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

A. Feder Cooper,[*] **Katherine Lee,** **Madiha Zahrah Choksi,** **Solon Barocas,**

Christopher De Sa, **James Grimmelmann,** **Jon Kleinberg,** **Siddhartha Sen,**

Baobao Zhang

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 [4]. 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 [30, 44], 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." 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

---

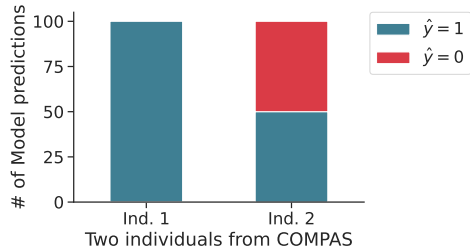[*]Corresponding author. `https://afedercooper.info`; `https://genlaw.org`

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).

2 exhibit so little confidence that their classification is effectively *arbitrary* [15, 16, 18]. 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* [34, 40, 50]. However, as is clear from the example learning process in Figure 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:

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 (Section 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 (Section 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** (Section 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 (Section 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.

2.1 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 group of *protected attributes* that we do not use for learning (e.g., race, gender).[2] 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.

---

[2] We examine the common setting in which $|\boldsymbol{g}| = 1$, and abuse notation by treating $\boldsymbol{g}$ like a scalar with $\mathbb{G} = \{0, 1\}$.

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 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 (Section 3).[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 $h_{\boldsymbol{D}_k}(\boldsymbol{x}) = \mathbf{1}\left[r_{\boldsymbol{D}_k}(\boldsymbol{x}) \geq \tau\right]$, 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* $f : \mathbb{Y} \times \mathbb{Y} \mapsto \mathbb{R}$. 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 (Definition 1), we consider the *expected error*, $\texttt{Err}(\mathcal{A},\mathbb{D},(\boldsymbol{x},\boldsymbol{g},o)) = \mathbb{E}_{\mathbf{D}}[f(o,\hat{y})|\mathbf{x} = \boldsymbol{x}]$.

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 [25]. These costs are related to the classifier threshold $\tau = \frac{C_{01}}{C_{01}+C_{10}}$, with $C_{01}, C_{10} \in \mathbb{R}^{+}$ (Appendix A.3). Common fairness metrics, such as Equality of Opportunity [34], further analyze error by computing disparities across group-specific error rates $\texttt{FPR}_g$ and $\texttt{FNR}_g$. For example, $\texttt{FPR}_g \triangleq p_\mu[r_{\mathbf{D}}(\mathbf{x}) \geq \tau | o=0, \mathbf{g}=\boldsymbol{g}] = p_\mu[\hat{y}=1|o=0,\mathbf{g}=\boldsymbol{g}]$. Model-specific $\texttt{FPR}_g$ and $\texttt{FNR}_g$ are further-conditioned on the dataset used in training, i.e., $\mathbf{D} = \boldsymbol{D}_k$.

2.2 Empirical approximation of the formulation

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 [11, 17, 38, e.g.]. **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 (Section 5). For more details, see Efron and Tibshirani [23, 24] and Wager [57].

To get around these reliability issues, one can *bootstrap*.[4] 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 $\hat{\boldsymbol{D}}_1, \hat{\boldsymbol{D}}_2, ..., \hat{\boldsymbol{D}}_B := \hat{\mathbb{D}}$. We use these replicates $\hat{\mathbb{D}}$ to approximate the learning process on $\mathbb{D}$ (Def. 1). We treat the resulting $\hat{h}_{\hat{\boldsymbol{D}}_1}, \hat{h}_{\hat{\boldsymbol{D}}_2}, ..., \hat{h}_{\hat{\boldsymbol{D}}_B}$ 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 (Appendix A.4).

3 Variance, Self-Consistency, and Arbitrariness

From these preliminaries, we can now pin down arbitrariness more precisely. We develop a quantitative proxy for measuring arbitrariness, called *self-consistency* (Section 3.2), which is derived from a definition of statistical *variance* between different model predictions (Section 3.1). We then illustrate how self-consistency is a simple-yet-powerful tool for revealing the role of arbitrariness in fair classification (Section 3.3). Next, we will introduce an algorithm to improve self-consistency (Section 4) and compute self-consistency on popular fair binary classification benchmarks (Section 5).

3.1 Arbitrariness resembles statistical variance

In Section 2, we discussed how common fairness metrics analyze error by computing false positive rate (FPR) and false negative rate (FNR). Another common way to formalize error is as a decomposition

---

[3]Model multiplicity has similar aims, but ultimately relocates the arbitrariness we describe to model selection (Section 6; Appendix C.3).

[4]We could use MCMC [60], but optimization is the standard tool that allows use of standard models in fairness.

of different statistical sources: *noise-*, *bias-*, and *variance*-induced error [2, 32]. To understand our metric for self-consistency (Section 3.2), 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 $h_{D_k} \sim \mu$. Variance is the error in the learning process that comes from training on different datasets $D_k \in \mathbb{D}$. In theory, we measure variance by imagining training all possible $h_{D_k} \sim \mu$, testing them all on the same test instance (x, g), and then quantifying how much the resulting classifications for (x, g) deviate *from each other*. More formally,

Definition 2. For all pairs of possible models $h_{D_i}, h_{D_j} \sim \mu \, (i \neq j)$, the **variance** for a test (x, g) is

$$\mathrm{var}\big(\mathcal{A}, \mathbb{D}, (x, g)\big) \triangleq \mathbb{E}_{h_{D_i} \sim \mu, h_{D_j} \sim \mu} \Big[f\Big(h_{D_i}(x), h_{D_j}(x) \Big) \Big].$$

We can approximate variance directly by using the bootstrap method (Section 2.2, Appendix B.1). For 0-1 and cost-sensitive loss with costs $C_{01}, C_{10} \in \mathbb{R}^+$ (Section 2.1), 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 (x, g),

$$\hat{\mathrm{var}}\big(\mathcal{A}, \hat{\mathbb{D}}, (x, g)\big) := \frac{1}{B(B-1)} \sum_{i \neq j} f\Big(\hat{h}_{\hat{D}_i}(x), \hat{h}_{\hat{D}_j}(x) \Big) = \frac{(C_{01} + C_{10}) B_0 B_1}{B(B-1)}. \tag{1}$$

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

3.2 Defining self-consistency from variance

It is clear from above that, in general, variance (1) is unbounded. We can always increase the maximum possible $\hat{\mathrm{var}}$ by increasing the magnitudes of our chosen C_{01} and C_{10}.[5] 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. For all pairs of possible models $h_{D_i}, h_{D_j} \sim \mu \, (i \neq j)$, the **self-consistency of the learning process** for a test (x, g) is

$$\mathrm{SC}\big(\mathcal{A}, \mathbb{D}, (x, g)\big) \triangleq \mathbb{E}_{h_{D_i} \sim \mu, h_{D_j} \sim \mu} \Big[h_{D_i}(x) = h_{D_j}(x) \Big] = p_{h_{D_i} \sim \mu, h_{D_j} \sim \mu} \big(h_{D_i}(x) = h_{D_j}(x) \big). \tag{2}$$

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.[6] Like variance, we can derive an empirical approximation of SC. Using the bootstrap method with $B = B_0 + B_1 > 1$,

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

For increasingly large B, $\hat{\mathrm{SC}}$ is defined on $[0.5 - \epsilon, 1]$ (Appendix B.3). Throughout, we use the shorthand *self-consistency*, but it is important to note that Definition 3 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 [31], Stalnaker [55]; Appendix B.3).

Interpretation. Definition 3 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 (Section 2.1), 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**

---

[5]Because $\tau = \frac{C_{01}}{C_{01} + C_{10}}$, for a given τ we can scale costs arbitrarily and have the same decision rule (Section 2.1). Relative, not absolute, costs affect the number of classifications B_0 and B_1.

[6](2) follows from it being equally likely to draw any two $D_i, D_j \in \mathbb{D}$ in a learning process (Appendix B.3).
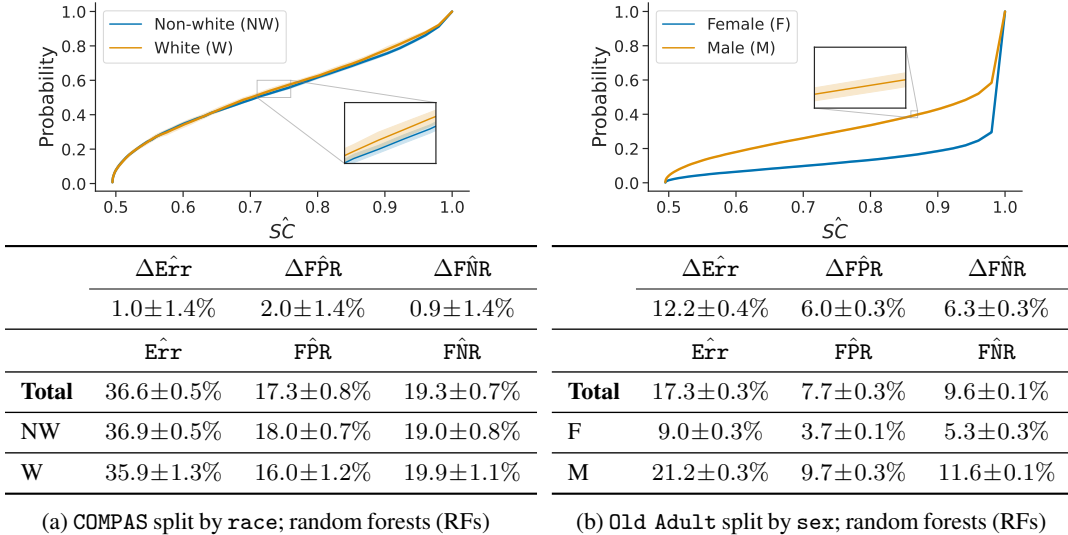
	$\Delta\hat{\mathtt{Err}}$	$\Delta\hat{\mathtt{FPR}}$	$\Delta\hat{\mathtt{FNR}}$
	1.0±1.4%	2.0±1.4%	0.9±1.4%

	$\hat{\mathtt{Err}}$	$\hat{\mathtt{FPR}}$	$\hat{\mathtt{FNR}}$
Total	36.6±0.5%	17.3±0.8%	19.3±0.7%
NW	36.9±0.5%	18.0±0.7%	19.0±0.8%
W	35.9±1.3%	16.0±1.2%	19.9±1.1%

	$\Delta\hat{\mathtt{Err}}$	$\Delta\hat{\mathtt{FPR}}$	$\Delta\hat{\mathtt{FNR}}$
	12.2±0.4%	6.0±0.3%	6.3±0.3%

	$\hat{\mathtt{Err}}$	$\hat{\mathtt{FPR}}$	$\hat{\mathtt{FNR}}$
Total	17.3±0.3%	7.7±0.3%	9.6±0.1%
F	9.0±0.3%	3.7±0.1%	5.3±0.3%
M	21.2±0.3%	9.7±0.3%	11.6±0.1%

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

Figure 2: $\hat{\mathtt{SC}}$ 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. $\hat{\mathtt{SC}}$ is effectively identical across subgroups g in `COMPAS`; `Old Adult` exhibits **systematic** differences in **arbitrariness** across g. Tables show mean ± STD of the relative disparities, e.g., $\Delta\hat{\mathtt{Err}}=|\hat{\mathtt{Err}}_0-\hat{\mathtt{Err}}_1|$ (top); and, the absolute $\hat{\mathtt{Err}},\hat{\mathtt{FPR}},\hat{\mathtt{FNR}}$, and $\hat{\mathtt{SC}}$, also broken down by g (bottom) (Appendix E).

indicates high variance, and vice versa. We derive empirical $\hat{\mathtt{SC}}$ (3) from $\hat{\mathtt{var}}$ (1) by leveraging observations about the definition of $\hat{\mathtt{var}}$ for 0-1 loss (Appendix B.3). While there are no costs C_{01}, C_{10} in computing (3), they still affect empirical measurements of $\hat{\mathtt{SC}}$. Because C_{01} and C_{10} affect τ (Section 2.1), they control the concrete number of B_0 and B_1, and thus the $\hat{\mathtt{SC}}$ 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 $\hat{\mathtt{SC}}$'s practical effects** (Section 5). In our experiments, we typically use $B=101$, which yields a $\hat{\mathtt{SC}}$ range of $[\approx 0.495, 1]$ in practice.[7]

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 [50]. 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 [12, 29]. 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., $\forall k, \hat{y}_k \neq o$; Section 6).

3.3 Illustrating self-consistency in practice

$\hat{\mathtt{SC}}$ enables us to evaluate arbitrariness in classification experiments. It is straightforward to compute $\hat{\mathtt{SC}}$ (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 $\hat{\mathtt{SC}}$ for individuals (Figure 1), we can also do so across sets of individuals. We plot the cumulative distribution (CDF) of $\hat{\mathtt{SC}}$ for the groups $\boldsymbol{g}$ in the test set (i.e., the x-axis shows the range of $\hat{\mathtt{SC}}$ for $B=101$, $[\approx 0.495, 1]$). In Figure 2, we provide illustrative examples from two of the most common fair classification benchmarks [26], `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 $\hat{h_1}, \hat{h_2}, ..., \hat{h_{101}}$ (Section 2.2). We repeat this process on 10 train/test splits, and the resulting confidence intervals (shown in the inset) indicate that our $\hat{\mathtt{SC}}$ estimates are stable. We group observations regarding these examples into two categories:

---

[7]Efron and Tibshirani [24] recommend $B \in \{50...200\}$.

Individual arbitrariness. Both CDFs show that $\hat{\mathrm{SC}}$ 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 $\hat{\mathrm{SC}}$ across the test set persist even though the 101 models exhibit relatively small average disparities $\Delta\hat{\mathrm{Err}}$, $\Delta\hat{\mathrm{FPR}}$, and $\Delta\hat{\mathrm{FNR}}$ (Figure 2a, bottom; Section 5.2). 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 (Section 1).

Systematic arbitrariness. We can also highlight $\hat{\mathrm{SC}}$ according to groups. The $\hat{\mathrm{SC}}$ plot for `Old Adult` shows that it is possible for the degree of arbitrariness to be *systematically worse* for a particular demographic $\boldsymbol{g}$ (Figure 2b). While the lack of $\hat{\mathrm{SC}}$ is not as extreme as it is for `COMPAS` (Figure 2a) — the majority of test instances exhibit over .9 $\hat{\mathrm{SC}}$ — 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 [52]. 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 (Appendix 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`, empirical $\hat{\mathcal{W}}_1=0.127$; for `COMPAS`, which does not show systematic arbitrariness, $\hat{\mathcal{W}}_1=0.007$.

4   Accounting for Self-Consistency

By definition, low $\hat{\mathrm{SC}}$ signals that there is high $\hat{\mathrm{var}}$ (Section 3.2). It is therefore a natural idea to use variance reduction techniques to improve $\hat{\mathrm{SC}}$ (and thus reduce arbitrariness).

As a starting point for improving $\hat{\mathrm{SC}}$, we perform variance reduction with Breiman [8]'s *bootstrap aggregation*, or *bagging*, ensembling algorithm. Bagging involves bootstrapping to produce a set of B models (Section 2.2), and then, for each test instance, producing an aggregated prediction $\hat{y}_A$, which takes the **majority vote** of the $\hat{y}_1,...,\hat{y}_B$ classifications. This procedure is practically effective for classifiers with high variance [8, 9]. **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 (Figure 1), where the classifiers in the ensemble are evenly split between classes. This means that bagging alone cannot overcome arbitrariness (Appendix D.1).

To remedy this, we add the option to abstain from prediction if $\hat{\mathrm{SC}}$ is low (Algorithm 1). A minor adjustment to (3) accounts for abstentions, and a simple proof follows that Algorithm 1 improves $\hat{\mathrm{SC}}$ (Appendix D). We bootstrap as usual, but pro-

---

Algorithm 1 $\hat{\mathrm{SC}}$ Ensembling with Abstention

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

1: $\hat{y}_A \coloneqq \mathrm{list}()$ $\quad\triangleright$ To store ensemble predictions
2: **for** $1...B$ **do**
3: $\quad$ $\boldsymbol{D}_B \leftarrow \mathrm{Bootstrap}\big((\boldsymbol{X},\boldsymbol{o})\big)$
4: $\quad$ $\triangleright$ $\hat{h}_{\boldsymbol{D}_B}$ can itself be a bagged model, with $\mathcal{A}$
5: $\quad\quad$ bagging on $\boldsymbol{D}_B$ as the dataset to bootstrap
6: $\quad$ $\hat{h}_{\boldsymbol{D}_B} \leftarrow \mathcal{A}(\boldsymbol{D}_B)$
7: $\quad$ $\hat{y}_A.\mathrm{append}\big(\hat{h}_{\boldsymbol{D}_B}(\boldsymbol{x}_{\text{test}})\big)$ $\quad\triangleright\hat{y}_A=[\hat{y}_1,...,\hat{y}_B]$
8: **end for**
9: **return** $\mathrm{Aggregate}(\hat{y}_A,\kappa)$

10: $\triangleright$ Returns κ-majority prediction or abstains
11: **function** $\mathrm{Aggregate}(\hat{y}_1,...,\hat{y}_B,\kappa)$
12: $\quad$ $\triangleright$ Compute $\hat{\mathrm{SC}}$ (3)
13: $\quad$ **if** $\mathrm{SelfConsistency}(\hat{y}_1,...,\hat{y}_B)\geq\kappa$
14: $\quad\quad$ **return** $\mathrm{argmax}_{y'\in\mathbb{Y}}\left[\sum_{i=1}^{B}\mathbf{1}[y'=\hat{y}_i]\right]$
15: $\quad$ **end if**
16: $\quad$ **return** `Abstain`
17: **end function**

---

duce a prediction $\hat{y}\in[0,1]$ for $\boldsymbol{x}$ only if $\boldsymbol{x}$ surpasses a user-specified minimum level κ of $\hat{\mathrm{SC}}$; otherwise, if an instance fails to achieve $\hat{\mathrm{SC}}$ 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 Appendix D) and 2) abstaining from instances that exhibit low $\hat{\mathrm{SC}}$ (thereby raising the overall amount of $\hat{\mathrm{SC}}$ for the prediction set, see Appendix D).

Further, since variance is a component of error (Section 3), variance reduction also tends to improve accuracy [8]. 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 (Section 5, Appendix E). So, while at first glance it may seem odd that our solution for arbitrariness is to *not*

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

(a) `Old Adult` split by `sex`

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

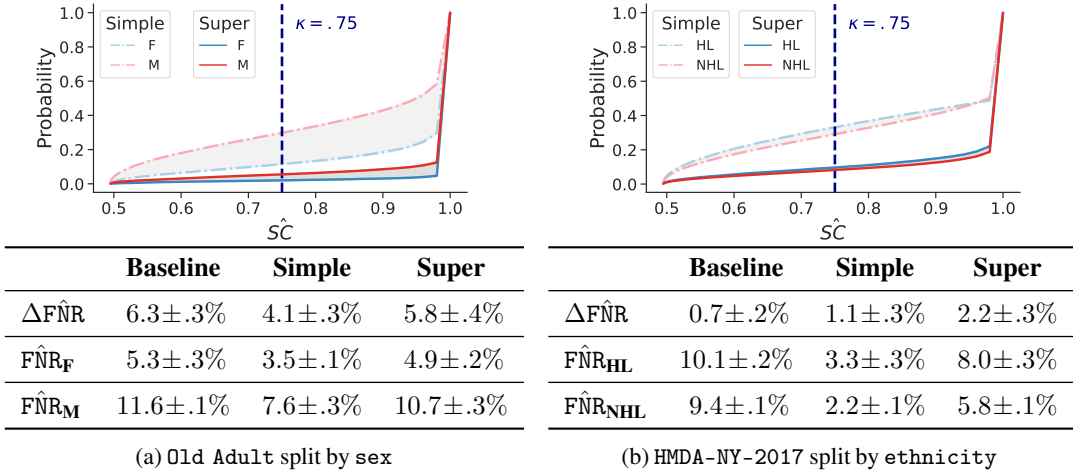(b) `HMDA-NY-2017` split by `ethnicity`

Figure 3: Algorithm 1: **simple** and **super ensembling** random forests (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 (Appendix E). To highlight systematic arbitrariness (Section 3.3), we shade in gray the area between group-specific $\hat{\text{SC}}$ CDFs for each method. An initial pass of variance reduction in **super** significantly decreases the systematic arbitrariness in `Old Adult`.

predict, it is worth noting that **we often would have predicted incorrectly on a large portion of the abstention set anyway** (Appendix 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 $\hat{\text{SC}}$. If overall $\hat{\text{SC}}$ 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 $\kappa = 0.75$ (Figure 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 $\hat{\text{SC}}$ (i.e., high $\hat{\text{var}}$) it can be beneficial to do an initial pass of variance reduction. We produce bagged classifiers using traditional bagging, but without abstaining (at Algorithm 1, lines 4-5); *then* we Aggregate using those bagged classifiers as the underlying models $\hat{h}$. The first round of bagging raises the overall $\hat{\text{SC}}$ 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 $\hat{\text{SC}}$ (Appendix D).[8] We also experiment with an Aggregate rule that averages the output probabilities of the underlying regressors $r_{\mathbf{D}_k}$, 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 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 (Appendix 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** (Section 5.2, Appendix 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

---

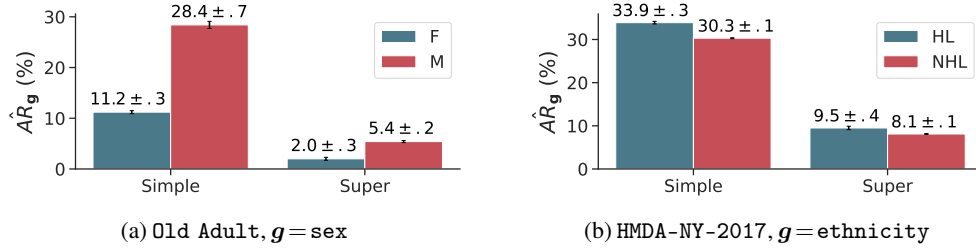[8]We could recursively super ensemble, but do not in this work.

