and 8% for simple and 14% for super prediction sets (§E.4.2). As expected, beyond reducing arbitrariness, we abstain from predicting for many instances for which we also would have been more inaccurate (§4).

A trade-off. Our results support that there is indeed a trade-off between abstention rate and error (§4). This is because Algorithm 1 identifies low-$\hat{SC}$ instances for which ML prediction does a poor job, and abstains from predicting on them. Nevertheless, it may be infeasible for some applications to tolerate a high $\hat{\text{AR}}$. Thus the choice of κ and ensembling method should be considered a context-dependent decision.

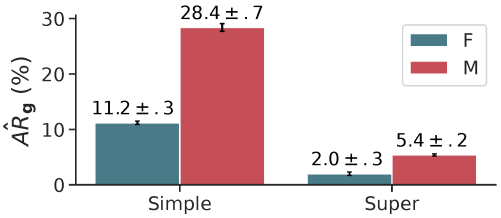
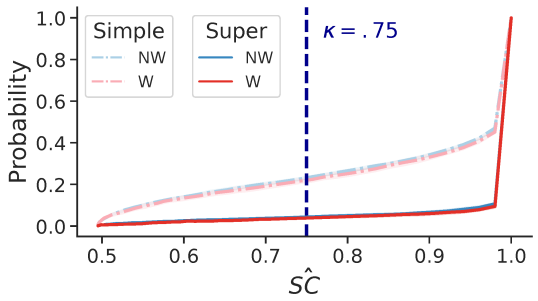


Figure 4: Group-specific abstention rates $\hat{\text{AR}}_{\boldsymbol{g}}$ for `Old Adult`. Super abstains less and more equally than simple.



	Baseline	Simple	Super
ΔFP̂R	$2.1 \pm 1.8\%$	$3.0 \pm 1.4\%$	$1.8 \pm 1.0\%$
FP̂R$_{NW}$	$14.7 \pm 1.3\%$	$11.4 \pm 1.0\%$	$12.9 \pm .8\%$
FP̂R$_{W}$	$12.6 \pm 1.3\%$	$8.4 \pm 1.0\%$	$11.1 \pm .6\%$

Figure 5: Algorithm 1, LR on COMPAS. $B = 101$, 10 train/test splits. Table shows mean FP̂R $\pm$ STD for individual models (Baseline) and ensembling methods' prediction sets.

Unequal abstention rates. When there is a high degree of systematic arbitrariness, ÂR can vary a lot by $\boldsymbol{g}$ (Fig. 4). With respect to improving SĈ, error, and fairness this may be a reasonable outcome: it is arguably better to abstain unevenly — deferring a final classification to non-ML decision processes — than to predict more inaccurately and arbitrarily for one group. More importantly, we rarely observe systematic arbitrariness; unequal ÂR is uncommon in practice (§6).

A Problem of Empirical Algorithmic Fairness

We also highlight results for COMPAS, 1 of the 3 most common fairness datasets (Fabris et al. 2022). Algorithm 1 is similarly very effective at reducing arbitrariness (Fig. 5), and is able to obtain state-of-the-art accuracy (Lin et al. 2020) with ΔFP̂R between $1.8 - 3\%$. Analogous results for German Credit indicate statistical equivalence in fairness metrics (§E.4.3, E.4.7). These low disparities do not cohere with much of the literature, which often reports much larger fairness violations (Larson et al. 2016, notably). However, most work on fair classification examines individual models, selected via cross-validation with a handful of random seeds (§2). Our results suggest that selecting between a few individual models in fair binary classification experiments is unreliable. When we instead estimate expected error by ensembling, we have difficulty reproducing unfairness in practice. Variance in the underlying models in $\hat{\mu}$ seems to be the culprit. The individual models we train exhibit radically different group-specific error rates. Our strategy of shifting focus to the overall behavior of $\hat{\mu}$ provides a solution: we not only mitigate arbitrariness, we also improve accuracy *and* usually average away most underlying, individual-model unfairness (§E.5).

6 Discussion and Related Work

In this paper, we advocate for a shift in thinking about *individual* models to the *distribution over possible models* in fair binary classification. This shift surfaces arbitrariness in underlying model decisions. We suggest a metric of *self-consistency* as a proxy for arbitrariness (§3) and an intuitive, elegantly simple extension of the classic bagging algorithm to mitigate it (§4). Our approach is tremendously effective with respect to improving SĈ, accuracy, and fairness metrics in practice (§5, §E.5).

Our findings contradict accepted truths in algorithmic fairness. For example, much work posits that there is an inherent analytical trade-off between fairness and accuracy (Corbett-Davies et al. 2017; Menon and Williamson 2018). Instead, our experiments complement prior work that disputes the practical relevance of this formulation (Rodolfa et al. 2021). We show it is in fact typically possible to achieve accuracy (via variance reduction) and close-to-fairness — and to do so *without* using fairness-focused interventions.