(a) `Old Adult`, $g = $ `sex`   (b) `HMDA-NY-2017`, $g = $ `ethnicity`

Figure 4: Group-specific abstention rates $\hat{\text{AR}}_g$ for each algorithm. **Super ensembling** abstains less overall, and more equally than **simple ensembling**. `HMDA-NY-2017`, which exhibits less systematic arbitrariness than `Old Adult`(Figure 3), exhibits roughly equal abstention rates across subgroups.

(`HMDA`) 2007-2017 datasets [27] — and release them with a standalone, easy-to-use software package.[9] 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** (Section 5.1, Appendix E).

Presentation. To visualize Algorithm 1, we plot the CDFs of the $\hat{\text{SC}}$ 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 $\hat{\text{SC}}$ 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 ($\hat{\text{AR}}$)**.

We necessarily defer most of our results to Appendix E. In the main text, we exemplify two overarching themes: the effectiveness of both ensembling variants (Section 5.1), and how our results reveal shocking insights about reliability in fair binary classification research (Section 5.2). For all experiments, we illustrate Algorithm 1 with $\kappa = 0.75$, but note that κ is task-dependent in practice.

5.1 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 $\hat{\text{SC}}$ 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 $\hat{\text{SC}}$ for these RFs (Figure 3a). When we apply **simple** to these RFs, overall $\hat{\text{Err}}$ decreases (Appendix E), shown in part by the decrease in $\hat{\text{FNR}}_F$ and $\hat{\text{FNR}}_M$. 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 ($g = $ M, Figure 4a).

As expected, **super** improves overall $\hat{\text{SC}}$ through a first pass of variance reduction (Section 4). The $\hat{\text{SC}}$ CDF curves are brought down, indicating a lower proportion of the test set exhibits low $\hat{\text{SC}}$. Abstention rate $\hat{\text{AR}}$ is lower and more equal (Figure 4a); however, error, while still lower than the baseline RFs, has gone up for all metrics. There is also a decrease in systematic arbitrariness (Section 3.3): 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` (Figure 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 g in comparison to the baseline, but the benefits are unequally applied: $\hat{\text{FNR}}_W$ 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

---

[9]It is repeatedly argued that the field needs such datasets [19, e.g.]. `HMDA` meets this need, but is less commonly used. It requires engineering effort to manipulate — a barrier we remove.
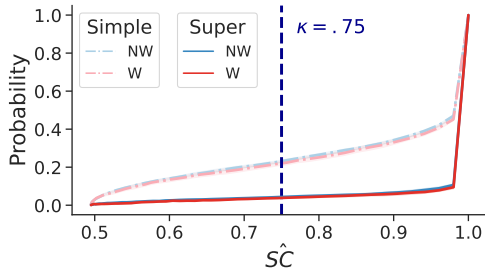
	Baseline	Simple	Super
$\Delta\hat{\text{FPR}}$	$2.1\pm1.8\%$	$3.0\pm1.4\%$	$1.8\pm1.0\%$
$\hat{\text{FPR}}_{\text{NW}}$	$14.7\pm1.3\%$	$11.4\pm1.0\%$	$12.9\pm.8\%$
$\hat{\text{FPR}}_{\text{W}}$	$12.6\pm1.3\%$	$8.4\pm1.0\%$	$11.1\pm.6\%$

Figure 5: Algorithm 1: **simple** and **super ensembling** logistic regression on `COMPAS`. $B = 101$, 10 train/test splits. Table shows mean $\hat{\text{FPR}} \pm$ STD for individual models (**Baseline**) and each ensembling method's prediction set. The $\hat{\text{SC}}$ CDFs are effectively identical across g.

sets (Appendix E.4.2). As expected, beyond reducing arbitrariness, we abstain from predicting for many instances for which we also would have been more inaccurate (Section 4).

A trade-off. Our results support that there is indeed a trade-off between abstention rate and error (Section 4). This is because Algorithm 1 identifies low-$\hat{\text{SC}}$ 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.

Unequal abstention rates. When there is a high degree of systematic arbitrariness, $\hat{\text{AR}}$ can vary a lot by g (Figure 4). With respect to improving $\hat{\text{SC}}$, 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 $\hat{\text{AR}}$ is uncommon in practice (Section 6).

5.2 A problem of empiriclal algorithmic fairness

We also highlight results for `COMPAS`, 1 of the 3 most common fairness datasets [26]. Algorithm 1 is similarly very effective at reducing arbitrariness (Figure 5), and is able to obtain state-of-the-art accuracy [45] with $\Delta\hat{\text{FPR}}$ between $1.8-3\%$. Analogous results for `German Credit` indicate statistical equivalence in fairness metrics (Appendices E.4.3 and E.4.7).

These low-single-digit disparities do not cohere with much of the literature on fair binary classification, which often reports much larger fairness violations [44, notably]. However, most work on fair classification examines individual models, selected via cross-validation with a handful of random seeds (Section 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 on these tasks exhibit radically different group-specific error rates (Appendix E.4.7). Our strategy of shifting focus to the overall behavior of the distribution $\hat{\mu}$ provides a solution: we not only mitigate arbitrariness, we **also improve accuracy** *and* **usually average away most underlying, individual-model unfairness** (Appendix 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 (Section 3) and an intuitive, elegantly simple extension of the classic bagging algorithm to mitigate it (Section 4). Our approach is tremendously effective with respect to improving $\hat{\text{SC}}$, accuracy, and fairness metrics in practice (Section 5, Appendix E).

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 [17, 48]. Instead, our experiments complement prior work that disputes the practical relevance of this formulation [53]. 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 calls attention to the need for metrics beyond fairness and accuracy. Model multiplicity reasons about sets of models that have similar accuracy [10], but differ in underlying properties due to variance in decision rules [7, 47, 58]. 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 selection criteria) to reason about arbitrariness (Appendix C.3). Some related work considers the role of uncertainty and variance in fairness [3, 5, 11, 39]. Notably, Black et al. [6] 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 (Appendix C). After our work, Ko et al. [41] built on prior work that studies fairness and variance in deep learning tasks [28, 51], and find that fairness emerges in deep ensembles (Appendix C.4).

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** (Appendix E.4). `Old Adult` (Figure 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.

What does this mean for fairness research?

While the field has put forth numerous theoretical results about (un)fairness regarding single models — impossibility of satisfying multiple metrics [40], post-processing individual models to achieve a particular metric [34] — 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 the fairness of machine-learning 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 (Appendix 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 [15, 18], 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 [37]. 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 [53].

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 baseline unfairness in fairness benchmarks (in order to demonstrate that proposed methods improve upon these baselines) should be subject to experimental scrutiny.

Further along these lines, **in our opinion, this project should not have been possible or necessary in 2022**. We believe that the novel findings we present here should have surfaced long ago, and likely would have surfaced if experimental contributions had been more evenly balanced with theoretical work.

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

References

[1] Daniel Abadi. Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story. *Computer*, 45(2):37–42, February 2012.

[2] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data: A Short Course*. AMLbook, USA, 2012.

[3] Junaid Ali, Preethi Lahoti, and Krishna P. Gummadi. Accounting for Model Uncertainty in Algorithmic Discrimination. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 336–345, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462630. URL `https://doi.org/10.1145/3461702.3462630`.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[5] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 401–413. Association for Computing Machinery, New York, NY, USA, 2021.

[6] Emily Black, Klas Leino, and Matt Fredrikson. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*, 2022.

[7] Emily Black, Manish Raghavan, and Solon Barocas. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533149. URL `https://doi.org/10.1145/3531146.3533149`.

[8] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, August 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350.

[9] Leo Breiman. Arcing classifiers. *Annals of Statistics*, 26:801–823, 1998.

[10] Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 08834237.

[11] Irene Chen, Fredrik D Johansson, and David Sontag. Why Is My Classifier Discriminatory? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[12] A. Feder Cooper and Ellen Abrams. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 46–54, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735.

[13] A. Feder Cooper, Karen Levy, and Christopher De Sa. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385534. URL `https://doi.org/10.1145/3465416.3483289`.

[14] A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde, and Christopher De Sa. Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

[15] A. Feder Cooper, Jonathan Frankle, and Christopher De Sa. Non-Determinism and the Lawlessness of Machine Learning Code. In *Proceedings of the 2022 Symposium on Computer Science and Law*, CSLAW '22, page 1–8, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392341. doi: 10.1145/3511265.3550446.

[16] A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 864–876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533150.

[17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874.

[18] Kathleen Creel and Deborah Hellman. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52 (1):26–43, 2022.

[19] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring Adult: New Datasets for Fair Machine Learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc., 2021.

[20] Pedro Domingos. A Unified Bias-Variance Decomposition and Its Applications. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, page 231–238, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

[21] Pedro Domingos. A unified bias-variance decomposition. Technical report, University of Washington, 2000. URL https://homes.cs.washington.edu/~pedrod/bvd.pdf.

[22] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1): 1 – 26, 1979. doi: 10.1214/aos/1176344552.

[23] Bradley Efron and Robert Tibshirani. Improvements on Cross-Validation: The 632+ Bootstrap Method. *Journal of the American Statistical Association*, 92(438):548–560, 1997. doi: 10.1080/01621459.1997.10474007.

[24] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

[25] Charles Elkan. The Foundations of Cost-Sensitive Learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, page 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[26] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. Algorithmic Fairness Datasets: the Story so Far. *Data Mining and Knowledge Discovery*, 36(6):2074–2152, September 2022. doi: 10.1007/s10618-022-00854-z.

[27] FFIEC 2017. HMDA Data Publication, 2017. URL https://www.consumerfinance.gov/data-research/hmda/historic-data/. Released due to the Home Mortgage Disclosure Act.

[28] Jessica Zosa Forde, A. Feder Cooper, Kweku Kwegyir-Aggrey, Chris De Sa, and Michael L. Littman. Model Selection's Disparate Impact in Real-World Deep Learning Applications, 2021. URL https://arxiv.org/abs/2104.00606.

[29] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016.

[30] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A Comparative Study of Fairness-Enhancing Interventions in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287589.

[31] Lon L. Fuller. *The Morality of Law*. Yale University Press, New Haven, 1965. ISBN 9780300010701.

[32] Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Comput.*, 4(1):1–58, January 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.1.1.

[33] Ulrike Grömping. South German Credit Data: Correcting a Widely Used Data Set. Technical report, Beuth University of Applied Sciences Berlin, 2019.

[34] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[35] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, USA, 2 edition, 2009.

[36] Jaakko Hintikka. *Knowledge and Belief*. Cornell University Press, 1962.

[37] Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901.

[38] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein Fair Classification. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. PMLR, 22–25 Jul 2020.

[39] Falaah Arif Khan, Denys Herasymuk, and Julia Stoyanovich. On Fairness and Stability: Is Estimator Variance a Friend or a Foe?, 2023.

[40] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[41] Wei-Yin Ko, Daniel D'souza, Karina Nguyen, Randall Balestriero, and Sara Hooker. FAIR-Ensemble: When Fairness Naturally Emerges From Deep Ensembling, 2023.

[42] Ron Kohavi. Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press, 1996.

[43] Eun Bae Kong and Thomas G. Dietterich. Error-Correcting Output Coding Corrects Bias and Variance. In Armand Prieditis and Stuart Russell, editors, *Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California, USA, July 9-12, 1995*, pages 313–321, 1995.

[44] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. Technical report, ProPublica, May 2016. URL https://www.prop ublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[45] Zhiyuan Jerry Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. The limits of human predictions of recidivism. *Science Advances*, 6(7):eaaz0652, 2020. doi: 10.1126/sciadv.aaz0652.

[46] Paul Mackun, Joshua Comenetz, and Lindsay Spell. Around Four-Fifths of All U.S. Metro Areas Grew Between 2010 and 2020. Technical report, United States Census Bureau, August 2021. URL https://www.census.gov/library/stories/2021/08/more-than-half-of-u nited-states-counties-were-smaller-in-2020-than-in-2010.html.

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

[48] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR, 23–24 Feb 2018.

[49] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.*, 12: 2825–2830, November 2011. ISSN 1532-4435.

[50] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On Fairness and Calibration. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus,

14

S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[51] Shangshu Qian, Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training. In *Advances in Neural Information Processing Systems*, volume 34, Red Hook, NY, USA, 2021. Curran Associates, Inc.

[52] Aaditya Ramdas, Nicolas Garcia, and Marco Cuturi. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests, 2015. URL `https://arxiv.org/abs/1509.02237`.

[53] K.T. Rodolfa, H. Lamba, and R Ghani. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3, 2021. doi: https://doi.org/10.1038/s42256-021-00396-x.

[54] Raymond M. Smullyan. Logicians Who Reason about Themselves. In *Proceedings of the 1986 Conference on Theoretical Aspects of Reasoning about Knowledge*, TARK '86, page 341–352, San Francisco, CA, USA, 1986. Morgan Kaufmann Publishers Inc. ISBN 0934613049.

[55] Robert Stalnaker. On Logics of Knowledge and Belief. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 128(1):169–199, 2006. ISSN 00318116, 15730883. URL `http://www.jstor.org/stable/4321718`.

[56] Brian Z. Tamanaha. *On the Rule of Law: History, Politics, Theory*. Cambridge University Press, Cambridge, 2004.

[57] Stefan Wager. Cross-Validation, Risk Estimation, and Model Selection: Comment on a Paper by Rosset and Tibshirani. *Journal of the American Statistical Association*, 115(529):157–160, 2020. doi: 10.1080/01621459.2020.1727235.

[58] Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive Multiplicity in Probabilistic Classification, 2022.

[59] I.C. Yeh and C.H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36:2473–2480, 2009.

[60] Ruqi Zhang, A. Feder Cooper, and Christopher De Sa. AMAGOLD: Amortized Metropolis Adjustment for Efficient Stochastic Gradient MCMC. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

Appendix Overview

The Appendix goes into significantly more detail than the main paper. It is organized as follows:

Appendix A: Extended Preliminaries

- A.1: Notes on notation and on our choice of terminology
- A.2: Constraints on our setup
- A.3: Costs and the classification decision threshold
- A.4: The bootstrap method

Appendix B: Additional Details on Variance and Self-Consistency

- B.1: Other statistical sources of error
- B.2: Our variance definition
- B.3: Deriving self-consistency from variance
    - B.3.1: Additional details on our choice of self-consistency metric

Appendix C: Related Work and Alternative Notions of Variance

- C.1: Defining variance in relation to a "main prediction"
- C.2: Why we choose to avoid computing the main prediction
    - C.2.1: The main prediction and cost-sensitive loss
- C.3: Putting our work in conversation with research on model multiplicity
- C.4: Concurrent work

Appendix D: Additional Details on Our Algorithmic Framework

- D.1: Self-consistent ensembling with abstention

Appendix E: Additional Experimental Results and Details for Reproducibility

- E.1: Hypothesis classes, datasets, and code
    - E.1.1: The standalone `HMDA` tookit
- E.2: Cluster environment details
- E.3: Details on motivating examples in the main paper
- E.4: Validating our algorithm in practice
    - E.4.1: `COMPAS`
    - E.4.2: `Old Adult`
    - E.4.3: `South German Credit`
    - E.4.4: `Taiwan Credit`
    - E.4.5: `New Adult - CA`
    - E.4.6: `HMDA`
    - E.4.7: Discussion of extended results for Algorithm 1
- E.5: Reliability and fairness metrics in `COMPAS` and `South German Credit`

Appendix F: Brief notes on future research

A Extended Preliminaries

A.1 Notes on notation and on our choice of terminology

Terminology. Traditionally, what we term "observed labels" o are often referred to instead as the "ground truth" or "correct" labels [2, 35, 43, e.g.]. We avoid this terminology because, as the work on label bias has explained, these labels are often unreliable or contested [12, 29].

Sets, random variables, and instances. We use bold non-italics letters to denote random variables (e.g., $\mathbf{x}$, $\mathbf{D}$), capital block letters to denote sets (e.g., $\mathbb{X}$, $\mathbb{Y}$), lower case italics letters to denote scalars (e.g., o), bold italics lower case letters to denote vectors (e.g., $\boldsymbol{x}$), and bold italics upper case to denote matrices (e.g., $\boldsymbol{D}_k$). For a complete example, $\boldsymbol{x}$ is an arbitrary instance's feature vector, $\mathbb{X}$ is the set representing the space of instances $\boldsymbol{x}$ ($\boldsymbol{x} \in \mathbb{X}$), and $\mathbf{x}$ is the random variable that can take on specific values of $\boldsymbol{x} \in \mathbb{X}$. We use this notation consistently, and thus do not always define all symbols explicitly.

A.2 Constraints on our setup

Our setup, per our definition of the learning process (Definition 1) is deliberately limited to studying the effects of variance due to changes in the underlying training dataset, with such datasets drawn from the same distribution. For this reason, Definition 1 does not include the data collection process or hyperparameter optimization (HPO), which can further introduce non-determinism to machine learning, and are thus assumed to have been already be completed.

Relatedly, variance-induced error can of course have other sources due to such non-determinism. For example, stochastic optimization methods, such as SGD and Adam, can cause fluctuations in test error; as, too, can choices in HPO configurations [14]. While each of these decision points is worthy of investigation with respect to their impact on fair classification outcomes, we aim to fix as many sources of randomness as possible in order to highlight the particular kind of arbitrariness that we describe in Sections 1 and 3. As such, we use the Limited-memory BFGS solver and fix our hyperparameters based on the results of an initial search (Section 5), for which we selected a search space through consulting related work such as Chen et al. [11].

A.3 Costs and the classification decision threshold

For reference, we provide a bit more of the basic background regarding the relationship between the classification decision threshold τ and costs of false positives FP (C_{01}) and false negatives FN (C_{10}). We visualize the loss as follows:

Table 1: Confusion matrix for cost-sensitive loss f, adapted from Elkan [25].

	$\hat{y}=0$	$\hat{y}=1$
$o=0$	TN: 0	FP: C_{01}
$o=1$	FN: C_{10}	TP: 0

0-1 loss treats the cost of different types of errors equally $C_{01} = C_{10} = 1$); false positives and false negatives are quantified as equivalently bad – they are *symmetric*; the case for which $C_{01} \neq C_{10}$ is *asymmetric* or *cost-sensitive*.

Altering the asymmetric of costs shifts the classification decision threshold τ applied to the underlying regressor $r_{\boldsymbol{D}_k}$. We can see this by examining the behavior of $r_{\boldsymbol{D}_k}$ that we learn. $r_{\boldsymbol{D}_k}$ estimates the probability of a each label given $\boldsymbol{x}$ (since we do not learn using $\boldsymbol{g}$), i.e., that we develop a good approximation of the distribution $p(\mathbf{y}|\mathbf{x})$. Ideally, $r_{\boldsymbol{D}_k}$ will be similar to the Bayes optimal classifier (for which the classification rule produces classifications y^* that yield the smallest weighted sum of the loss, where the weights are the probabilities of a particular label $\mathbf{y} = i$ for a given $(\boldsymbol{x}, \boldsymbol{g})$, i.e., sums over

$$p(\mathbf{y}=i|\mathbf{x}=\boldsymbol{x})\, f(i,y'). \tag{4}$$

For binary classification, the terms of (4) in the sum for a particular y' yield two cases:

- $i = y'$: By definition, $f(i,y') = 0$; therefore, $(4) = 0$.
- $i \neq y'$: By definition, $f(i,y') = C_{01}$ or $\ell(i,y') = C_{10}$. So, (4) will weight the cost by the probability $p(\mathbf{y} = i | \mathbf{x} = \boldsymbol{x})$.

We can therefore break down the Bayes optimal classifier into the following decision rule, which we hope to approximate through learning. For an arbitrary $(\boldsymbol{x}, \boldsymbol{g})$ and $\mathbb{Y} = \{0,1\}$,

$$\min\Big(\overbrace{\underbrace{p(\mathbf{y}{=}0|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of FP}} \times C_{01} + \underbrace{p(\mathbf{y}{=}1|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of TP}} \times 0}^{\text{Weighted cost of predicting positive (1) class}}, \overbrace{\underbrace{p(\mathbf{y}{=}0|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of TN}} \times 0 + \underbrace{p(\mathbf{y}{=}1|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of FN}} \times C_{10}}^{\text{Weighted cost of predicting negative (0) class}}\Big)$$

$$= \min\Big(\underbrace{p(\mathbf{y}{=}0|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of FP}} \times C_{10}, \underbrace{p(\mathbf{y}{=}1|\mathbf{x}{=}\boldsymbol{x})}_{\text{Probability of FN}} \times C_{10}\Big)$$

That is, to predict label 1, the cost of mis-predicting 1 (i.e., the cost of a false positive FP) must be be smaller than the cost of mis-predicting 0 (i.e, the cost of a false negative FN). In binary classification $p(\mathbf{y}|\mathbf{x} = \boldsymbol{x}) = p(\mathbf{y} = 1|\mathbf{x} = \boldsymbol{x}) + p(\mathbf{y} = 0|\mathbf{x} = \boldsymbol{x}) = 1$. So, we can assign $p(\mathbf{y} = 1|\mathbf{x} = \boldsymbol{x}) = \tau$ and $p(\mathbf{y} = 0|\mathbf{x} = \boldsymbol{x}) = 1 - \tau$, and rewrite the above as

$$\min\Big((1-\tau)C_{01}, \tau C_{10}\Big). \tag{5}$$

The decision boundary is the case for which both of the arguments to $\min$ in (5) are equivalent (i.e., the costs of predicting a false positive and a false negative are equal), i.e.,

$$(1-\tau)C_{01} = \tau C_{10} \Rightarrow \tau = \frac{C_{01}}{C_{01} + C_{10}}, \text{ so,}$$

$$h_{\boldsymbol{D}_k}(\boldsymbol{x}) = \mathbf{1}[r_{\boldsymbol{D}_k}(\boldsymbol{x}) \geq \tau] = \begin{cases} 1, & \text{if } p(\mathbf{y} = 1|\mathbf{x} = \boldsymbol{x}) \geq \tau = \frac{C_{01}}{C_{01} + C_{10}} \\ 0, & \text{otherwise.} \end{cases}$$

For 0-1 loss, in which $C_{01} = C_{10} = 1$, τ evaluates to $\frac{1}{2}$. If we want to model asymmetric costs, then we need to change this decision threshold to account for which type of error is more costly. For example, let us say that false negatives are more costly than false positives, with $C_{01} = 1$ and $C_{10} = 3$. This leads to a threshold of $\frac{1}{4}$, which biases $h_{\boldsymbol{D}_k}$ toward choosing the (generally cheaper to predict/more conservative) positive class.

A.4 The bootstrap method

In the bootstrap method, we treat each dataset $\hat{\boldsymbol{D}}_k \in \hat{\mathbb{D}}$ as equally likely. For each set aside test example $(\boldsymbol{x}, \boldsymbol{g}, o)$, we can approximate $\text{Err}(\mathcal{A}, \mathbb{D}, (\boldsymbol{x}, \boldsymbol{g}, o))$ empirically by computing

$$\hat{\text{Err}}\big(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g}, o)\big) = \frac{1}{B}\sum_{i=1}^{B}\ell\big(o, \hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x})\big) \tag{6}$$

for a concrete number of replicates B. We estimate overall error $\hat{\text{Err}}(\mathcal{A}, \hat{\mathbb{D}})$ for the test set by additionally summing over each example instance $(\boldsymbol{x}, \boldsymbol{g}, o)$, which we can further delineate into $\hat{\text{FPR}}$ and $\hat{\text{FNR}}$, or into group-specific $\hat{\text{Err}}_g$, $\hat{\text{FPR}}_g$, and $\hat{\text{FNR}}_g$ by computing separate averages according to $\boldsymbol{g}$.

The bootstrap method exhibits less variance than cross-validation, but can be biased — in particular, pessimistic — with respect to estimating expected error. To reduce this bias, one can follow our setup in Definition 1, which splits into train and test sets before resampling. For more information comparing the two methods, see Efron and Tibshirani [23, 24]. Further, recent work shows that, in relation to studying individual models, CV is in fact asymptotically uninformative regarding expected error [57].

B Additional Details on Variance and Self-Consistency

In this appendix, we provide more details on other types of statistical error (Appendix B.1), on variance (Appendix B.2) and self-consistency (Appendix B.3). Following this longer presentation of our

metrics, we then provide some additional information on other definitions of variance that have been used in work on fair classification, and contextualize issues with these definitions that encouraged us to deviate from them in order to derive our definition of self-consistency (Appendix C).

B.1 Other statistical sources of error

Noise. Noise is traditionally understood as *irreducible error*; it is due to inherent randomness in the data, which cannot be captured perfectly accurately by a deterministic decision rule h_{D_k}. Notably, noise is an aspect of the data collection pipeline, not the learning process (Definition 1). It is *irreducible* in the sense that it does not depend on our choice of training procedure $\mathcal{A}$ or how we draw datasets for training from $\mathbb{D}$, either in theory or in practice. Heteroskedastic noise across demographic groups is often hypothesized to be a source of unfairness in machine learning [11, 12]. Importantly, albeit somewhat confusingly, this is commonly referred to as label bias, where "bias" connotes discrimination, as opposed to the statistical bias that we mention here.

Unlike noise, bias and variance are traditionally understood as sources of epistemic uncertainty. These sources of error are *reducible* because they are contingent on the modeling choices we make in the learning process; if we knew how to model the task at hand more effectively, in principle, we could reduce bias and variance error.

Bias. Within the amount of reducible error, bias reflects the error associated with the chosen hypothesis class $\mathbb{H}$, and is therefore governed by decisions concerning the training procedure $\mathcal{A}$ in the learning process (Definition 1). This type of error is persistent because it takes effect at the level of possible models in $\mathbb{H}$; in expectation, all models $h_{D_k} \in \mathbb{H}$ have the same amount of bias-induced error.

Whereas variance depends on stochasticity in the underlying training data, noise and bias error are traditionally formulated in relation to the Bayes optimal classifier — the best possible classifier that machine learning could produce for a given task [2, 20, 32]. Since the Bayes optimal classifier is typically not available in practice, we often cannot estimate noise or bias directly in experiments.

Of the three types of statistical error, it is only variance that seems to reflect the intuition in Figure 1 concerning the behavior of different possible models h_{D_k}. This is because noise is a property of the data distribution; for a learning process (Definition 1), in expectation we can treat noise error as constant. Bias can similarly be treated as constant for the learning process: It is a property of the chosen hypothesis class $\mathbb{H}$, and thus is in expectation the same for each $h_{D_k} \in \mathbb{H}$. In Figure 1, we are keeping the data distribution constant and $\mathbb{H}$ constant; we are only changing the underlying subset of training data to produce different models h_{D_k}.

B.2 Our variance definition

We first provide a simple proof that explains the simplified version for our empirical approximation for variance in (1).

Proof. For the models $\{h_{D_b}\}_{b=1}^{B}$ that we produce, we denote $\hat{\mathbb{Y}}$ to be the multiset of their predictions on $(\boldsymbol{x}, \boldsymbol{g})$. $|\hat{\mathbb{Y}}| = B = B_0 + B_1$, where B_0 and B_1 represent the counts of 0 and 1-predictions, respectively. We also set the cost of false positives to be $f(0,1) = C_{01}$ and the cost of false negatives to be $f(1,0) = C_{10}$.

Looking at the sum in vâr (i.e., $\sum_{i \neq j}$), each of the B_0 0-predictions will get compared to the other $B_0 - 1$ 0-predictions and to the B_1 1-predictions. By the definition of f, each of the $B_0 - 1$ computations of $f(0,0)$ evaluates to 0 and each of the B_1 computations of $f(0,1)$ evaluates to C_{01}. Therefore, the B_0 0-predictions contribute

$$B_0 \times \big[\big(0 \times (B_0 - 1) \big) + C_{01} \times B_1 \big] = C_{01} B_0 B_1$$

19

to the sum in vâr, and, by similar reasoning, $B_1 \times \left[\left(0 \times (B_1 - 1) \right) + C_{10} \times B_0 \right] = C_{10} B_0 B_1$. It follows that the total sum in vâr is

$$\sum_{i \neq j} f\left(\hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x}), \hat{h}_{\hat{\boldsymbol{D}}_j}(\boldsymbol{x}) \right) = (C_{01} + C_{10}) B_0 B_1. \text{ Therefore}$$

$$\overbrace{\frac{1}{B(B-1)} \sum_{i \neq j} f\left(\hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x}), \hat{h}_{\hat{\boldsymbol{D}}_j}(\boldsymbol{x}) \right)}^{\text{vâr}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g}) \right)} = \overbrace{\frac{(C_{01} + C_{10}) B_0 B_1}{B(B-1)}}^{(1)}$$

$\square$

The effect of τ on variance. As discussed in Appendix A.3, C_{01} and C_{10} can be related to changing τ applied to $r_{\boldsymbol{D}_k}$ to produce classifier $h_{\boldsymbol{D}_k}$. We analyze the range of minimal and maximal empirical variance by examining the behavior of $B \to \infty$, i.e.,

$$\lim_{B \to \infty} \frac{(C_{01} + C_{10}) B_0 B_1}{B(B-1)}. \tag{7}$$

Minimal variance. Either B_0 or B_1 (exclusively, since $B_0 + B_1 > 1$) will be $= 0$, with the other being $= B$, making (7) equivalent to

$$\lim_{B \to \infty} \frac{(C_{01} + C_{10}) \times 0}{B(B-1)} = 0, \text{regardless of the value of } C_{01} + C_{10}.$$

Maximal variance. B_0 will represent half of B, with B_1 representing the other half. More particularly, $B_0 = \frac{B}{2}$ and $B_1 = \frac{B}{2}$; or, without loss of generality, $B_0 = \frac{B-1}{2}$ and $B_1 = \frac{B+1}{2}$. This means that

$$\frac{(C_{01} + C_{10}) B_0 B_1}{B(B-1)} = \frac{(C_{01} + C_{10}) (\frac{B}{2})^2}{B(B-1)} \qquad \left(\text{Or, } = \frac{(C_{01} + C_{10}) (\frac{B-1}{2})(\frac{B+1}{2})}{B(B-1)} \right)$$

$$= \frac{(C_{01} + C_{10}) (\frac{B^2}{4})}{B^2 - B} \qquad \left(\text{Or, } = \frac{(C_{01} + C_{10}) (\frac{(B^2-1)}{4})}{B(B-1)}; \text{ it will not matter in the limit} \right)$$

$$= \frac{(C_{01} + C_{10}) B^2}{4B^2 - 4B}.$$

And, therefore,

$$\lim_{B \to \infty} \frac{(C_{01} + C_{10}) B^2}{4B^2 - 4B} = \frac{C_{01} + C_{10}}{4}. \tag{8}$$

It follows analytically that variance will be in the range $[0, \frac{C_{01} + C_{10}}{4})$. However, empirically, for concrete B, $\text{vâr}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g}) \right) \to [0, \frac{C_{01} + C_{10}}{4} + \epsilon]$, for smaller positive ϵ as the number of models B increases. The maximal variance will better approximate $\frac{C_{01} + C_{10}}{4}$ as B gets larger, but will be $> \frac{C_{01} + C_{10}}{4}$. For example, for 0-1 loss $\frac{C_{01} + C_{10}}{4} = \frac{2}{4} = 0.5$. For $B = 100$, the maximal $\text{vâr}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g}) \right) = \frac{2 \times 50 \times 50}{100 \times 99} = \frac{50}{99} \approx .505$.

B.3 Deriving self-consistency from variance

In this appendix, we describe the relationship between variance (Definition 2) and self-consistency (Definition 3) in more detail, and show that $\hat{\text{SC}}\left(\mathcal{A}, \{\boldsymbol{D}_b\}_{b=1}^B, (\boldsymbol{x}, \boldsymbol{g}) \right) \to [0.5 - \epsilon, 1]$ for small positive ϵ as the number of models B increases.

Proof. Note that, by the definition of 0-1 loss, $C_{01} = C_{10} = 1$, so

$$\text{vâr}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g}) \right)_{\text{0-1}} = \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{1}[h_{\boldsymbol{D}_i}(\boldsymbol{x}) \neq h_{\boldsymbol{D}_j}(\boldsymbol{x})] = \frac{2 B_0 B_1}{B(B-1)}. \tag{9}$$

By the definition of the indicator function $\mathbf{1}$,

$$1 = \frac{1}{B(B-1)} \sum_{i \neq j} \left[\overbrace{\mathbf{1}[h_{\boldsymbol{D}_i}(\boldsymbol{x}) \neq h_{\boldsymbol{D}_j}(\boldsymbol{x})]}^{\text{From } \hat{\text{var}}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\right)_{0\text{-}1}} + \overbrace{\mathbf{1}[h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})]}^{\text{From } \hat{\text{SC}}\left(\mathcal{A}, \{\hat{\boldsymbol{D}}_b\}_{b=1}^{B}, (\boldsymbol{x}, \boldsymbol{g})\right)} \right]$$

$$= \overbrace{\frac{2B_0 B_1}{B(B-1)}}^{(9)} + \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{1}[h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})].$$

Therefore, rearranging,

$$\hat{\text{SC}}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\right) = \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{1}[h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})] = 1 - \frac{2B_0 B_1}{B(B-1)}.$$

$\square$

We note that $\hat{\text{SC}}$ (2) is independent of specific costs C_{01} and C_{10}. Nevertheless, the choice of decision threshold τ will of course impact the values of B_0 and B_1 in practice. In turn, this will impact the degree of self-consistency that a learning process exhibits empirically. In short, the measured degree of self-consistency in practice will depend on the choice of f. Further, following an analysis similar to what we can show that $\hat{\text{SC}}$ will be a value in $[0.5+\epsilon, 1]$, for small positive ϵ. This reality is reflected in the results that we report for our experiments, for which $B = 101$ yields minimal $\hat{\text{SC}} \approx 0.495$.

Cost-independence of self-consistency   Intuitively, *self*-consistency of a learning process is a relative metric; it is a quantity that is measured relative to the learning process. We therefore conceive of it as a metric that is normalized with respect to the learning process (Definition 1). Such a process can be maximally 100% self-consistent, but it does not make sense for it to be more than that (reflected by the maximum value of 1).

In contrast, as discussed in Appendix B, variance can measure much greater than 1, depending on the magnitude of the sum of the costs C_{01} and C_{10}, in particular, for $C_{01} + C_{10} > 4$ (8). However, it is not necessarily meaningful to compare the magnitude of variance across classifiers. Recall that the effect of changing costs C_{01} and C_{10} corresponds to a change in the binary classification decision threshold, with $\tau = \frac{C_{01}}{C_{01} + C_{10}}$. It is the *relative* costs that change the decision threshold; not the costs themselves. For example, the classifier with costs $C_{01} = 1$ and $C_{10} = 3$ is equivalent to the classifier with costs $C_{01} = \frac{1}{2}$ and $C_{10} = \frac{3}{2}$ (for both, $\tau = \frac{1}{4}$), but the former would measure a larger magnitude for variance.

It is this observation that grounds our cost-independent definition of self-consistency in Section 3 and Appendix B.3. Given the fact that the magnitude of variance measurements can complicate our comparisons of classifiers, as discussed above, we focus on the part of variance that encodes information about arbitrariness in a learning process: its measure of (dis)agreement between classification decisions that result from changing the training dataset. We could alternatively conceive of self-consistency as the additive inverse of normalized variance, but this is more complicated because it would require a computation that depends on the specific costs, $\hat{\text{var}}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\right)_{\text{normalized}} = \frac{\hat{\text{var}}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\right)}{\hat{\text{var}}\left(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})\right)_{\text{max}}}$.

B.3.1   Additional details on our choice of self-consistency metric

Terminology.   In logic, the idea of consistent belief has to do with ensuring that we do not draw conclusions that contradcit each other. This is much like the case that we are modeling with self-consistency — the idea that underlying changes in the dataset can lead to predictions that are directly in contradiction [36, 54, 55]. Ideas of consistency in legal rules have a similar flavor; legal rules should not contradict each other; legal judgments should not contradict each other (this is at least an aspiration for the law, based on common ideas in legal theory [31, 56]. For both of these reasons, the term "consistent" has a natural mapping to our usage of it in this paper. This is especially true in the legal theory case, given that inconsistency in the law is often considered arbitrary and a source of discrimination.

We nevertheless realize that the word "consistent" is overloaded with many meanings in statistics and different subfields computer science like distributed computing [1, 60, e.g.,]. Nevertheless, due to the clear relationship between our purposes concerning arbitrariness and discrimination, and definitions in logic and the law, we believe that it is the most appropriate term for our work.

Quantifying systematic arbitrariness. We depict *systematic arbitrariness* using the Wasserstein-1 distance [52]. This is the natural distance for us to consider because it has a closed form when being applied to CDFs. For our purposes, it should be interpreted as computing the total disparity in self-consistency by examining all possible self-consistency levels κ at once.

Formally,[10] for two groups $g = 0$ and $g = 1$ with respective SC CDFs F_0 and F_1,

$$\mathcal{W}_1 = \int_{\mathbb{R}} |F_0(\kappa) - F_1(\kappa)| \, d\kappa.$$

For self-consistency, which we have defined on $[0.5, 1]$, this is just

$$\mathcal{W}_1 = \int_{0.5}^{1} |F_0(\kappa) - F_1(\kappa)| \, d\kappa.$$

Empirically, we can approximate this with

$$\hat{\mathcal{W}}_1 := \frac{1}{|\hat{\mathbb{K}}|} \sum_{\hat{\mathbb{K}}} |\hat{F}_0(\hat{\kappa}) - \hat{F}_1(\hat{\kappa})|, \ \ \text{where } \hat{\mathbb{K}} = \left\{ 1 - \frac{2B_0 B_1}{B(B-1)} \, \middle| \, B_0 \in \{0...B\} \wedge B_1 \in \{0...B\} \wedge B_0 + B_1 = B \right\}.$$

We typically set $B = 101$, and thus

$\hat{\mathbb{K}} = [0.49505, 0.49545, 0.49624, 0.49743, 0.49901, 0.50099, 0.50337, 0.50614, 0.50931, 0.51287,$
  $0.51683, 0.52119, 0.52594, 0.53109, 0.53663, 0.54257, 0.54891, 0.55564, 0.56277, 0.57030,$
  $0.57822, 0.58653, 0.59525, 0.60436, 0.61386, 0.62376, 0.63406, 0.64475, 0.65584, 0.66733,$
  $0.67921, 0.69149, 0.70416, 0.71723, 0.73069, 0.74455, 0.75881, 0.77347, 0.78851, 0.80396,$
  $0.81980, 0.83604, 0.85267, 0.86970, 0.88713, 0.90495, 0.92317, 0.94178, 0.96079, 0.9802, 1.0],$

which we use to produce our CDF plots.

When measuring systematic arbitrariness with abstention, we set the probability mass for $< \kappa$ to 0 it. This makes sense because we are effectively re-defining the $\hat{\text{SC}}$ CDFs to not include instances that exhibit below a minimal amount of $\hat{\text{SC}}$. This also makes comparing systematic arbitrariness across CDFs for different interventions more interpretable. It allows us to keep the number of experimental samples for the empirical CDF measures constant when computing averages, so abstaining would then always have the effect of decreasing systematic arbitrariness. If we did not do this, because the Wasserstein-1 distance is an average, changing the set $\hat{\mathbb{K}}$, of course, would change the amount of Wasserstein-1 distance — possibly leading to a relative *increase* (if there are greater discrepancies between g-condition CDF curves at $\geq \kappa$).

C Related Work and Alternative Notions of Variance

As noted in Section 6, prior work that discusses variance and fair classification often relies on the definition of variance from Domingos [20]. We deviate from prior work and provide our own definition for two reasons: 1) variance in Domingos [20, 21] does not cleanly extend to cost-sensitive loss, and 2) the reference point for measuring variance in Domingos [20, 21] — the *main prediction* — can be unstable/ brittle in practice. We start by explaining the Domingos [20, 21] definitions, and then use these definitions to support our rationale.

C.1 Defining variance in relation to a "main prediction"

To begin, we restate the definitions from Domingos [20, 21] concerning the expected model (called the *main predictor*). We change the notation from Domingos to align with our own, as we believe these changes provide greater clarity concerning meaning, significance, and consequent takeaways.

---

[10]We consider the Wasserstein distance for one-dimensional distributions. More generally, the p-th Wasserstein distance for such distributions, $\mathcal{W}_p$, requires the inverse CDFs to be well-defined (i.e., the CDFs need to be strictly monotonic). This is fine to assume for our purposes. We have to relax the formal definition of the Wasserstein distance, anyway, when we estimate it in practice with a discrete number of samples.

Nevertheless, these definitions for quantifying error are equivalent to those in Domingos [21], and they fundamentally depend on human decisions for setting up the learning process.

Domingos [20, 21] define predictive variance in relation to this single point of reference. This reference point captures the general, expected behavior of models that could be produced by the chosen learning process. We can think of each prediction of this point of reference as the "central tendency" of the predictions made by all possible models in μ for $(\boldsymbol{x}, \boldsymbol{g})$. Formally,

Definition 4. The **main prediction** $\hat{y}$ is the prediction value $y' \in \mathbb{Y}$ that generates the minimum average loss with respect to all of the predictions $\hat{y} \in \hat{\mathbb{Y}}$ generated by the different possible models in μ. It is defined as the expectation over training sets $\mathbb{D}$ for a loss function f, given an example instance $(\boldsymbol{x}, \boldsymbol{g})$. That is,

$$\overline{y} = \underset{y'}{\arg\min} \mathbb{E}_{\mathbf{D}}[f(\hat{y}, y') | \mathbf{x} = \boldsymbol{x}, \mathbf{g} = \boldsymbol{g}]. \tag{10}$$

The *main predictor* $\overline{h} : \mathbb{X} \to \mathbb{Y}$ produces the main prediction $\overline{y}$ for each $(\boldsymbol{x}, \boldsymbol{g})$.