Other research also highlights the need for metrics beyond fairness and accuracy. Model multiplicity reasons about sets of models that have similar accuracy (Breiman 2001), but differ in underlying properties due to variance in decision rules (Black et al. 2022a; Marx et al. 2020). This work emphasizes developing criteria for selecting an *individual* model from that set. Instead, our work uses the *distribution over possible models* (with no normative claims about model accuracy or other criteria) to reason about arbitrariness (App C.3). Some related work considers the role of uncertainty and variance in fairness (Chen et al. 2018; Khan et al. 2023). Notably, Black et al. (2022b) concurrently investigates abstention-based ensembling, employing a strategy that (based on their choice of variance definition) ultimately does not address the arbitrariness we describe and mitigate (§C).

Most importantly, we take a comprehensive experimental approach missing from prior work. It is this approach that uncovers our alarming results: almost all tasks and settings demonstrate close-to or complete statistical equality in fairness metrics, after accounting for arbitrariness (§E.4). Old Adult (Fig. 3a) is one of two exceptions. These results hold for larger, newer datasets like HMDA, which we clean and release. Altogether, our findings indicate that variance is undermining the reliability of conclusions in fair binary classification experiments. It is worth revisiting all prior experiments that depend on cross validation or few models.

The future of fairness research. While the field has put forth numerous theoretical results about (un)fairness regarding single models — impossibility of satisfying multiple metrics (Kleinberg et al. 2017), post-processing individual models to achieve a particular metric (Hardt et al. 2016) — these results seem to miss the point. By examining individual models, arbitrariness remains latent; when we account for arbitrariness in practice, most measurements of unfairness vanish. We are not suggesting that there are no reasons to be concerned with fairness of ML models. We are not challenging the idea that actual, reliable violations of standard fairness metrics should be of concern. Instead, we are suggesting that common formalisms and methods for measuring fairness can lead to false conclusions about the degree to which such violations are happening in practice (§F). Worse, they can conceal a tremendous amount of arbitrariness, which should itself be an important concern when examining the social impact of automated decision-making.

Ethical Statement

This work raises important ethical concerns regarding the practice of fair-binary-classification research. We organize these concerns into several themes below.

Arbitrariness and legitimacy. On common research benchmarks, we show that many classification decisions are effectively arbitrary. Intuitively, this is unfair, but is a type of unfairness that largely has gone unnoticed in the algorithmic-fairness community. Such arbitrariness raises serious concerns about the legitimacy of automated decision-making. Fully examining these implications is the subject of current work that our team is completing. Complementing prior work on ML and arbitrariness (Creel and Hellman 2022; Cooper et al. 2022b), we are working on a law-review piece that clarifies the due process implications of arbitrariness in ML-decision outcomes. For additional notes on future work in this area, see Appendix F.

Misspecification, mismeasurement, and fairness. Much prior work has emphasized theoretical contributions and problem formulations for how to study fairness in ML. A common pattern is to study unequal model error rates between demographic subgroups in the available data. Typically, experimental validation of these ideas has relied on using just a handful of models. Our work shows that this is not empirically sound: it can lead to drawing unreliable conclusions about the degree of unfairness (defined in terms of error rates). Most observable unfairness seems due to inadequately modeling or measuring the role of variance in learned models on common benchmark tasks.

Other than indicating serious concerns about the rigor of experiments in fairness research, our findings suggest ethical issues about the role of mismeasurement in identifying and allocating resources to specific research problems (Jacobs and Wallach 2021). A lot of resources and research effort have been allocated to the study of these problem formulations. In turn, they have had profound social influence and impact, both in research and in the real world, with respect to how we reason broadly about fairness in automated decision-making. In response to the heavy investment in these ideas, many have noted that there are normative and ethical reasons why such formulations are inadequate for the task of aligning with more just or equitable outcomes in practice. Our work shows that normative and ethical considerations extend even further. Even if we take these formulations at face value in theory, they are very difficult to replicate in practice on common fairness benchmarks when we account for variance in predictions across trained models. When we perform due diligence with our experiments, we have trouble validating the hypothesis that popular ML-theoretical formulations of fairness are capturing a meaningful practical phenomenon.

This should be an incredibly alarming finding to anyone in the community that is concerned about the practice, not just the theory, of fairness research. Using bootstrapping, we observe serious problems with respect to the reliability of how fairness and accuracy are measured in work that relies on cross-validation of just a few models. We also find little empirical evidence of a trade-off between fairness and accuracy (another common formulation in the community), which complements prior work that has made similar observations (Rodolfa et al. 2021).