What (10) evaluates to in practice of course depends on the loss function f. For squared loss, the main prediction is defined as the mean prediction of all the $h_{\boldsymbol{D}_k}$ [20, 43]. Following Kong and Dietterich [43], for 0-1 loss Domingos [20] defines the main prediction as the mode/majority vote — the most frequent prediction for an example instance $(\boldsymbol{x}, \boldsymbol{g})$. We provide a more formal discussion of why this is the case when we discuss problems with the main prediction for cost-sensitive loss (Appendix C.2). Domingos [20, 21] then define variance in relation to specific models $h_{\boldsymbol{D}_k}$ and the main predictor $\overline{h}$:

Definition 5. The *variance*-induced error for fresh example instance $(\boldsymbol{x}, \boldsymbol{g})$ is
$$\mathtt{var}(\mathcal{A}, \mathbb{D}, (\boldsymbol{x}, \boldsymbol{g})) = \mathbb{E}_{\mathbf{D}}[f(\overline{y}, \hat{y}) | \mathbf{x} = \boldsymbol{x}, \mathbf{g} = \boldsymbol{g}],$$
where $\overline{y} = \overline{h}(\boldsymbol{x})$ is the main prediction and the $\hat{y}$ are the predictions for the different $h_{\boldsymbol{D}_k} \sim \mu$.

That is, for a specific $(\boldsymbol{x}, \boldsymbol{g})$, it is possible to compare the individual predictions $\hat{y} = h_{\boldsymbol{D}_k}(\boldsymbol{x})$ to the main prediction $\overline{y} = \overline{h}(\boldsymbol{x})$. Using the main prediction as a reference point, one can compute the extent of disagreement of individual predictions with the main prediction as a source of error. It is this definition (Definition 5) that prior work on fair classification tends to reference when discussing variance [6, 11]. However, as we discuss in more detail below (Appendix C.2), many of the theoretical results in Chen et al. [11] follow directly from the definitions in Domingos [20], and the experiments do not actually use those results in practice. Black et al. [6], in contrast, presents results that rely heavily on the main prediction in Domingos [20].

C.2 Why we choose to avoid computing the main prediction

We now compare our definition of variance (Definition 2) to the one in Domingos [20, 21] (Definition 5). This comparison makes clear in detail why we deviate from prior work that relies on Domingos [20, 21].

No decomposition result. Following from above, it is worth noting that by not relying on the main prediction, we lose the applicability of the decomposition result that Domingos [20, 21] develop. However, we believe that this is fine for our purposes, as we are interested in the impact of empirical variance specifically on fair classification outcomes. We do not need to reason about bias or noise in our results to understand the arbitrariness with which we are concerned (Section 3.1). It is also worth noting that prior work on fair classification that leverages Domingos [20] also does not leverage the decomposition, either. Chen et al. [11] extends the decomposition to subgroups in the context of algorithmic fairness,[11] and then informally translates the takeaways of the Domingos [20] result to a notion of a "level of discrimination." Moreover, unlike our work, these prior studies do not actually measure variance directly in its experiments.

No need to compute a "central tendency". In Domingos [20, 21], variance is defined in terms of both the loss function f and the main prediction $\overline{y}$. This assumes that the main prediction is well-defined for the loss function, and that it is well-behaved. While there is a simple interpretation of the main prediction for squared loss (the mean) and for 0-1 loss (the mode/majority vote), it is significantly messier for cost-sensitive loss, which is a more general formulation that includes 0-1 loss. Domingos [20, 21] does not discuss this explicitly, so we derive the main prediction for cost-sensitive loss ourselves below. In summary:

---

[11]This just involves splitting the conditioning on an example instance of features $\boldsymbol{x}$ into conditioning on an example instance whose features are split into $(\boldsymbol{x}, \boldsymbol{g})$.

- The behavior of the main prediction for cost-sensitive loss reveals that the decomposition result provided in the extended technical report (Theorem 4, Domingos [21]) is in fact very carefully constructed. We believe that this construction is so specific that it is not practically useful (it is, in our opinion, hardly "unified" in a more general sense, as it is so carefully adapted to specific loss functions and their behavioral special cases).
- By decoupling from the need to compute a main prediction as a reference point, our variance definition is ultimately much simpler and more general, with respect to how it accommodates different loss functions.[12]

Brittleness of the main prediction. For high variance instances, the main prediction can flip-flop from $\hat{y} = 1$ to $\hat{y} = 0$ and back. While the strategy in Black et al. [6] is to abstain on the prediction in these cases, we believe that a better alternative is to understand that the main prediction is not very meaningful more generally for high-variance examples. That is, for these examples, the ability (and reliability) of breaking close ties to determine the main (simple majority) prediction is not the right approach. Instead, we should ideally be able to embed more confidence into our process than a simple-majority-vote determination.[13] Put different, in cases for which we can reliably estimate the main prediction, but the vote margin is slim, we believe that the main prediction is still uncertain, based on our understanding of variance, intuited in Figure 1. **The main prediction can be reliable, but it can still, in this view, be arbitrary** (Section 6). With a simple-majority voting scheme, there can be huge differences between predictions that are mostly in agreement, and those that are just over the majority reference point. Freeing ourselves of this reference point via our self-consistency metric, we can define thresholds of self-consistency as our criterion for abstention (where simple-majority voting is one instantiation of that criterion).[14]

C.2.1 The main prediction and cost-sensitive loss

We show here that, for cost-sensitive loss, the main prediction depends on the majority class being predicted, the asymmetry of the costs, and occasional tie-breaking, such that the main prediction can either be the majority vote or the minority vote. Domingos [21] provides an error decomposition in Theorem 4, but does not explain the effects on the main prediction. We do so below, and also call attention to 0-1 loss as a special case of cost-sensitive loss, for which the costs are symmetric (and equal to 1). We first summarize the takeaways of the analysis below:

- **Symmetric loss**: The main prediction is the **majority vote**.
- **Asymmetric loss**: Compute 1) the relative cost difference (i.e., $\frac{C_{01}-C_{10}}{C_{10}}$), 2) the majority class (and, as a result, the minority class) for the $\hat{y} \in \hat{\mathbb{Y}}$, and 3) the relative difference in the number of votes in the majority and minority classes (i.e., what we call the *vote margin*; below, $\frac{(i+2j+1)-i}{i}$)

  – If the **majority class** in $\hat{\mathbb{Y}}$ has the **lower cost** of misclassification, then the main prediction is the **majority vote**.
  – If the **majority class** in $\hat{\mathbb{Y}}$ has the **higher cost** of misclassification, then the main prediction **depends on the asymmetry of the costs and the vote margin**, i.e.,
    * If $\frac{C_{01}-C_{10}}{C_{10}} = \frac{(i+2j+1)-i}{i}$, we can choose the main prediction to be **either class** (but must make this choice consistently).
    * If $\frac{C_{01}-C_{10}}{C_{10}} > \frac{(i+2j+1)-i}{i}$, the **minority vote** is the main prediction.

---

[12]This reveals a subtle ambiguity in the definition of the loss f in Domingos [20, 21]. Neither paper explicitly defines the signature of f. For the main prediction (Definition 4) and variance (Definition 5), there is a lack of clarity in what constitutes a valid domain for f. Computing the main prediction $\overline{y}$ suggests $f : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}_{\geq 0}$, where $\overline{y} \in \mathbb{Y}$, but, since $\hat{\mathbb{Y}} \subseteq \mathbb{Y}$, it is possible that $\overline{y} \notin \mathbb{Y}$. However, the definition of variance suggests that $f : \mathbb{Y} \times \hat{\mathbb{Y}} \to \mathbb{R}_{\geq 0}$. Since $\hat{\mathbb{Y}} \subseteq \mathbb{Y}$, it is not guaranteed that $\hat{\mathbb{Y}} = \mathbb{Y}$. This may be fine in practice, especially for squared loss and 0-1 loss (the losses with which Domingos [20] explicitly contends), but it does arguably present a problem formally with respect to generalizing.

[13]This is also another aspect of the simplicity of not needing to define and compute a "central tendency" prediction. We do not need to encode a notion of a tie-breaking vote to determine a "central tendency." The main prediction can be unclear in cases for which there is no "main outcome" (e.g., Individual 2 in Figure 1), as the vote is split exactly down the middle. By avoiding the need to vote on a main reference point, we also avoid having to ever choose that reference point arbitrarily.

[14]This problem is worse for cost-sensitive loss, where the main prediction is not always the majority vote (see below).

$*$ If $\frac{C_{01}-C_{10}}{C_{10}} < \frac{(i+2j+1)-i}{i}$, the **majority vote** is the main prediction.

Proof. Let us consider cost-sensitive loss for binary classification, for which $f(0,0)=f(1,1)=0$ and we have potentially-asymmetric loss for misclassifications, i.e. $f(1,0)=C_{10}$ and $f(0,1)=C_{01}$, with $C_{01},C_{10}\in\mathbb{R}^+$. 0-1 loss is a special case for this type of loss, for which $C_{01}=C_{10}=1$.

Let us say that the total number of models trained is k, which we evaluate on an example instance $\boldsymbol{x}$. Let us set $|\hat{\mathbb{Y}}|=k=2i+2j+1$, with $i\geq 0$ and $j\geq 0$. We can think of i as the common number of votes that each class has, and $2j+1$ as the margin of votes between the two classes. Given this setup, this means that $k\geq 1$, i.e., we always have the predictions of at least 1 model to consider, and k is always odd. This means that there is always a strict majority classification.

Without loss of generality, on $\boldsymbol{x}$, of these k model predictions $\hat{y}\in\hat{\mathbb{Y}}$, there are i class-0 predictions and $i+2j+1$ class-1 predictions (i.e., we do our analysis with class 1 as the majority prediction). To compute the main prediction $\overline{y}$, each $\hat{y}\in\hat{\mathbb{Y}}$ will get compared to the values of possible predictions $y'\in\mathbb{Y}=\{0,1\}$. That is, there are two cases to consider:

- **Case $y'=0$:** $y'=0$ will get compared i times to the i $\hat{y}=0$s in $\hat{\mathbb{Y}}$, for which $f(0,0)=0$; $y'=0$ will similarly get compared $i+2j+1$ times to the 1s in in $\hat{\mathbb{Y}}$, for which (by Definition 4) the comparison is $f(1,0)=C_{10}$. By definition of expectation, the expected loss is
$$\frac{i\times 0+(i+2j+1)\times C_{10}}{2i+2j+1}=\frac{C_{10}(i+2j+1)}{2i+2j+1}. \tag{11}$$

- **Case $y'=1$:** Similarly, the label 1 will also get compared i times to the 0s in $\hat{\mathbb{Y}}$, for which the comparison is $f(0,1)=C_{01}$; $y'=1$ will also be compared $i+2j+1$ times to the 1s in $\hat{\mathbb{Y}}$, for which $f(1,1)=0$. The expected loss is
$$\frac{i\times C_{01}+(i+2j+1)\times 0}{2i+2j+1}=\frac{C_{01}i}{2i+2j+1}. \tag{12}$$

We need to compare these two cases for different possible values of C_{10} and C_{01} to understand which expected loss is minimal, which will determine the main prediction $\overline{y}$ that satisfies Equation (10). The three different possible relationships for values of C_{10} and C_{01} are $C_{10}=C_{01}$ (symmetric loss), and $C_{10}>C_{01}$ and $C_{10}<C_{01}$ (asymmetric loss). Since the results of the two cases above share the same denominator, we just need to compare their numerators, $C_{10}(i+2j+1)$ (11) and $C_{01}i$ (12).

Symmetric Loss (0-1 Loss). When $C_{10}=C_{01}=1$, the numerators in (11) and (12) yield expected losses $i+2j+1$ and i, respectively. We can rewrite the numerator for (12) as
$$i+\overbrace{2j+1}^{\geq 1,\text{ given }j\geq 0}\geq i+1,$$

which makes the comparison of numerators $i<i+1$, i.e., we are in the case (12) < (11). This means that the case of $y'=1$ (12) is the minimal one; the expected loss for class 1, the most frequent class, is the minimum, and thus the most frequent/ majority vote class is the main prediction. An analogous result holds if we instead set the most frequent class to be 0. More generally, this holds for all symmetric losses, for which $C_{10}=C_{01}$.

▶ For **symmetric losses**, the main prediction $\overline{y}$ is **majority vote** of the predictions in $\hat{\mathbb{Y}}$.

Asymmetric Loss. For asymmetric/ cost-sensitive loss, we need to examine two sub-cases: $C_{10}>C_{01}$ and $C_{10}<C_{01}$.

- **Case $C_{10}>C_{01}$:** $C_{01}i<C_{10}(i+\overbrace{2j+1}^{\geq 1})$, given that $j\geq 0$. Therefore, since $C_{01}i$ is minimal and associated with class 1 (the most frequent class in our setup), the majority vote is the main prediction. We can achieve an analogous result if we instead set 0 as the majority class.

  ▶ For **asymmetric losses**, the main prediction $\overline{y}$ is the **majority vote** of the predictions in $\hat{\mathbb{Y}}$, **if the majority class has a cheaper cost associated with misclassification** (i.e., if the majority class is 1 and $C_{10}<C_{01}$, or if the majority class is 0 and $C_{01}<C_{10}$).

- **Case** $C_{10} < C_{01}$: If $C_{10} < C_{10}$, it depends on how asymmetric the costs are and how large the vote margin (i.e., $2j+1$) between class votes is. There are 3 sub-cases:

  - **Case** $C_{01}i = C_{10}(i+2j+1)$**, i.e. cost equality**: We can look at the relative asymmetric cost difference of the minority class cost (above C_{01}, without loss of generality) and the majority class cost (above C_{10}, without loss of generality), (above $\frac{C_{01}-C_{10}}{C_{10}}$, without loss of generality). If that relative cost difference is equal to the relative difference of the votes between the majority and minority classes (i.e., $\frac{(i+2j+1)-i}{i}$), then the costs of predicting either 1 or 0 are equal. That is, we can rearrange terms as a ratio of costs to votes:

$$C_{01}i = C_{10}(i+\overbrace{2j+1}^{\geq 1}) \qquad \text{(The terms in this equality are} > 0)$$

$$\frac{C_{01}}{C_{10}} = \frac{i+2j+1}{i} \qquad \text{(Given the above, } C_{01}i > 0 \text{ so } i > 0)$$

$$= 1 + \frac{2j+1}{i}$$

$$\frac{C_{01}}{C_{10}} - 1 = \frac{2j+1}{i}$$

$$\frac{C_{01}-C_{10}}{C_{10}} = \frac{2j+1}{i} = \frac{(i+2j+1)-i}{i} \geq \frac{1}{i} \tag{13}$$

  ► For asymmetric loss **when the majority-class-associated cost is less than the minority-class associated cost and if the expected losses are equal**, then the **main prediction** $\overline{y}$ **is either** 1 **or** 0, (and we must make this choice consistently).

  - **Case** $C_{01}i > C_{10}(i+2j+1)$: We can look at the relative asymmetric cost difference of the minority class cost (above C_{01}, without loss of generality) and the majority class cost (above C_{10}, without loss of generality), (above $\frac{C_{01}-C_{10}}{C_{10}}$, without loss of generality). If that relative cost difference is greater than the relative difference of the votes between the majority and minority classes (i.e., $\frac{(i+2j+1)-i}{i}$), then the *minority vote* yields the minimum cost and is the main prediction $\overline{y}$ (above $\overline{y}=0$, without loss of generality; an analogous result holds if we had set the majority vote to be 0 and the minority vote to be 1). Following (13) above, this is the same as

$$\frac{C_{01}-C_{10}}{C_{10}} > \frac{(i+2j+1)-i}{i}$$

  ► For asymmetric loss **when the majority-class-associated cost is less than the minority-class associated cost**, it is possible for the **minority class** to have a greater associated loss. In this case, the *minority vote* **is the main prediction** $\overline{y}$.

  - **Case** $C_{01}i < C_{10}(i+2j+1)$: We can look at the relative asymmetric cost difference of the minority class cost (above C_{01}, without loss of generality) and the majority class cost (above C_{10}, without loss of generality), (above $\frac{C_{01}-C_{10}}{C_{10}}$, without loss of generality). If that relative cost difference s less than the relative difference of the votes between the majority and minority classes (i.e., $\frac{(i+2j+1)-i}{i}$), then the majority vote yields to minimum cost and is the main prediction $\overline{y}$ (above $\overline{y}=1$, without loss of generality; an analogous result holds if we had set the majority vote to be 0 and the minority vote to be 1). Following (13) above, this is the same as

$$\frac{C_{01}-C_{10}}{C_{10}} < \frac{(i+2j+1)-i}{i}$$

  ► For asymmetric loss **when the majority-class-associated cost is less than the minority-class associated cost**, it is possible for the **majority class** to have a greater associated loss. In this case, the *majority vote* **is the main prediction** $\overline{y}$.

$\square$

C.3 Putting our work in conversation with research on model multiplicity

A line of related work to ours concerns *model multiplicity* and fairness [7, 47, 58]. This work builds off of an observation made by Breiman [10] regarding how there are multiple possible models of the

same problem that exhibit similar degrees of accuracy. This set of multiple possible models of similar accuracy is referred to as the Rashomon set [10].

Work on model multiplicity has recently become fashionable in algorithmic fairness. In an effort to develop more nuanced model selection metrics beyond looking at just fairness and accuracy for different demographic groups, work at the intersection of model multiplicity and fairness tends to examine other properties of models in the Rashomon set in order to surface additional metrics for determining which model to use in practice.

At first glance, this work may seem similar to what we investigate here, but we observe four key differences:[15]

1. Model multiplicity places conditions on accuracy and fairness in order to determine the Rashomon set. We place no such conditions on the models that a learning process (Definition 1) produces; we simulate the distribution over possible models μ without making any claims about the associated properties of those models.

2. Model multiplicity makes observations about the Rashomon set with the aim of still ultimately putting forward criteria for helping to select *a single model*. While the metrics used to inform these criteria include variance, most often work on model multiplicity still aims to choose one model to use in practice.

3. Much of the work on model multiplicity emphasizes theoretical contributions, whereas our emphasis is on more experimental contributions. In conjunction with the first point, of ultimately trying to arrive at a single model, this work is also trying to make claims with respect to the Bayes-optimal model. Given our empirical focus — of what we can actually produce in practice — claims about optimality are not our concern.

4. We focus specifically on variance reduction as a way to mitigate arbitrariness. We rely on other work, coincidentally contributions also made by Breiman, to study arbitrariness [8], and emphasize the importance of using ensemble models to produce predictions or abstention from prediction. We do not study the development of model selection criteria to pick a single model to use in practice; we use self-consistency to give a sense of predictive confidence about when to predict or not. We always select an ensemble model — regardless of whether that model is produced by simple or super ensembling (Section 4) — and then use a user-specified level of self-consistency κ to determine when that model actually produces predictions.

These differences ultimately lead to very different methods for making observations about fairness. Importantly, we can study the arbitrariness of the underlying laerning process with a bit more nuance. For example, it could be the case that a particular task is just impossible to get right for some large subset of the test data (and this would be reflected in the Rashomon set of models), but for some portion of it there is a high amount of self-consistency for which we may still want to produce predictions.

Further, based on our experimental approach, we highlight completely different normative problems than those highlighted in work on model multiplicity (notably, see Black et al. [7]). So, in short, while model multiplicity deals with related themes as our work — issues of model selection, problem formulation, variance, etc. — the goals of that work are ultimately different, but potentially complementary, from those in our paper.

For example, a potentially interesting direction for future work would be to measure how metrics from work on model multiplicity behave in practice in light of the ensembling methods we present here. We could run experiments using Algorithm 1 and investigate model multiplicity metrics for the underlying ensembled models. However, we ultimately do not see a huge advantage to doing this. Our empirical results indicate that variance is generally high, and has led to reliability issues regarding conclusions about fairness and accuracy. In fairness settings and available benchmarks, we find that the most important point is that variance has muddled conclusions. Under these circumstances, ensembling with abstention based on self-consistency seems a reasonable solution, in contrast to finding a single best model in the Rashomon set that attains other desired criteria.

---

[15]We defer discussion of Black et al. [6] to C.4.

C.4 Concurrent work

There are several related papers that either preceded or came after this work's public posting. Some of this work is clearly concurrent, given the time frame. Other works that came after ours are not necessarily concurrent, but are either independent and unaware of our paper, or build on our work.

Setting the stage in 2021. The present work was scoped in 2021, in direct response to the initial study by Forde et al. [28] and critical review by Cooper and Abrams [12]. Forde et al. [28] was one of the first (if not the first) paper to note that variance is overlooked in problem formulations that consider fairness. However, it was limited in scope and also dealt with deep learning settings, which have multiple sources of non-determinism that can be difficult to tease apart with respect to their effects on variance.

Cooper and Abrams [12] notes important, overlooked normative assumptions in the fairness-accuracy trade-off problem formulation, and suggests that this formulations is tautological. Our work is a natural direction for future research, in this respect – to see how, in practice, the fairness-accuracy trade-off behaves after we account for variance. Indeed, we find that there is often no such trade-off, but for different reasons than those suggested by Cooper and Abrams [12]. We expected there to be residual label bias that contributes to noise-induced error, but ultimately did not really observe this in practice. In these respects, our work both strengthens and complements these prior works. We support their claims, and go significantly beyond the work they did in order to provide such support. Further, our results suggest additional conclusions about experimental reliability in algorithmic fairness.

Variance and abstention-based ensembling. Black et al. [6] is concurrent work that slightly preceded our public posting. This work is similarly is interested in variance reduction, ensembling, and abstention in fairness settings, but fundamentally studies these topics in a different manner. We address four differences:

1. Black et al. [6] does not take the wide-ranging experimental approach that we take. While we both study variance and fairness, our work also considers *the practice of fair classification research* as an object of study. It is for these reasons that we do so many experiments on benchmark datasets, and clean and release another dataset for others to use.

2. They rely on the definition of variance from Domingos [20] in their work, likely building on the choice made by Chen et al. [11] to use this defintion. Much of this Appendix is devoted to discussing Domingos [20, 21] and his definition of variance. The overarching takeaway from our discussion is that 1) there are technical problems with this definition (which have been noted by others that investigated the bias-variance-noise trade-off for 0-1 loss in the early 2000s), 2) the definition does not naturally extend to cost-sensitive loss, 3) the main prediction can be unstable in practice and thus should not be the criterion for investigating arbitrariness (indeed, relying on the main prediction just pushes arbitrariness into that definition). While Black et al. [6] observes that variance is an important consideration for fairness, they ultimately focus on reliable estimation of the main prediction as the criterion for abstention in their ensembling method. While this kind of reliability is important, it does not deal with the general problem of arbitrary predictions (i.e., it is possible to have a reliable main prediction that is still effectively arbitrary). As a result, the nature of when and how to abstain is very different from ours. We instead base our criterion on a notion of confidence in the prediction, and we allow for flexibility around when to abstain when predictions are too arbitrary.

3. As a result of the above two differences, the claims and conclusions in both of our works are different. While there are similar terms used in both works (e.g., variance, abstention), which may make the works seem overlapping with a cursory read, our definitions, methods, claims, and conclusions are non-overlapping. For example, as stated in 1., while Black et al. [6]'s use of successful ensembles is intended to address individual-level arbitrariness, by relying on traditional bagging (simple-majority vote ensembling) and the definition of variance from Domingos [20] that encodes a main prediction, arbitrariness gets pushed into the aggregation rule. If they can estimate the mode prediction reliably, they do not abstain; the mode, however, may still be effectively arbitrary. Our measure of arbitrariness is more direct and more configurable. We can avoid such degenerate situations, as in the example we give for making reliable but arbitrary predictions in Black et al. [6].

4. We also describe a method for recursively ensembling in order to achieve different trade-offs between abstention and prediction. This type of strategy is absent from Black et al. [6].

Deep learning. Qian et al. [51] is work that came after Forde et al. [28]. They, too, do a wide-ranging empirical study of variance and fairness, but focus on deep learning settings. As a result, they are not examining the fair classification experimental setup that is most common in the field. They therefore make different claims about reliability, which have a similar flavor as those that we make here. However, because of our setup, we are able to probe these claims much deeper (due in part to model/ problem size and being able to limit non-determinism solely to sampling the training data). We mention this work because of its close relationship to Forde et al. [28], which in part inspired this study.

Ko et al. [41] is another deep learning fairness paper. It was posted publicly months after our study, and examines non-overlapping settings and tasks. While the results are similar — we find fairness after ensembling — it is again fundamentally different (along the lines of Qian et al. [51] and Forde et al. [28]) because it does not study common non-deep-learning setups. They also do not study arbitrariness, which is one of the main purposes of our paper.

Variance in fair classification. Khan et al. [39] is concurrent work that studies the same problem that we study, but also takes a different approach. For one, they bake in a notion of 0-1 loss into their definitions. In this respect, our definition of self-consistency generalizes the definitions in their paper. While they run more types of models than we do (we initially ran more, but ultimately stopped because the results were largely similar with more common model types), they do not cover as many datasets as we do. They also do not study arbitrariness or abstention-based ensembling to deal with it, and they do not release a dataset. Further, based on the fact that they study fewer empirical tasks than we do, and that they do not examine abstention-based ensembling, they do not surface or make claims about the experimental reliability issues that we observe. They do not make claims about the fundamental problem that we observe: **That variance is the culprit for much observed algorithmic unfairness in classification; in practice, we do not seem to learn very confident decisions for large portions of the datasets we examine, and this is a key problem that has been masked by current common experimental practices in the field**. We make notes about this in our Ethics Statement.

Other work. Any other work on variance and fairness **comes after** the present study. We have made a significant attempt to keep our related work section up-to-date in response to this new work. We have used a detailed and robust mixed of Google alerts and scraping arXiv to find new related work. We used this same procedure to make sure we found (ideally) all related work on fairness and variance when we conducted this project. There are some studies, which directly build on ours, which we choose not to cite.

D   Additional Details on Our Algorithmic Framework

A natural question is to see if we can improve self-consistency, with the hope that doing so would reduce arbitrariness in the learning process, improve accuracy, and, for the cases in which there is different self-consistency across subgroups, also perhaps improve fairness. To do so, we consider ways of reducing variance, as, based on our definitions (Definition 2 and 3), doing so should improve self-consistency.

We consider the classic *b*ootstrap *agg*regation — or, *bagging* — algorithm [8] as a starting point. It has been well-known since Breiman [8] that *bagging* can improve the performance of unstable predictors. That is, for models produced by a learning process that is sensitive to the underlying training data, it is (theoretically-grounded) good practice to train an ensemble of models using bootstrapping (Appendix A.4; Efron [22], Efron and Tibshirani [24]). When classifying an example instance, we then leverage the whole ensemble by aggregating the predictions produced by its members. This aggregation process identifies the most common prediction in the ensemble, and returns that label as the classification. Put differently, we have combined the information of a lot of unstable classifiers, and averaged over their behavior in order to generate more stable classifications.

Given the the relationship between variance (Definition 2) and self-consistency (Definition 3), reducing variance will improve self-consistency. However, rather than relying on a simple-majority-vote to decide the aggregated prediction, we also will instill a notion of confidence in our predictions by requiring a minimum level of self-consistency, which is described in Algorithm 1.

D.1   Self-consistent ensembling with abstention

We present a framework that alters the semantics of classification outputs to 0, 1, and `Abstain`, and employ ensembling to determine the $\hat{SC}$-level that guides the output process. We modify bagging from using a simple-majority-vote because this type of aggregation rule still allows for arbitrariness. If, for

example, we happen to train $B = 101$ classifiers, it is possible that 50 of them yield one classification and the other 51 yield the other classification for a particular example. Bagging would select the classification that goes along with the 51 underlying models; however, if we happened to train $B = 103$ models, it is perhaps the case that the majority vote would flip. In short, the bagging aggregation rule bakes in the idea that simple-majority voting is a sufficient strategy for making decisions. And while this may generally be true for variance reduction in high-variance classifiers, it does not address the problem of arbitrariness that we study. It just encodes arbitrariness in the aggregation rule — it picks classifications, in some cases, that are no better than a coin flip.

Instead, Algorithm 1 is more flexible. It suggests many possible ways to produce bagged classifiers that do not have to rely on simple-majority voting, by allowing for abstentions. For example, we can change the aggregation rule in regular bagging to use a self-consistency level κ rather than majority vote. Instead of relying on votes, we can bag the underlying prediction probabilities and then apply κ a filter. We could take the top-n most consistent predictions and let a super-ensemble of underlying bagged classifiers decide whether to abstain or predict.

In the experiments in the paper, we provide two examples: Changing the underlying bagging vote aggregation rule (simple ensembling), and applying a round of regular bagging to do variance reduction and then bagging the bagged outputs (super ensembling) to apply a self-consistency threshold. Our ensemble model will not produce predictions for examples for which the lack of self-consistency is too high. We describe our procedure more formally in Algorithm 1.

Simple proof that abstention improves self-consistency (by construction). We briefly show the simple proof that any method that meets the semantics of Algorithm 1 will be more self-consistent than its counterpart that cannot `Abstain`.

We define abstentions to be in agreement with both 0 and 1 predictions. This makes sense intuitively: Algorithm 1 abstains to avoid making predictions that lack self-consistency, so abstaining should not increase disagreement between predictions.

It follows that we can continue to use Definition 3 and associated empirical approximations $\hat{\text{SC}}$ (3), but with one small adjustment. Instead of the total number of predictions $B = B_0 + B_1$, with B_0 and B_1 corresponding to 0 and 1 predictions, respectively, we now allow for $B \geq B_0 + B_1$, in order to account for possibly some non-zero number of abstentions.

In more detail, let us denote $\hat{\mathbb{Y}}$ to be the multiset of predictions for models $h_{\boldsymbol{D}_1}, h_{\boldsymbol{D}_2}, \ldots, h_{\boldsymbol{D}_B}$ on $(\boldsymbol{x}, \boldsymbol{g})$, with $|\hat{\mathbb{Y}}| = B = B_0 + B_1 + B_{\texttt{Abstain}}$. This is where we depart from our typical definition of self-consistency, for which $B = B_0 + B_1$ (Section 3, Appendix B.3). We continue to let B_0 and B_1 represent the counts of 0 and 1 predictions, respectively, and now include $B_{\texttt{Abstain}}$ to denote the (possibly nonzero) number of abstentions. This leads to the following adjustment of (3):

$$\hat{\text{SC}}\big(\mathcal{A}, \{\hat{\boldsymbol{D}}_b\}_{b=1}^B, (\boldsymbol{x}, \boldsymbol{g})\big) = 1 - \frac{2(B_0 B_1 + B_0 B_{\texttt{Abstain}} + B_1 B_{\texttt{Abstain}})}{B(B-1)}. \tag{14}$$

Equation (14) follows from a similar analysis of comparing 0s, 1s, and abstentions for Definition 3, which lead us to derive (3) in Appendix B.3. However, since the costs of 0-to-`Abstain` comparisons and 1-to-`Abstain` comparisons are both 0, the $B_0 B_{\texttt{Abstain}}$ and $B_1 B_{\texttt{Abstain}}$ terms in (14) reduce to 0. As a result, we yield our original definition for self-consistency (3), with the possibility that $B = B_0 + B_1 + B_{\texttt{Abstain}} > B_0 + B_1$, if there is a nonzero number of abstentions $B_{\texttt{Abstain}}$.

Since $B > 1$ and $B_0, B_1, B_{\texttt{Abstain}} \geq 0$, it is always the case that option to `Abstain` is at least as self-consistent as not having the option to do so. This follows from the fact that $B_0 + B_1 + B_{\texttt{Abstain}} = B \geq B_0 + B_1$, which would make the denominator in (14) greater than or equal to the corresponding method that cannot `Abstain`; when subtracted from 1, this would produce a $\hat{\text{SC}}$ that is no smaller than the value for the corresponding method without that cannot `Abstain`.

Now, it follows that, given the choice between `Abstain` and predicting a label that is in disagreement with an existing prediction label, choosing to `Abstain` will always lead to higher self-consistency. This is because the cost to `Abstain` is less than disagreeing, so it will always be the minimal choice that maximizes $\hat{\text{SC}}$.

Error and the abstention set. It is very straightforward to see that the *abstention set* will generally exhibit higher than the *prediction set*. When we ensemble and measure $\hat{\text{SC}}$, the exmaples that exhibit

low $\hat{\text{SC}}$ contain higher variance-induced error. Let us call the size of the abstention set U (which incurs error u), the size of the prediction set V (which incurs error v), and the size of the test set T (which incurs error t). We can relate the total number of misclassified examples as $T*t = U*u + V*v$, with $T = U+V$. If we assume the bias and noise are equally distributed across the test and abstention sets (this is a reasonable assumption, on average, in our setup), then splitting off the high variance instances from the low variance (high $\hat{\text{SC}}$ instances) requires that $u > v$. The error on the abstention set necessarily has to be larger than the error on the prediction set, in order to retain the above relationship.

E   Additional Experimental Results and Details for Reproducibility

The code for the examples in Sections 1, 3 and 5 can be found in https://github.com/pasta41/variance. This repository also contains necessary and sufficient information concerning reproducibility. At the time of writing, we use `Conda` to produce environments with associated package-versioning information, so that our results can be exactly replicated and independently verified. We also use the `Scikit-Learn` [49] toolkit for modeling and optimization. More details on our choice of models and hyperparameter optimization can be found in our code repository, cited above. In brief, we consulted prior related work (e.g., Chen et al. [11]) and performed our own validation for reasonable hyperparameters per model type. We keep these settings fixed to reduce impact on our results, in order to observe in isolation how different training data subsets impact our results.

During these early runs, we collected information on train accuracy, not just test accuracy; while models ultimately have similar test accuracy in most cases for the same task, they can vary significantly in terms of train accuracy (e.g., for logistic regression, `COMPAS` is in the low .70s; for random forests, it is in the mid .90s). We do not include these results for the sake of space.

This section is organized as follows. We first present information on our datasets, models and code, including our `HDMA` toolkit (Appendix E.1). We then provide details on our setup for running experiments on our cluster (Appendix E.2). Appendix E.3 contains more detailed information concerning the experiments performed to produce Figures 1 and 2 in the main paper. In Appendix E.4, we provide more details on the results presented in Section 5, as well as additional experiments. Lastly, in ppendix E.5, we discuss implications of these results for common fairness Abenchmarks like `South German Credit`. We conclude that in many cases, without adequate attention to error estimation, it is likely that training and post-processing a single model for fairness on these models likely is a brittle approach to achieve generalizable fairness (and accuracy) performance. Based on our experiments, it seems like high variance can be a significant confounding factor when using a small set of models to draw conclusions about performance — whether fairness or accuracy. There is an urgent need for future work concerning reproducibility. More specifically, our results indicate that it would be useful to revisit key algorithmic strategies in fair classification to see how they perform in context with more reliable expected error estimation and variance reduction.

Note on CDF figures. We show our results in terms of the $\hat{\text{SC}}$ of the underlying bagged models because doing so conveys how Algorithm 1 makes decisions to predict or abstain.[16] For both types of ensembling, Algorithm 1 predicts for all examples captured by the area to the right of the κ reference line, and abstains for all examples on the left.

It is also worth noting (though hopefully obvious) that our CDF plots of $\hat{\text{SC}}$ are not continuous, yet we choose to plot them as interpolated curves. This are discrete because we train a concrete number of models (individual models or bags) — typically 101 of them — that we treat as our approximation for B when computing $\hat{\text{SC}}$. This means that there are a finite number of κ-values for $\hat{\text{SC}}$, for which we plot a corresponding concrete number of heights y corresponding to the cumulative proportion of the test set. In this respect, it would perhaps be more precise to plot our curves using a step function, exemplified below (see Appendix B.3 for the values in $\hat{\mathbb{K}}$):

We opted not to do this for two reasons. First, plotting steps for some of our figures, in our opinion, can make the figures more difficult to understand. Second, in experiments for which we increase the number of models used to estimate $\hat{\text{SC}}$ (e.g., Appendix E.5), we found that the curves for 101 models were a reasonable approximation of the overall CDF. We therefore concluded that plotting

---

[16] The $\hat{\text{SC}}$ CDF of Algorithm 1, computed via a *third* round of bootstrapping, has nearly all mass at $\hat{\text{SC}} = 1$; it is difficult to visualize.
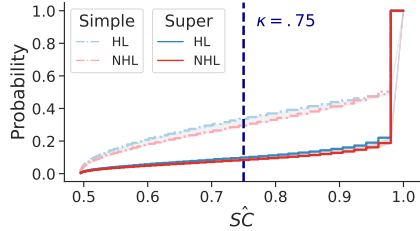
Figure 6: Plotting $\hat{SC}$ with an emphasis on discrete levels κ.

the figures without steps was worth the clarity of presentation, with a sacrifice in correctness for the overall takeaways that we intend with these figures.

A remark on cost. It can be considerably more computationally intensive to train an ensemble of models to compute $\hat{SC}$ than to train a handful of models and perform cross-validation, as is the standard practice in fair classification. However, as our empirical analysis demonstrates, this cost comes with a huge benefit: It enables us to improve self-consistency and to root out the arbitrariness of producing predictions that are effectively close-to-random, which is especially important in high-stakes fairness settings [13]. Moreover, for common fair classification datasets, the increased cost on modern hardware is relatively small; (super-) ensembling with confidence takes under an hour to execute (Appendix E.4).

E.1 Hypothesis classes, datasets, and code

Models. According to a comprehensive recent survey study [26], as well as related work like Chen et al. [11], we conclude that some of the most common models used in fair classification are logistic regression, decision tree classifiers, random forest classifiers, SVMs, and MLPs. We opted to include comprehensive results for the first three, since they capture different complexities, and therefore encode different degrees of statistical bias, that we expected to have an impact on the underlying sources of error. We provide some results for SVMs and MLPs, which we include in this Appendix. Since we choose not to use stochastic optimizers to reduce the sources of randomness, for our results, training MLPs is slower than it could be. We consistently use a decision threshold of 0.5 (i.e., 0-1 loss) for our experiments, though our results can easily be extended to other thresholds, as discussed in Section 3. Depending on the dataset, we reserve between 20% and 30% of the available data for the test set. This is consistent with standard fair classification training settings, which we validated during our initial experiments to explore the space (for which we also did preliminary hyperparameter optimization, before fixing the hyperparameters for our presented results).[17]

Datasets. Also according to Fabris et al. [26], the most common tasks in fair classification are `Old Adult` [42], `COMPAS` [44], and `South German Credit` [33].[18] These three datasets arguably serve as a *de facto* benchmark in the community, so we felt the need to include them in the present work. In recognition of the fact that these three datasets, however standard, have problems, we also run experiments on 3 tasks in the `New Adult` dataset, introduced by Ding et al. [19] to replace `Old Adult`. We subset to the `CA` (California) subset of the dataset, and run on `Income`, `Employment`, and `Public Coverage`, and consider `sex` and `race` as protected attributes, which we binarize into {Male, Female} and {White, Non-white}. These are all large-scale tasks, at least in the domain of algorithmic fairness — on the order of hundreds of thousands of example instances. However, the 3 tasks do share example instances and some features. In summary, concerning common tasks in fair classification:

- `COMPAS` [44]. We run on the commonly-used version of this dataset from Friedler et al. [30], which has 6167 example instances with 404 features. The target is to predict recidivism within 2 years (1 corresponding to Yes, and 0 to No). The protected attribute is `race`, binarized into "Non-white" (0) and "White" (1) subgroups.

- `Old Adult` [42]. We run on the commonly-used version of this dataset from Friedler et al. [30], which has 30,162 examples with 97 features. This version of the dataset removes instances with missing values from the original dataset, and changes the encoding of some

---

[17]Please refer to `https://github.com/pasta41/variance` for more details.

[18]Technically, Grömping [33] is an updated and corrected version of the dataset from 2019.

of the features (Kohavi [42] has 48842 example instnaces with 88 features). The target is to predict $< \$50,000$ income (0) $>= \$50,000$ income (1). The protected attribute is `sex`, binarized into "Female" (0) and "Male" (1) subgroups.

- `South German Credit` [33]. We download the dataset from UCI[19] and process the data ourselves. We use the provided `codetable.txt` to "translate" the features from German to English. We say "translate" because the authors took some liberties, e.g., the column converted to "credit_history" is labeled "moral" in the German, which is not a translation. There are four categories in the protected attribute "personal_status_sex" column, one of which (2) is used for both "Male (single)" and "Female (non-single)." We therefore remove rows with this value, and binarize the remaining three categories into "Female" (0) and "Male" (1). What results is a dataset with 690 example instances (of the original 1000) with 19 features. The target is "good" credit (1) and "bad" credit (0).

- `Taiwan Credit` [59]. This task is to predict default on credit card payments (1) or not (0). There are 30,000 example instances and 24 features. The protected attribute is binary `sex`. We download this dataset from UCI.[20].

- `New Adult` [19]. This dataset contains millions of example instances from US Census data, which can be used for several different targets/tasks. We select three of them (listed below). These tasks share some features, and therefore are not completely independent. Further, given the size of the whole dataset, we subset to `CA` (California), the most populous state in the US. There are two protected attribute columns that we use: sex, which is binarized "Female" (0) and "Male" (1) subgroups, and race, which we binarize into "Non-white" (0) and "White" (1). In future work, we would like to explore extending our results beyond binary subgroups.

    - `Income`. This task is designed to be analogous to `Old Adult` [42]. As a result, the target is to predict $< \$50,000$ income (0) $>= \$50,000$ income (1). In the `CA` subset, there are 195,665 example instances with 8 features.
    - `Employment`. This task is to predict whether an individual is employed (1) or not (0). In the `CA` subset, there are 378,817 example instances with 14 features.
    - `Public Coverage`. This task is to predict whether an individual is on public health insurance (1) or not (0). In the `CA` subset, there are 138,554 example instances with 17 features.

E.1.1   The standalone `HMDA` tookit

In addition to the above standard tasks, we include experiments that use the `NY` and `TX` 2017 subsets of the the Home Mortgage Data Disclosure Act (`HMDA`) 2007-2017 dataset [27]. These two datasets have 244,107 and 576,978 examples, respectively, with 18 features. The `HMDA` datasets together contain over 140 million examples of US home mortgage loans from 2007-2017 (newer data exists, but in a different format). We developed a toolkit, described below, to make this dataset easy to use for classification experiments. Similar to `New Adult`, we enable subsetting by US state. For the experiments in this paper, we run on the `NY` (New York) and `TX` (Texas) 2017 subset, in order to add some geographic diversity to complement our `New Adult` experiments. We additionally chose `NY` and `TX` because they are two of the most populous states in the US, alongside `CA`.[21]

The target variable, `action_taken`, concerning loan origination has 8 values, 2 of which we cannot meaningful conclude approval or denial decisions. They are: Action Taken: 1 – Loan originated, 2 – Application approved but not accepted, 3 – Application denied by financial institution, 4 – Application withdrawn by applicant, 5 – File closed for incompleteness, 6 – Loan purchased by the institution, 7 – Preapproval request denied by financial institution, and 8 – Preapproval request approved but not accepted (optional reporting). We filter out 4 and 6, and binarize into $grant = \{1,2,8\} = 1$ and $reject = \{3,5,7\} = 0$. There are three protected attributes that we consider: `sex`, `race`, and `ethnicity`:

- `sex` has 5 possible values, 2 of which correspond to categories/non-missing values: Male $- 1$ and Female $- 2$. We binarize `sex` into $F = 0$ and $M = 1$.
- `race` has 8 possible values, 5 of which correspond to categories/ non-missing information: 1 – American Indian or Alaska Native, 2 – Asian, 3 – Black or African American, 4 – Native

---

[19]See `https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29`

[20]See `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

[21]Per the 2020 Census, the top-4-most-populous states are `CA`, `TX`, `FL`, and `NY` [46].