Project aims, reduction of scope. We emphasize that this was an unintended outcome of our original research objectives. We aimed to study arbitrariness as a latent issue in problem formulations that have to do with fair classification. This included broader methodological aims: studying many sources of non-determinism that could impact arbitrariness in practice (e.g., minibatching, example ordering). Instead, our initial results of close-to-fair expected performance for individual models made us refocus our work on issues of mismeasurement and fairness. We did not expect to find that dealing with arbitrariness would make almost all unfairness (again, as measured by common definitions) vanish in practice. Regardless of our intention, these results indicate the limits of theory in a domain that, by centering social values like fairness, cannot be separated from practice. We believe that problem formulations are only as good as they are useful. Based on our work, it is unclear how useful our existing formulations are if they do not bear out in experiments.

Reproducibility and project aims. In the course of this study, we also tried to reproduce the experiments of many well-cited theory-focused works. We almost always could not do so: code was almost always unavailable. We therefore pivoted from making reproducibility an explicit aspect of the present paper, which we will pursue in future work that focuses solely on reproducibility and fairness. Nevertheless, our work raises concerns about the validity of some of this past work. At the very least, past work that makes claims about preexisting unfairness in fairness benchmarks (in order to demonstrate that proposed methods provide improvements) should be subject to experimental scrutiny. Further along these lines, we believe that the novel findings we present here should have surfaced long ago. They likely would have surfaced if experimental contributions had been more evenly balanced with theoretical ones, or if exact Bayesian inference (rather than optimization) had been employed as the chosen algorithmic approach in the problem formulation.

The limits of prediction. Lastly, it has not escaped our notice that our results signal severe limits to prediction in social settings. It is true that our method performs reasonably well with respect to both fairness and accuracy metrics; however, arbitrariness is such a rampant problem, it is arguably unreasonable to assign these metrics much value in practice.

Acknowledgments

This work was done as part of an internship at Microsoft Research. A. Feder Cooper is supported by Prof. Christopher De Sa's NSF CAREER grant, Prof. Baobao Zhang, and Prof. James Grimmelmann. A. Feder Cooper is affiliated with The Berkman Klein Center for Internet & Society at Harvard University. The authors would like to thank The Internet Society Project at Yale Law School, Artificial Intelligence Policy and Practice at Cornell University, Jack Balkin, Emily Black, danah boyd, Sarah Dean, Fernando Delgado, Hoda Heidari, Ken Holstein, Jessica Hullman, Abigail Z. Jacobs, Meg Leta Jones, Michael Littman, Kweku Kwegyir-Aggrey, Rosanne Liu, Emanuel Moss, Kathy Strandburg, Hanna Wallach, and

Simone Zhang for their feedback.

References

Abu-Mostafa, Y. S.; et al. 2012. *Learning From Data: A Short Course*.

Barocas, S.; et al. 2019. *Fairness and Machine Learning: Limitations and Opportunities*.

Black, E.; et al. 2022a. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Black, E.; et al. 2022b. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.

Breiman, L. 1996. Bagging Predictors. *Machine Learning*.

Breiman, L. 1998. Arcing classifiers. *Annals of Statistics*.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*.

Chen, I.; et al. 2018. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems*.

Cooper, A. F.; and Abrams, E. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*.

Cooper, A. F.; et al. 2022a. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*.

Cooper, A. F.; et al. 2022b. Non-Determinism and the Lawlessness of Machine Learning Code. In *Proceedings of the 2022 Symposium on Computer Science and Law*.

Corbett-Davies, S.; et al. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Creel, K.; and Hellman, D. 2022. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*.

Efron, B.; and Tibshirani, R. 1997. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*.

Efron, B.; and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap*.

Fabris, A.; et al. 2022. Algorithmic Fairness Datasets: the Story so Far. *Data Mining and Knowledge Discovery*.

FFIEC 2017. 2017. HMDA Data Publication.

Friedler, S. A.; et al. 2016. On the (im)possibility of fairness.

Friedler, S. A.; et al. 2019. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Fuller, L. L. 1965. *The Morality of Law*.

Geman, S.; et al. 1992. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.*

Hardt, M.; et al. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*.

Khan, F. A.; et al. 2023. On Fairness and Stability: Is Estimator Variance a Friend or a Foe? arXiv:2302.04525.

Kleinberg, J. M.; et al. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference*.

Larson, J.; et al. 2016. How We Analyzed the COMPAS Recidivism Algorithm. Technical report, ProPublica.

Lin, Z. J.; et al. 2020. The limits of human predictions of recidivism. *Science Advances*.

Marx, C.; et al. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning*.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*.

Pleiss, G.; et al. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*.

Ramdas, A.; et al. 2015. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests.

Rodolfa, K.; et al. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*.

Stalnaker, R. 2006. On Logics of Knowledge and Belief. *Philosophical Studies*.

Wager, S. 2020. Cross-Validation, Risk Estimation, and Model Selection: Comment on a Paper by Rosset and Tibshirani. *Journal of the American Statistical Association*.