Hawaiian or Other Pacific Islander, and 5 – White. There are 5 fields for applicant race, which model an applicant belonging to more than one racial group. For our experiments, we only look at the first field. When we binarize `race`, NW $= 0$ and W $= 1$.

- `ethnicity` has 5 possible values, 2 of which correspond to categories/ non-missing information: 1 – Hispanic or Latino and 2 – Not Hispanic or Latino. We binarize ethnicity to be HL $= 0$ and NHL $= 1$.

After subsetting to only include examples that have values that do not correspond to missing information, `HMDA` has 18 features. The `NY` dataset has 244,107 examples; the `TX` dataset has 576,978 examples, making it the largest dataset in our experiments. As with our experiments using `New Adult`, we would like to extend our results beyond binary subgroups and binary classification in future work.

Releasing a standalone toolkit.   These datasets are less-commonly used in current algorithmic fairness literature [26]. We believe this is likely due to the fact that the over-100-million data examples are only available in bulk files, which are on the order of 10s of gigabytes and therefore not easily downloadable or explorable on most personal computers. Following the example of Ding et al. [19], one of our contributions is to pre-process all of these datasets — all locations and years — and release them with a software toolkit. The software engineering effort to produce this toolkit was substantial. Our hope is that wider access to this dataset will further reduce the community's dependency on small (and dated) datasets. Please refer to `https://github.com/pasta41/hmda` for the latest information on this standalone software package. Our release aligns with the terms of service for this dataset.

E.2   Cluster environment details

While most of the experiments run in this paper can be easily reproduced on a modern laptop, for efficiency, we ran all of our experiments (except the one to produce Figure 1) in a cluster environment. This enabled us to easily execute train/test splits n in parallel on different CPUs, serialize our results, and then reconstitute and combine them to produce plots locally. Our cluster environment runs Ubuntu 20.04 and uses Slurm v20.11.8 to manage jobs. We ran all experiments using `Anaconda3`, which is why we used `Conda` to reproduce environments for easy replicability.

The experiments using `New Adult` and `HMDA` rely on datasets that are (in some cases) orders of magnitude larger than the traditional algorithmic fairness tasks. This is one of the reasons why we recommend running on a cluster, and therefore do not include Jupyter notebooks in our repository for these tasks. We also limit our modeling choices to logistic regression, decision tree classifiers, and random forest classifiers for these results due to the expense of training on the order of thousands of models for each experiment.

E.3   Details on motivating examples in the main paper

This appendix provides extended results for the experiments associated in Sections 1 and 3, which give an intuition for individual- and subgroup-level consistency. The experimental results in the main paper are for logistic regression. We expand the set of models we examine, and associated discussion of how to interpret comparisons between these results.

Reproducing Figure 1.   The experiment to produce this figure in Section 1 (also shown in Appendix B.3) trains $B = 10$ logistic regression models on the `COMPAS` dataset (Appendix E.1) using 0-1 loss. We use the bootstrap method to produce each model, which we evaluate on the same test set. We then search for a maximally consistent and minimally consistent individual in the test set, i.e., an individual with 10 predictions that agree and an individual with 5 predictions in each class, which we plot in the bar graph. Please refer to the README in `https://github.com/pasta41/variance` regarding which `Jupyter` notebook to run to produce the underlying results and figure. The experiments to reproduce this figure can be easily replicated on a laptop.

Reproducing Figure 2. These figures were produced by executing $S = 10$ runs of $B = 101$ bootstrap training replicates to train random forest classifiers for `Old Adult` and `COMPAS`. We reproduce these figures below, so that they can be examined and treated in relation to our additional results for decision tree classifiers and logistic regression. For each s run, we take train/test split, bootstrap the train split $B = 101$ times, and evaluate the resulting model classification decisions on the test set. $\hat{SC}$ can be estimated from the results across those 101 models. We Run this process $S = 10$ times to produce confidence intervals, shown in the figures below. The intervals are not always clearly visible; there is

not a lot of variance at the level of comparing whole runs to each other. Please refer to the README in `https://github.com/pasta41/variance` regarding which `Jupyter` notebook to run to produce the underlying results and figure. There are also scripted version of these experiments, which enable them to be run in parallel in a cluster environment.

Self-consistency of incorrectly-classified instances. Last, we include figures that underscore how self-consistency is independent from correctness that is measured in terms of observed label alignment. That is, it is possible for an instance $(\boldsymbol{x}, \boldsymbol{g})$ to be self-consistent and classified incorrectly, with respect to its observed label o. We show this using stacked bar plots. For the above experiments, we find the test examples that have the majority of their classifications incorrect ($\hat{y} \neq o$, for $B = 101$, we find the instances with ≥ 51 incorrect classifications) and the majority of their classification correct (similarly), and we examine how self-consistent they are. We bucket self-consistency into different levels, and then plot the relative proportion of majority-incorrectly and majority-correctly classified examples according to subgroup. Subgroups in `COMPAS` exhibit a similar trend, while subgroups in `Adult Old` exhibit differences, with the heights of the bars corresponding to the trends we plot in our CDF plots. As we note briefly in Section 3, it may be interesting to examine patterns in examples about which learning processes are confident (i.e., highly self-consistent) but wrong in terms of label alignment. If such issues correlate with subgroup, it may be worth testing the counterfactual that such labels are indicative of label bias. We leave such thoughts to future work.
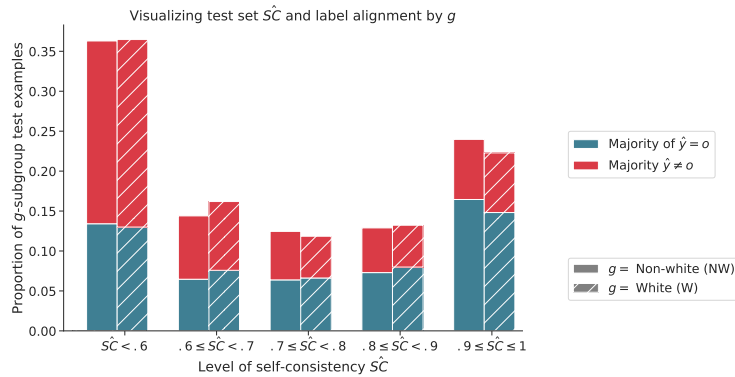
(a) `COMPAS`
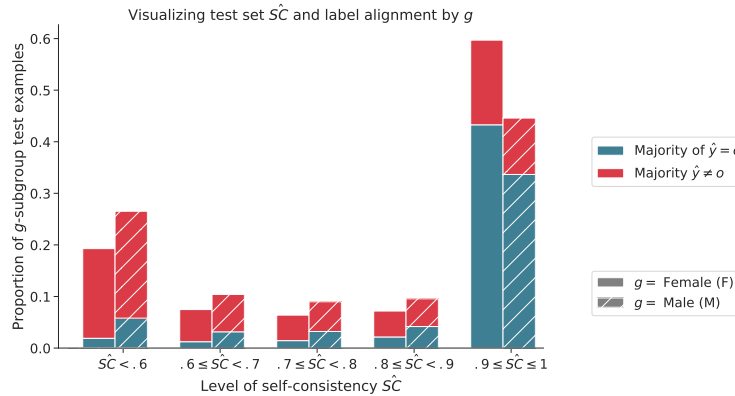
(b) `Adult Old`

Figure 7: $\hat{SC}$ broken down by $\boldsymbol{g}$ and label alignment with the observed label o. For each train/test split, and for each $\hat{SC}$ range (x-axis), we find the examples that are incorrectly classified the majority of time (≥ 5 splits, we find that $\hat{y} \neq o$), and the examples that are correctly classified the majority of the time (> 5, we find that $\hat{y} = o$). We compute the average the proportion over (over splits) in each $\hat{SC}$ range (y-axis). We plot these proportions with respect to subgroup $\boldsymbol{g}$ (where the sums of the heights of bars for by each $\boldsymbol{g}$ is equal to 1).

E.4 Validating our algorithm in practice

E.4.1 `COMPAS`

$\hat{SC}$ CDFs for `COMPAS` ($g = \texttt{race}$) and associated error metrics on the prediction set. **Baseline** metrics computed with $B = 101$ models. For **simple**, $B = 101$ models; for **super**, $B = 101$ ensemble models, each composed of 51 underlying models. We repeat for 10 test/train splits. We also report abstention rate $\hat{AR}$.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	14.5±0.3%	18.7±0.5%	15.6±0.1%
$\hat{PR}_{NW}$	45.3±1.2%	43.8±1.1%	44.2±0.7%
$\hat{PR}_{W}$	30.8±1.5%	25.1±1.6%	28.6±0.6%
$\Delta\hat{Err}$	0.2±0.2%	1.1±1.5%	0.9±1.1%
$\hat{Err}_{NW}$	33.0±1.3%	27.9±0.9%	31.0±1.0%
$\hat{Err}_{W}$	33.2±1.1%	29.0±2.4%	31.9±2.1%
$\Delta\hat{FPR}$	2.1±0.0%	3.0±0.0%	1.8±0.2%
$\hat{FPR}_{NW}$	14.7±1.3%	11.4±1.0%	12.9±0.8%
$\hat{FPR}_{W}$	12.6±1.3%	8.4±1.0%	11.1±0.6%
$\Delta\hat{FNR}$	2.4±0.0%	4.0±1.1%	2.8±0.8%
$\hat{FNR}_{NW}$	18.3±1.1%	16.5±1.9%	18.0±1.3%
$\hat{FNR}_{W}$	20.7±1.1%	20.5±3.0%	20.8±2.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	1.1±0.9%	0.5±0.0%
$\hat{AR}_{NW}$	23.2±1.3%	4.3±0.5%
$\hat{AR}_{W}$	22.1±2.2%	3.8±0.5%

Figure 8: **Logistic regression** on `COMPAS`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	10.1±0.6%	22.9±1.7%	15.8±0.5%
$\hat{PR}_{NW}$	47.9±0.7%	43.4±3.1%	48.5±1.2%
$\hat{PR}_{W}$	37.8±1.3%	20.5±1.4%	32.7±1.7%
$\Delta\hat{Err}$	0.6±0.9%	1.7±0.7%	1.2±0.8%
$\hat{Err}_{NW}$	38.8±0.5%	24.0±0.9%	32.8±0.4%
$\hat{Err}_{W}$	38.2±1.4%	22.3±1.6%	31.6±1.2%
$\Delta\hat{FPR}$	0.2±0.4%	4.0±0.4%	2.5±0.9%
$\hat{FPR}_{NW}$	18.8±0.8%	10.4±1.8%	16.1±0.9%
$\hat{FPR}_{W}$	18.6±1.2%	6.4±1.4%	13.6±1.8%
$\Delta\hat{FNR}$	0.3±0.3%	2.3±1.3%	1.4±0.1%
$\hat{FNR}_{NW}$	19.9±0.7%	13.6±1.0%	16.6±1.3%
$\hat{FNR}_{W}$	19.6±1.0%	15.9±2.3%	18.0±1.2%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	1.9±1.0%	2.3±0.1%
$\hat{AR}_{NW}$	62.3±1.8%	12.3±0.8%
$\hat{AR}_{W}$	64.2±2.8%	14.6±0.9%

Figure 9: **Decision trees** on `COMPAS`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	13.0±0.7%	24.3±0.4%	18.6±0.5%
$\hat{\mathrm{PR}}_{\mathbf{NW}}$	48.0±0.6%	45.6±1.7%	47.8±0.9%
$\hat{\mathrm{PR}}_{\mathbf{W}}$	35.0±1.3%	21.3±1.3%	29.2±1.4%
$\Delta\hat{\mathrm{Err}}$	1.0±0.8%	0.6±0.8%	2.1±1.0%
$\hat{\mathrm{Err}}_{\mathbf{NW}}$	36.9±0.5%	23.3±0.8%	32.3±0.4%
$\hat{\mathrm{Err}}_{\mathbf{W}}$	35.9±1.3%	23.9±1.6%	30.2±1.4%
$\Delta\hat{\mathrm{FPR}}$	2.0±0.4%	3.2±0.0%	4.5±0.4%
$\hat{\mathrm{FPR}}_{\mathbf{NW}}$	18.0±0.8%	10.0±1.3%	15.3±1.2%
$\hat{\mathrm{FPR}}_{\mathbf{W}}$	16.0±1.2%	6.8±1.3%	10.8±0.8%
$\Delta\hat{\mathrm{FNR}}$	0.9±0.4%	3.7±1.2%	2.4±0.8%
$\hat{\mathrm{FNR}}_{\mathbf{NW}}$	19.0±0.7%	13.4±1.2%	16.9±1.2%
$\hat{\mathrm{FNR}}_{\mathbf{W}}$	19.9±1.1%	17.1±2.4%	19.3±2.0%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	0.3±0.6%	0.2±0.7%
$\hat{\mathrm{AR}}_{\mathbf{NW}}$	53.9±1.6%	10.6±0.5%
$\hat{\mathrm{AR}}_{\mathbf{W}}$	53.6±2.2%	10.8±1.2%

Figure 10: **Random forests** on COMPAS

E.4.2 Old Adult

$\hat{\mathrm{SC}}$ CDFs for Old Adult (g = sex) and associated error metrics on the prediction set. **Baseline** metrics computed with B = 101 models. For **simple**, B = 101 models; for **super**, B = 101 ensemble models, each composed of 51 underlying models. We repeat for 10 test/train splits. We also report abstention rate $\hat{\mathrm{AR}}$.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	18.3±0.2%	17.8±0.1%	18.1±0.1%
$\hat{\mathrm{PR}}_{\mathbf{F}}$	8.2±0.3%	7.1±0.4%	7.6±0.4%
$\hat{\mathrm{PR}}_{\mathbf{M}}$	26.5±0.5%	24.9±0.5%	25.7±0.5%
$\Delta\hat{\mathrm{Err}}$	11.3±0.1%	10.8±0.1%	11.4±0.2%
$\hat{\mathrm{Err}}_{\mathbf{F}}$	7.8±0.4%	7.0±0.3%	7.5±0.2%
$\hat{\mathrm{Err}}_{\mathbf{M}}$	19.1±0.3%	17.8±0.4%	18.9±0.4%
$\Delta\hat{\mathrm{FPR}}$	4.7±0.0%	4.4±0.2%	4.8±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{F}}$	2.3±0.3%	1.6±0.1%	1.8±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{M}}$	7.0±0.3%	6.0±0.3%	6.6±0.3%
$\Delta\hat{\mathrm{FNR}}$	6.7±0.1%	6.5±0.1%	6.6±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{F}}$	5.5±0.3%	5.4±0.2%	5.7±0.2%
$\hat{\mathrm{FNR}}_{\mathbf{M}}$	12.2±0.2%	11.9±0.1%	12.3±0.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	2.6±0.0%	0.5±0.0%
$\hat{\mathrm{AR}}_{\mathbf{F}}$	1.8±0.2%	0.3±0.1%
$\hat{\mathrm{AR}}_{\mathbf{M}}$	4.4±0.2%	0.8±0.1%

Figure 11: **Logistic regression** on Old Adult

Decision tree prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\mathrm{PR}}$	20.3±0.1%	18.2±0.1%	19.9±0.1%
$\hat{\mathrm{PR}}_{\mathbf{F}}$	12.1±0.4%	4.5±0.3%	7.8±0.5%
$\hat{\mathrm{PR}}_{\mathbf{M}}$	32.4±0.5%	22.7±0.2%	27.7±0.4%
$\Delta\hat{\mathrm{Err}}$	12.3±0.0%	6.0±0.1%	10.9±0.2%
$\hat{\mathrm{Err}}_{\mathbf{F}}$	10.8±0.3%	3.0±0.2%	6.6±0.4%
$\hat{\mathrm{Err}}_{\mathbf{M}}$	23.1±0.3%	9.0±0.3%	17.5±0.2%
$\Delta\hat{\mathrm{FPR}}$	6.2±0.1%	2.5±0.1%	5.4±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{F}}$	5.7±0.2%	0.4±0.0%	1.9±0.3%
$\hat{\mathrm{FPR}}_{\mathbf{M}}$	11.9±0.3%	2.9±0.1%	7.3±0.1%
$\Delta\hat{\mathrm{FNR}}$	6.1±0.1%	3.4±0.0%	5.5±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{F}}$	5.1±0.3%	2.7±0.2%	4.7±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{M}}$	11.2±0.2%	6.1±0.2%	10.2±0.2%

Abstention set metrics

	Simple	Super
$\Delta\hat{\mathrm{AR}}$	19.5±0.2%	3.5±0.1%
$\hat{\mathrm{AR}}_{\mathbf{F}}$	18.2±0.4%	3.4±0.2%
$\hat{\mathrm{AR}}_{\mathbf{M}}$	37.7±0.6%	6.9±0.3%

Figure 12: **Decision trees** on `Old Adult`

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\mathrm{PR}}$	20.0±0.2%	17.1±0.3%	19.0±0.2%
$\hat{\mathrm{PR}}_{\mathbf{F}}$	9.8±0.2%	4.8±0.2%	7.7±0.3%
$\hat{\mathrm{PR}}_{\mathbf{M}}$	29.8±0.4%	21.9±0.5%	26.7±0.5%
$\Delta\hat{\mathrm{Err}}$	12.2±0.0%	6.5±0.1%	10.7±0.0%
$\hat{\mathrm{Err}}_{\mathbf{F}}$	9.0±0.3%	4.2±0.2%	6.6±0.2%
$\hat{\mathrm{Err}}_{\mathbf{M}}$	21.2±0.3%	10.7±0.3%	17.3±0.2%
$\Delta\hat{\mathrm{FPR}}$	6.0±0.2%	2.5±0.1%	5.0±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{F}}$	3.7±0.1%	0.7±0.1%	1.7±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{M}}$	9.7±0.3%	3.2±0.2%	6.7±0.3%
$\Delta\hat{\mathrm{FNR}}$	6.3±0.2%	4.1±0.2%	5.8±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{F}}$	5.3±0.3%	3.5±0.1%	4.9±0.2%
$\hat{\mathrm{FNR}}_{\mathbf{M}}$	11.6±0.1%	7.6±0.3%	10.7±0.3%

Abstention set metrics

	Simple	Super
$\Delta\hat{\mathrm{AR}}$	17.2±0.4%	3.4±0.1%
$\hat{\mathrm{AR}}_{\mathbf{F}}$	11.2±0.3%	2.0±0.3%
$\hat{\mathrm{AR}}_{\mathbf{M}}$	28.4±0.7%	5.4±0.2%

Figure 13: **Random forests** on `Old Adult`

E.4.3 South German Credit

$\hat{SC}$ CDFs for `German Credit` ($g = $ sex) and associated error metrics on the prediction set. **Baseline** metrics computed with $B = 101$ models. For **simple**, $B = 101$ models; for **super**, $B = 101$ ensemble models, each composed of 51 underlying models. We repeat for 10 test/train splits. We also report abstention rate $\hat{AR}$.
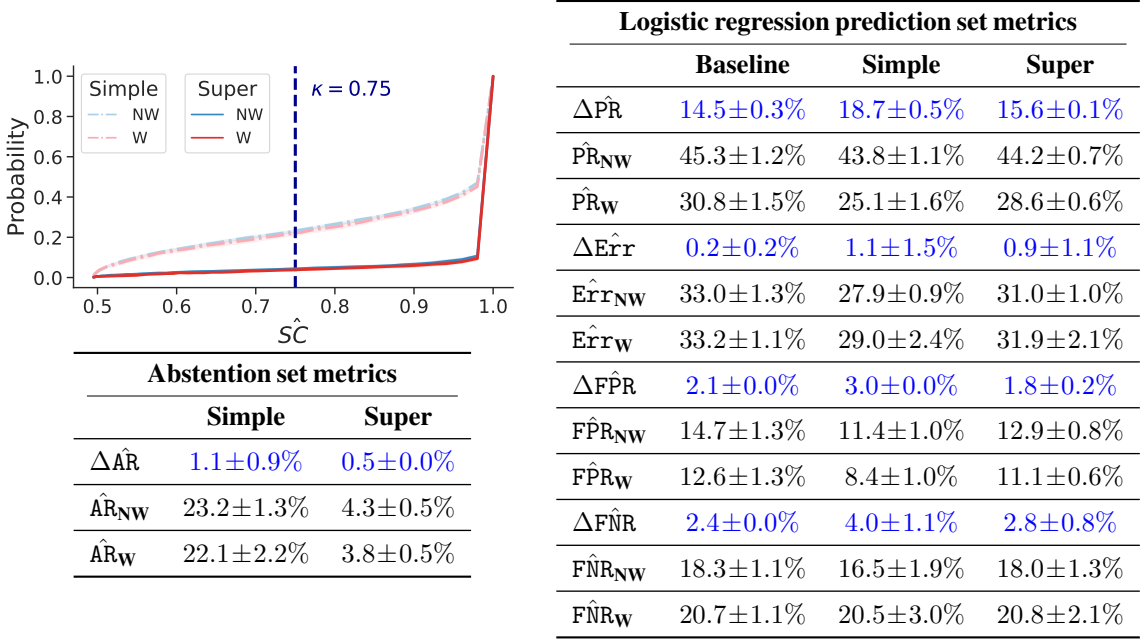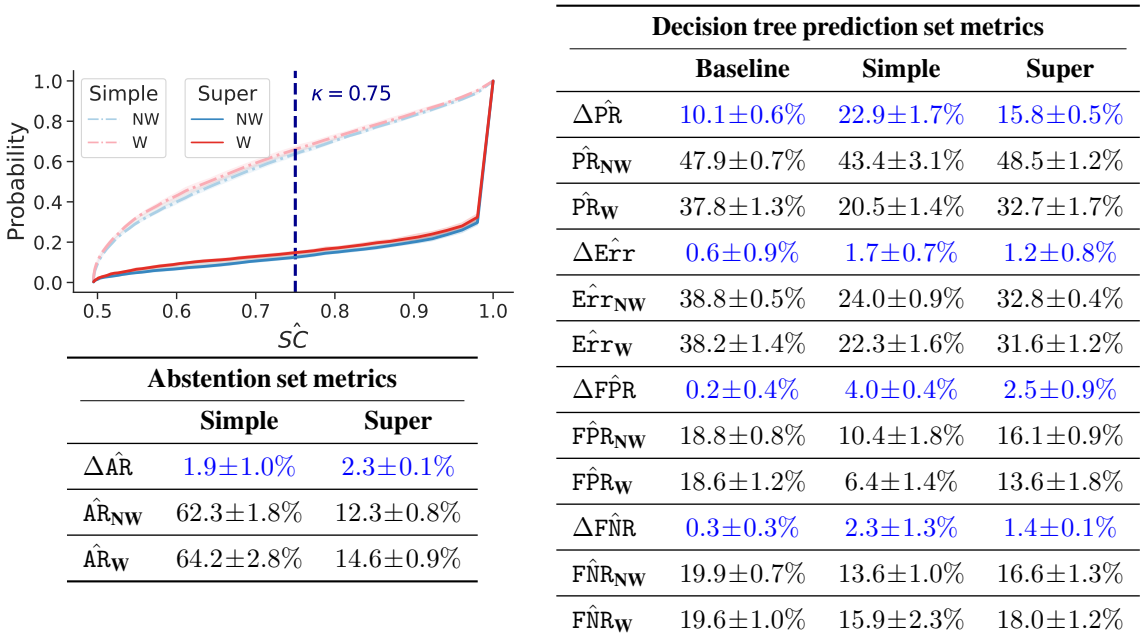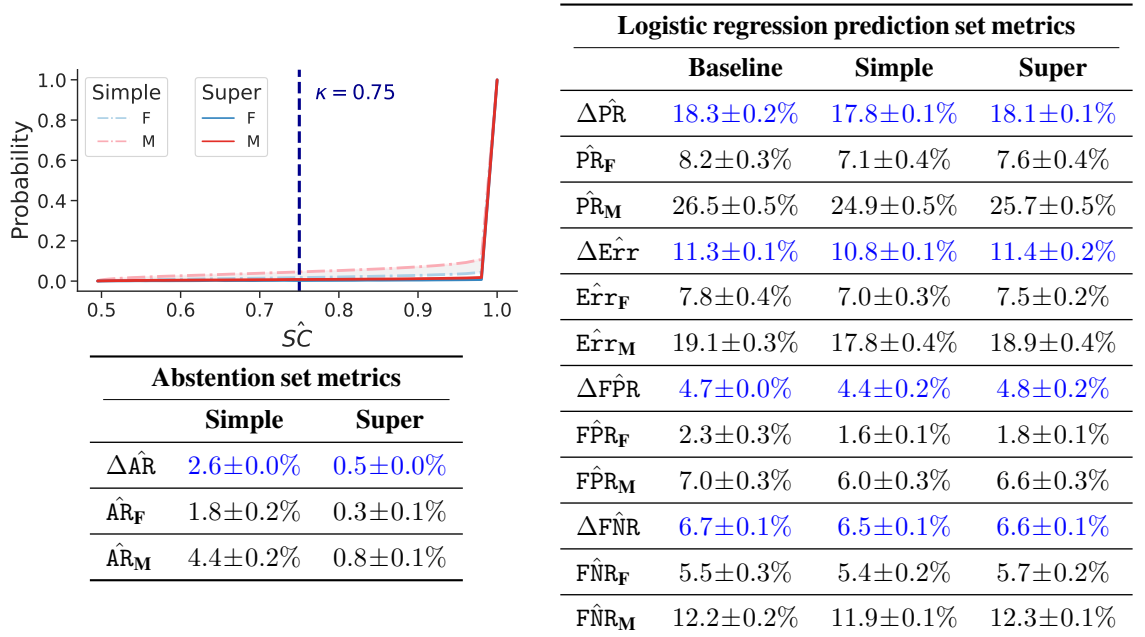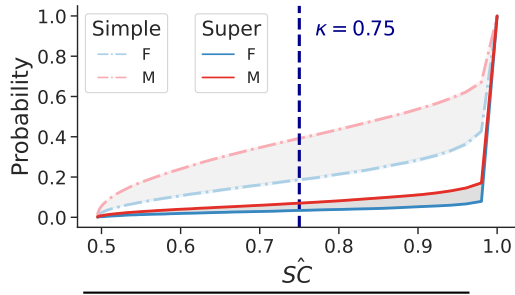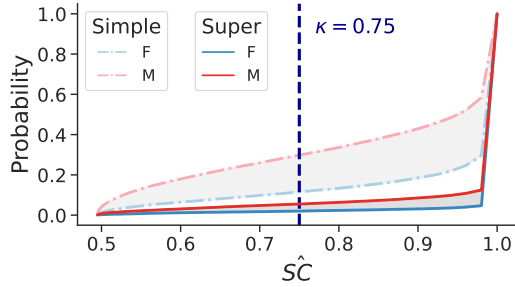

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.4±3.9%	0.1±1.8%
$\hat{AR}_\mathbf{F}$	20.8±6.6%	4.1±3.1%
$\hat{AR}_\mathbf{M}$	21.2±2.7%	4.0±1.3%

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	8.8±1.4%	9.1±1.2%	9.7±1.7%
$\hat{PR}_\mathbf{F}$	88.8±4.7%	96.0±4.1%	91.7±5.0%
$\hat{PR}_\mathbf{M}$	80.0±3.3%	86.9±2.9%	82.0±3.3%
$\Delta\hat{Err}$	0.9±4.0%	4.6±6.1%	3.4±4.7%
$\hat{Err}_\mathbf{F}$	23.3±6.9%	22.8±8.7%	25.5±7.4%
$\hat{Err}_\mathbf{M}$	24.2±2.9%	18.2±2.6%	22.1±2.7%
$\Delta\hat{FPR}$	0.7±3.9%	5.4±5.8%	5.5±5.4%
$\hat{FPR}_\mathbf{F}$	16.2±6.2%	19.6±8.4%	21.1±7.8%
$\hat{FPR}_\mathbf{M}$	15.5±2.3%	14.2±2.6%	15.6±2.4%
$\Delta\hat{FNR}$	1.6±1.1%	0.8±1.9%	2.1±1.8%
$\hat{FNR}_\mathbf{F}$	7.1±3.7%	3.2±3.5%	4.4±3.8%
$\hat{FNR}_\mathbf{M}$	8.7±2.6%	4.0±1.6%	6.5±2.0%
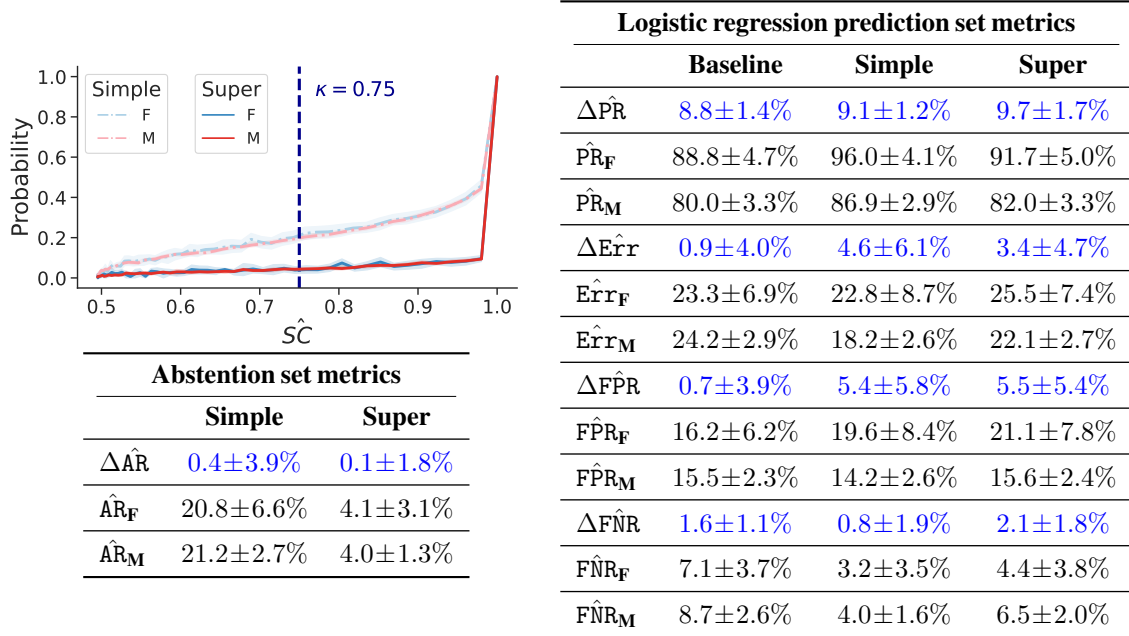
Figure 14: **Logistic regression** on `German Credit`

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.0±2.6%	2.8±2.6%
$\hat{AR}_\mathbf{F}$	65.2±6.0%	19.9±5.9%
$\hat{AR}_\mathbf{M}$	65.2±3.4%	17.1±3.3%

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	0.3±2.5%	1.9±0.4%	3.7±3.7%
$\hat{PR}_\mathbf{F}$	71.2±4.6%	99.6±0.8%	87.9±6.0%
$\hat{PR}_\mathbf{M}$	70.9±2.1%	97.7±1.2%	84.2±2.3%
$\Delta\hat{Err}$	1.1±2.9%	0.3±5.2%	0.1±4.9%
$\hat{Err}_\mathbf{F}$	33.0±4.8%	9.8±8.0%	20.3±8.2%
$\hat{Err}_\mathbf{M}$	31.9±1.9%	9.5±2.8%	20.2±3.3%
$\Delta\hat{FPR}$	0.7±3.5%	0.8±5.1%	0.4±3.9%
$\hat{FPR}_\mathbf{F}$	15.6±5.9%	9.6±7.9%	15.2±7.0%
$\hat{FPR}_\mathbf{M}$	14.9±2.4%	8.8±2.8%	14.8±3.1%
$\Delta\hat{FNR}$	0.5±2.3%	0.4±0.0%	0.2±3.6%
$\hat{FNR}_\mathbf{F}$	17.4±4.4%	0.2±0.7%	5.2±5.1%
$\hat{FNR}_\mathbf{M}$	16.9±2.1%	0.6±0.7%	5.4±1.5%

Figure 15: **Decision trees** on `German Credit`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	3.9±1.6%	1.9±0.6%	4.9±2.1%
$\hat{PR}_F$	81.7±3.3%	99.9±0.4%	94.5±4.0%
$\hat{PR}_M$	77.8±1.7%	98.0±1.0%	89.6±1.9%
$\Delta\hat{Err}$	2.3±3.1%	0.0±4.7%	2.4±6.7%
$\hat{Err}_F$	25.8±5.1%	11.9±7.8%	23.5±9.6%
$\hat{Err}_M$	28.1±2.0%	11.9±3.1%	21.1±2.9%
$\Delta\hat{FPR}$	2.5±3.0%	0.3±4.7%	2.2±6.2%
$\hat{FPR}_F$	13.8±4.8%	11.8±7.8%	20.5±9.1%
$\hat{FPR}_M$	16.3±1.8%	11.5±3.1%	18.3±2.9%
$\Delta\hat{FNR}$	0.2±1.7%	0.3±0.1%	0.1±2.3%
$\hat{FNR}_F$	11.9±3.1%	0.1±0.3%	3.0±3.5%
$\hat{FNR}_M$	11.7±1.4%	0.4±0.4%	2.9±1.2%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.1±2.8%	1.2±3.3%
$\hat{AR}_F$	47.5±6.7%	9.9±5.7%
$\hat{AR}_M$	47.4±3.9%	8.7±2.4%

Figure 16: **Random forests** on `German Credit`

E.4.4 `Taiwan Credit`

$\hat{SC}$ CDFs for `Taiwan Credit` ($g = \text{sex}$) and associated error metrics on the prediction set. **Baseline** metrics computed with $B = 101$ models. For **simple**, $B = 101$ models; for **super**, $B = 101$ ensemble models, each composed of 41 underlying models. We repeat for 10 test/train splits. We also report abstention rate $\hat{AR}$.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	1.5±0.1%	1.0±0.1%	1.0±0.1%
$\hat{PR}_F$	6.7±0.3%	6.2±0.1%	6.9±0.1%
$\hat{PR}_M$	8.2±0.4%	7.2±0.2%	7.9±0.2%
$\Delta\hat{Err}$	3.1±0.1%	3.1±0.3%	3.2±0.3%
$\hat{Err}_F$	17.8±0.5%	17.0±0.2%	17.5±0.3%
$\hat{Err}_M$	20.9±0.4%	20.1±0.5%	20.7±0.6%
$\Delta\hat{FPR}$	0.7±0.2%	0.3±0.0%	0.3±0.1%
$\hat{FPR}_F$	1.8±0.1%	1.7±0.1%	2.0±0.1%
$\hat{FPR}_M$	2.5±0.3%	2.0±0.1%	2.3±0.2%
$\Delta\hat{FNR}$	2.4±0.2%	2.7±0.4%	2.8±0.3%
$\hat{FNR}_F$	16.0±0.6%	15.3±0.2%	15.6±0.3%
$\hat{FNR}_M$	18.4±0.4%	18.0±0.6%	18.4±0.6%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.4±0.2%	0.0±0.2%
$\hat{AR}_F$	2.1±0.1%	0.4±0.0%
$\hat{AR}_M$	2.5±0.3%	0.4±0.2%
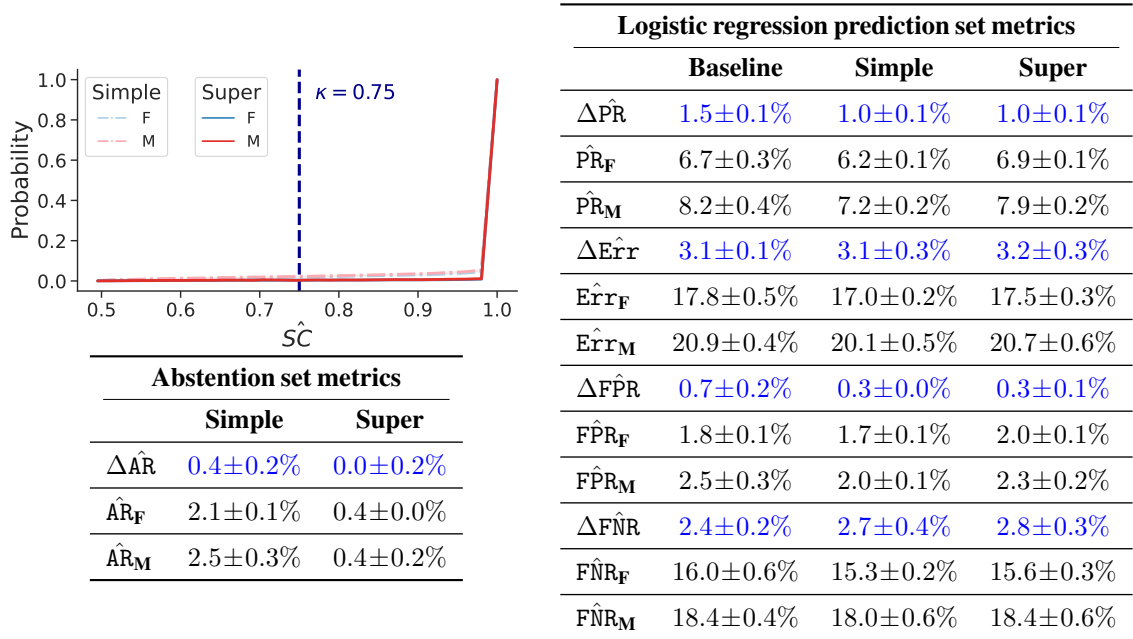
Figure 17: **Logistic regression** on `Taiwan Credit`

Decision tree prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\mathrm{PR}}$	2.1±0.1%	1.2±0.0%	2.0±0.2%
$\hat{\mathrm{PR}}_{\mathbf{F}}$	22.9±0.2%	3.0±0.4%	9.9±0.3%
$\hat{\mathrm{PR}}_{\mathbf{M}}$	25.0±0.3%	4.2±0.4%	11.9±0.5%
$\Delta\hat{\mathrm{Err}}$	2.3±0.1%	1.6±0.0%	2.5±0.1%
$\hat{\mathrm{Err}}_{\mathbf{F}}$	26.8±0.2%	9.6±0.4%	15.3±0.3%
$\hat{\mathrm{Err}}_{\mathbf{M}}$	29.1±0.3%	11.2±0.4%	17.8±0.4%
$\Delta\hat{\mathrm{FPR}}$	0.6±0.1%	0.2±0.1%	0.7±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{F}}$	14.4±0.2%	0.6±0.1%	3.0±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{M}}$	15.0±0.3%	0.8±0.2%	3.7±0.3%
$\Delta\hat{\mathrm{FNR}}$	1.7±0.2%	1.3±0.1%	1.9±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{F}}$	12.4±0.4%	9.0±0.4%	12.3±0.3%
$\hat{\mathrm{FNR}}_{\mathbf{M}}$	14.1±0.2%	10.3±0.5%	14.2±0.4%
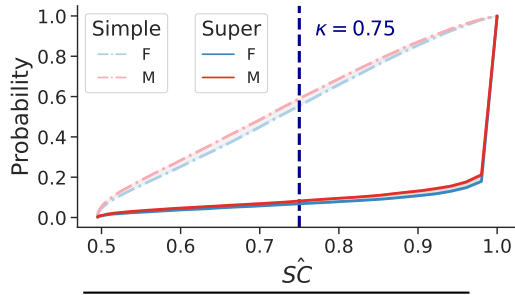
Abstention set metrics

	Simple	Super
$\Delta\hat{\mathrm{AR}}$	3.2±0.1%	1.3±0.1%
$\hat{\mathrm{AR}}_{\mathbf{F}}$	56.7±0.6%	6.7±0.2%
$\hat{\mathrm{AR}}_{\mathbf{M}}$	59.9±0.5%	8.0±0.1%

Figure 18: **Decision trees** on `Taiwan Credit`

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\mathrm{PR}}$	2.5±0.1%	1.0±0.1%	2.1±0.2%
$\hat{\mathrm{PR}}_{\mathbf{F}}$	14.9±0.2%	4.1±0.3%	10.3±0.2%
$\hat{\mathrm{PR}}_{\mathbf{M}}$	17.4±0.3%	5.1±0.2%	12.4±0.4%
$\Delta\hat{\mathrm{Err}}$	2.8±0.1%	1.9±0.0%	2.5±0.1%
$\hat{\mathrm{Err}}_{\mathbf{F}}$	20.5±0.3%	12.0±0.4%	15.8±0.4%
$\hat{\mathrm{Err}}_{\mathbf{M}}$	23.3±0.4%	13.9±0.4%	18.3±0.5%
$\Delta\hat{\mathrm{FPR}}$	1.0±0.1%	0.3±0.1%	0.6±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{F}}$	7.2±0.2%	0.9±0.1%	3.3±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{M}}$	8.2±0.3%	1.2±0.2%	3.9±0.2%
$\Delta\hat{\mathrm{FNR}}$	1.7±0.1%	1.6±0.1%	1.8±0.0%
$\hat{\mathrm{FNR}}_{\mathbf{F}}$	13.3±0.4%	11.0±0.3%	12.6±0.4%
$\hat{\mathrm{FNR}}_{\mathbf{M}}$	15.0±0.3%	12.6±0.4%	14.4±0.4%
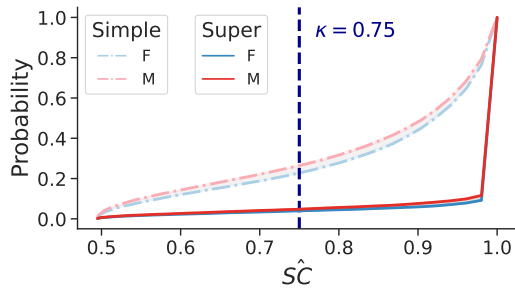
Abstention set metrics

	Simple	Super
$\Delta\hat{\mathrm{AR}}$	4.1±0.1%	0.8±0.0%
$\hat{\mathrm{AR}}_{\mathbf{F}}$	24.0±0.8%	3.9±0.3%
$\hat{\mathrm{AR}}_{\mathbf{M}}$	28.1±0.7%	4.7±0.3%

Figure 19: **Random forests** on `Taiwan Credit`

E.4.5 `New Adult - CA`

$\hat{\text{SC}}$ CDFs for three tasks (`Income`, `Employment`, `Public Coverage`) in `New Adult - CA`, using $g=$ `sex` and `race`, and associated error metrics on the prediction set. **Baseline** metrics computed with $B=101$ models. For **simple**, $B=101$ models; for **super**, $B=101$ ensemble models, each composed of 21 underlying models for `Income` and `Public Coverage`; 15 for `Employment`. We repeat for 5 test/train splits. We also report abstention rate $\hat{\text{AR}}$.
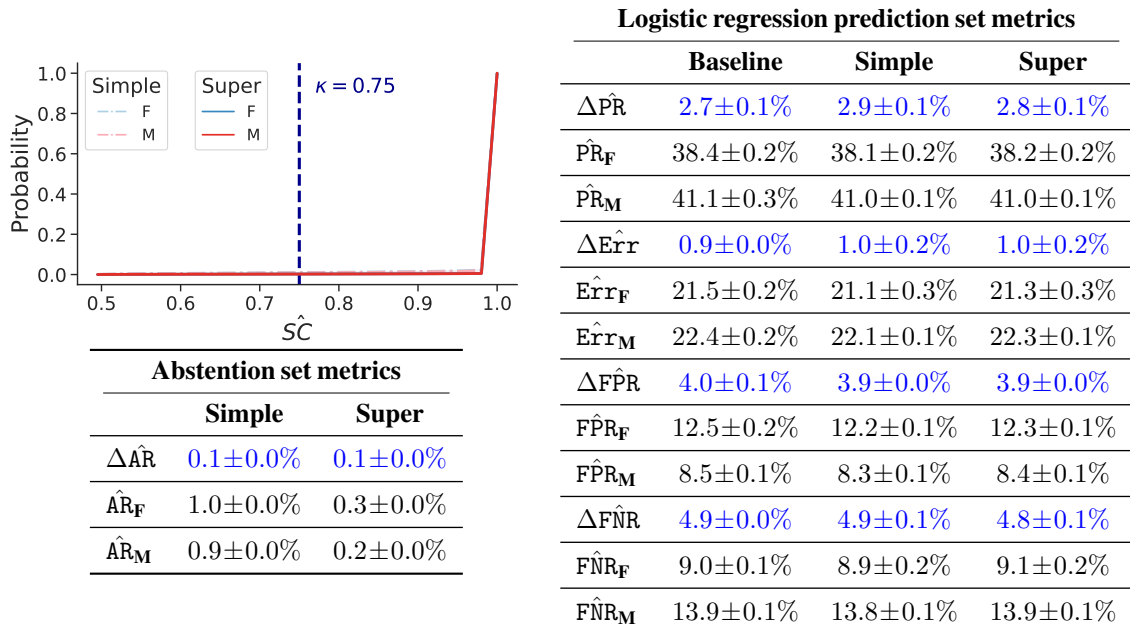
`Income` - **by** `sex`.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	2.7±0.1%	2.9±0.1%	2.8±0.1%
$\hat{\text{PR}}_{\textbf{F}}$	38.4±0.2%	38.1±0.2%	38.2±0.2%
$\hat{\text{PR}}_{\textbf{M}}$	41.1±0.3%	41.0±0.1%	41.0±0.1%
$\Delta\hat{\text{Err}}$	0.9±0.0%	1.0±0.2%	1.0±0.2%
$\hat{\text{Err}}_{\textbf{F}}$	21.5±0.2%	21.1±0.3%	21.3±0.3%
$\hat{\text{Err}}_{\textbf{M}}$	22.4±0.2%	22.1±0.1%	22.3±0.1%
$\Delta\hat{\text{FPR}}$	4.0±0.1%	3.9±0.0%	3.9±0.0%
$\hat{\text{FPR}}_{\textbf{F}}$	12.5±0.2%	12.2±0.1%	12.3±0.1%
$\hat{\text{FPR}}_{\textbf{M}}$	8.5±0.1%	8.3±0.1%	8.4±0.1%
$\Delta\hat{\text{FNR}}$	4.9±0.0%	4.9±0.1%	4.8±0.1%
$\hat{\text{FNR}}_{\textbf{F}}$	9.0±0.1%	8.9±0.2%	9.1±0.2%
$\hat{\text{FNR}}_{\textbf{M}}$	13.9±0.1%	13.8±0.1%	13.9±0.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	0.1±0.0%	0.1±0.0%
$\hat{\text{AR}}_{\textbf{F}}$	1.0±0.0%	0.3±0.0%
$\hat{\text{AR}}_{\textbf{M}}$	0.9±0.0%	0.2±0.0%

Figure 20: **Logistic regression** on `New Adult - CA - Income`, by `sex`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	7.5±0.1%	12.5±0.1%	9.7±0.1%
$\hat{\text{PR}}_{\textbf{F}}$	37.4±0.2%	26.8±0.4%	34.1±0.3%
$\hat{\text{PR}}_{\textbf{M}}$	44.9±0.1%	39.3±0.3%	43.8±0.2%
$\Delta\hat{\text{Err}}$	1.4±0.0%	1.0±0.0%	1.4±0.0%
$\hat{\text{Err}}_{\textbf{F}}$	24.4±0.1%	6.9±0.1%	14.5±0.2%
$\hat{\text{Err}}_{\textbf{M}}$	25.8±0.1%	7.9±0.1%	15.9±0.2%
$\Delta\hat{\text{FPR}}$	1.4±0.0%	0.1±0.1%	0.5±0.1%
$\hat{\text{FPR}}_{\textbf{F}}$	13.5±0.1%	3.6±0.1%	7.6±0.1%
$\hat{\text{FPR}}_{\textbf{M}}$	12.1±0.1%	3.5±0.2%	7.1±0.2%
$\Delta\hat{\text{FNR}}$	2.9±0.0%	1.1±0.0%	1.9±0.1%
$\hat{\text{FNR}}_{\textbf{F}}$	10.9±0.1%	3.3±0.1%	6.9±0.1%
$\hat{\text{FNR}}_{\textbf{M}}$	13.8±0.1%	4.4±0.1%	8.8±0.2%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	2.1±0.2%	0.8±0.0%
$\hat{\text{AR}}_{\textbf{F}}$	49.2±0.3%	13.3±0.2%
$\hat{\text{AR}}_{\textbf{M}}$	51.3±0.1%	14.1±0.2%
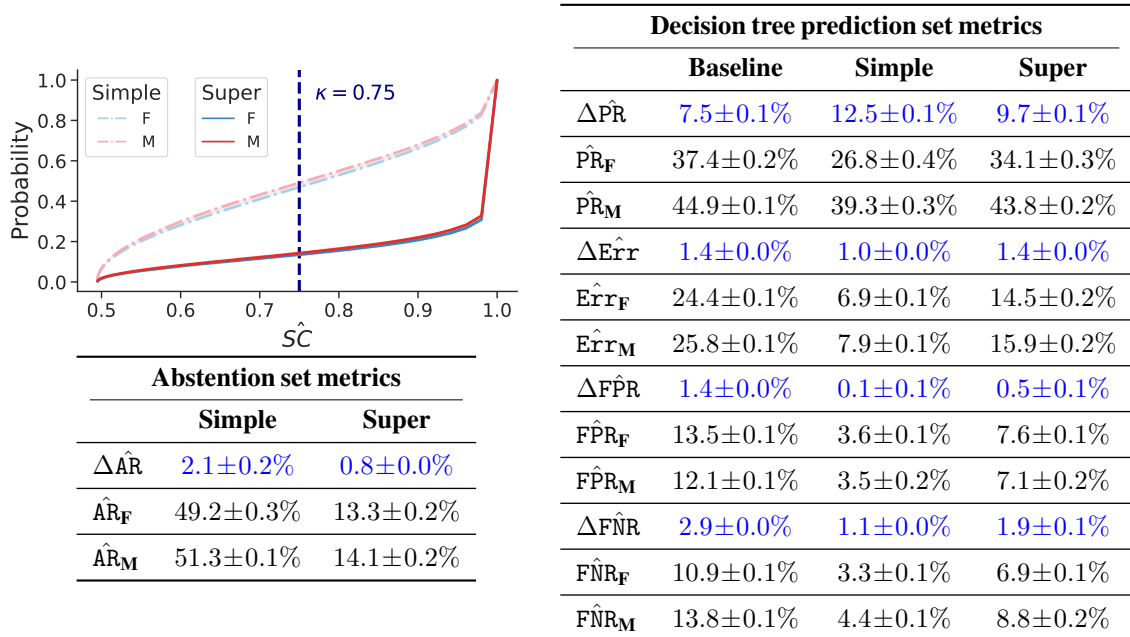
Figure 21: **Decision trees** on `New Adult - CA - Income`, by `sex`

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	7.4±0.1%	10.3±0.2%	8.6±0.1%
$\hat{PR}_F$	36.7±0.2%	30.6±0.4%	34.9±0.3%
$\hat{PR}_M$	44.1±0.1%	40.9±0.2%	43.5±0.2%
$\Delta\hat{Err}$	1.4±0.0%	1.2±0.0%	1.4±0.1%
$\hat{Err}_F$	21.0±0.1%	9.3±0.2%	15.3±0.1%
$\hat{Err}_M$	22.4±0.1%	10.5±0.2%	16.7±0.2%
$\Delta\hat{FPR}$	1.4±0.0%	0.5±0.0%	0.9±0.1%
$\hat{FPR}_F$	11.4±0.1%	4.9±0.1%	8.1±0.1%
$\hat{FPR}_M$	10.0±0.1%	4.4±0.1%	7.2±0.2%
$\Delta\hat{FNR}$	2.8±0.0%	1.8±0.0%	2.3±0.0%
$\hat{FNR}_F$	9.6±0.1%	4.4±0.1%	7.2±0.1%
$\hat{FNR}_M$	12.4±0.1%	6.2±0.1%	9.5±0.1%
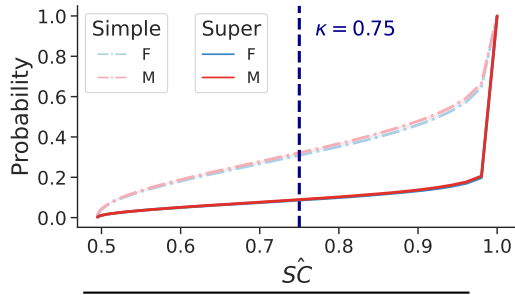
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	1.4±0.1%	0.2±0.0%
$\hat{AR}_F$	32.8±0.2%	8.6±0.1%
$\hat{AR}_M$	34.2±0.1%	8.8±0.1%

Figure 22: **Random forests** on `New Adult - CA - Income`, by `sex`

`Income` - **by** `race`.

Logistic regression prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	9.2±0.1%	9.2±0.3%	9.2±0.2%
$\hat{PR}_{NW}$	34.1±0.3%	33.9±0.3%	34.0±0.3%
$\hat{PR}_W$	43.3±0.2%	43.1±0.0%	43.2±0.1%
$\Delta\hat{Err}$	0.6±0.1%	0.4±0.1%	0.4±0.1%
$\hat{Err}_{NW}$	21.6±0.2%	21.4±0.2%	21.6±0.2%
$\hat{Err}_W$	22.2±0.1%	21.8±0.1%	22.0±0.1%
$\Delta\hat{FPR}$	0.6±0.1%	0.5±0.0%	0.5±0.0%
$\hat{FPR}_{NW}$	10.0±0.2%	9.8±0.1%	9.9±0.1%
$\hat{FPR}_W$	10.6±0.1%	10.3±0.1%	10.4±0.1%
$\Delta\hat{FNR}$	0.0±0.1%	0.1±0.2%	0.1±0.3%
$\hat{FNR}_{NW}$	11.6±0.2%	11.6±0.3%	11.7±0.3%
$\hat{FNR}_W$	11.6±0.1%	11.5±0.1%	11.6±0.0%
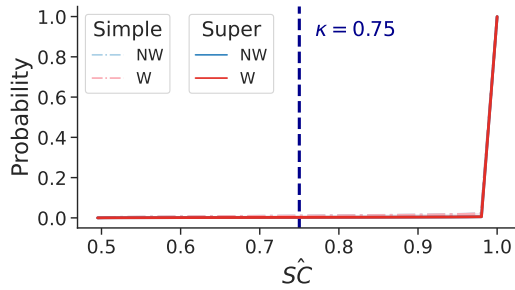
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	0.0±0.0%	0.0±0.0%
$\hat{AR}_{NW}$	1.0±0.0%	0.2±0.0%
$\hat{AR}_W$	1.0±0.0%	0.2±0.0%

Figure 23: **Logistic regression** on `New Adult - CA - Income`, by `race`
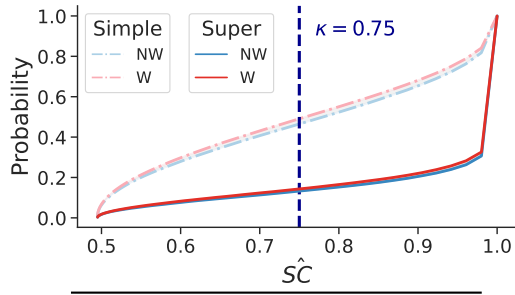
Abstention set metrics

	Simple	Super
$\Delta\hat{\text{AR}}$	2.5±0.1%	1.0±0.1%
$\hat{\text{AR}}_{\text{NW}}$	48.8±0.3%	13.1±0.2%
$\hat{\text{AR}}_{\text{W}}$	51.3±0.2%	14.1±0.1%

Decision tree prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\text{PR}}$	7.0±0.0%	10.3±0.0%	9.9±0.1%
$\hat{\text{PR}}_{\text{NW}}$	37.0±0.2%	27.0±0.2%	33.1±0.3%
$\hat{\text{PR}}_{\text{W}}$	44.0±0.2%	37.3±0.2%	43.0±0.2%
$\Delta\hat{\text{Err}}$	1.1±0.0%	0.2±0.1%	0.6±0.0%
$\hat{\text{Err}}_{\text{NW}}$	24.5±0.1%	7.3±0.0%	14.8±0.2%
$\hat{\text{Err}}_{\text{W}}$	25.6±0.1%	7.5±0.1%	15.4±0.2%
$\Delta\hat{\text{FPR}}$	0.2±0.0%	0.5±0.0%	0.7±0.0%
$\hat{\text{FPR}}_{\text{NW}}$	12.9±0.1%	3.2±0.1%	6.9±0.1%
$\hat{\text{FPR}}_{\text{W}}$	12.7±0.1%	3.7±0.1%	7.6±0.1%
$\Delta\hat{\text{FNR}}$	1.4±0.0%	0.3±0.0%	0.2±0.0%
$\hat{\text{FNR}}_{\text{NW}}$	11.6±0.1%	4.1±0.1%	8.0±0.1%
$\hat{\text{FNR}}_{\text{W}}$	13.0±0.1%	3.8±0.1%	7.8±0.1%

Figure 24: **Decision trees** on `New Adult - CA - Income`, by `race`
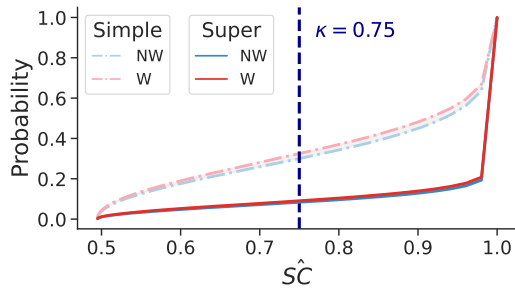
Abstention set metrics

	Simple	Super
$\Delta\hat{\text{AR}}$	2.4±0.2%	0.8±0.0%
$\hat{\text{AR}}_{\text{NW}}$	32.1±0.3%	8.2±0.1%
$\hat{\text{AR}}_{\text{W}}$	34.5±0.1%	9.0±0.1%

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{\text{PR}}$	8.4±0.0%	11.0±0.0%	10.1±0.2%
$\hat{\text{PR}}_{\text{NW}}$	35.4±0.2%	29.3±0.2%	33.2±0.3%
$\hat{\text{PR}}_{\text{W}}$	43.8±0.2%	40.3±0.2%	43.3±0.1%
$\Delta\hat{\text{Err}}$	1.2±0.0%	0.2±0.1%	0.5±0.0%
$\hat{\text{Err}}_{\text{NW}}$	21.0±0.1%	9.8±0.1%	15.7±0.2%
$\hat{\text{Err}}_{\text{W}}$	22.2±0.1%	10.0±0.2%	16.2±0.2%
$\Delta\hat{\text{FPR}}$	0.6±0.0%	0.7±0.0%	0.9±0.1%
$\hat{\text{FPR}}_{\text{NW}}$	10.3±0.1%	4.2±0.1%	7.1±0.2%
$\hat{\text{FPR}}_{\text{W}}$	10.9±0.1%	4.9±0.1%	8.0±0.1%
$\Delta\hat{\text{FNR}}$	0.7±0.0%	0.5±0.0%	0.3±0.1%
$\hat{\text{FNR}}_{\text{NW}}$	10.7±0.1%	5.6±0.1%	8.6±0.2%
$\hat{\text{FNR}}_{\text{W}}$	11.4±0.1%	5.1±0.1%	8.3±0.1%

Figure 25: **Random forests** on `New Adult - CA - Income`, by `race`

Employment - **by** sex.
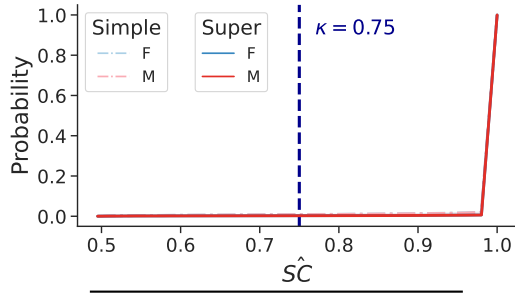
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.1±0.0%	0.0±0.0%
$\hat{AR}_\mathbf{F}$	0.8±0.0%	0.3±0.0%
$\hat{AR}_\mathbf{M}$	0.7±0.0%	0.3±0.0%

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	4.3±0.1%	4.5±0.0%	4.4±0.0%
$\hat{PR}_\mathbf{F}$	56.6±0.1%	56.8±0.1%	56.7±0.1%
$\hat{PR}_\mathbf{M}$	52.3±0.2%	52.3±0.1%	52.3±0.1%
$\Delta\hat{Err}$	5.0±0.0%	4.9±0.0%	4.9±0.0%
$\hat{Err}_\mathbf{F}$	25.8±0.1%	25.5±0.1%	25.6±0.1%
$\hat{Err}_\mathbf{M}$	20.8±0.1%	20.6±0.1%	20.7±0.1%
$\Delta\hat{FPR}$	8.1±0.0%	8.1±0.0%	8.1±0.0%
$\hat{FPR}_\mathbf{F}$	20.1±0.1%	20.1±0.0%	20.1±0.0%
$\hat{FPR}_\mathbf{M}$	12.0±0.1%	12.0±0.0%	12.0±0.0%
$\Delta\hat{FNR}$	3.1±0.0%	3.2±0.1%	3.2±0.1%
$\hat{FNR}_\mathbf{F}$	5.7±0.1%	5.4±0.0%	5.5±0.0%
$\hat{FNR}_\mathbf{M}$	8.8±0.1%	8.6±0.1%	8.7±0.1%

Figure 26: **Logistic regression** on New Adult - CA - Employment, by sex
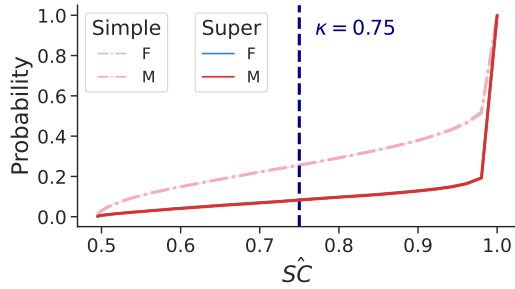
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.2±0.0%	0.1±0.1%
$\hat{AR}_\mathbf{F}$	22.5±0.1%	8.2±0.2%
$\hat{AR}_\mathbf{M}$	22.3±0.1%	8.1±0.3%

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	0.5±0.0%	0.7±0.2%	0.6±0.1%
$\hat{PR}_\mathbf{F}$	50.3±0.2%	50.0±0.3%	51.2±0.2%
$\hat{PR}_\mathbf{M}$	49.8±0.2%	49.3±0.1%	50.6±0.1%
$\Delta\hat{Err}$	4.8±0.0%	5.8±0.1%	5.8±0.1%
$\hat{Err}_\mathbf{F}$	24.8±0.1%	17.8±0.1%	20.9±0.2%
$\hat{Err}_\mathbf{M}$	20.0±0.1%	12.0±0.0%	15.1±0.1%
$\Delta\hat{FPR}$	6.1±0.0%	6.2±0.1%	6.4±0.1%
$\hat{FPR}_\mathbf{F}$	16.5±0.1%	13.8±0.1%	15.2±0.1%
$\hat{FPR}_\mathbf{M}$	10.4±0.1%	7.6±0.0%	8.8±0.0%
$\Delta\hat{FNR}$	1.3±0.0%	0.3±0.0%	0.7±0.1%
$\hat{FNR}_\mathbf{F}$	8.3±0.1%	4.0±0.1%	5.6±0.2%
$\hat{FNR}_\mathbf{M}$	9.6±0.1%	4.3±0.1%	6.3±0.1%

Figure 27: **Decision trees** on New Adult - CA - Employment, by sex

45

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	$0.3\pm0.1\%$	$0.5\pm0.1\%$	$0.6\pm0.1\%$
$\hat{PR}_{\mathbf{F}}$	$49.2\pm0.1\%$	$48.7\pm0.2\%$	$50.3\pm0.2\%$
$\hat{PR}_{\mathbf{M}}$	$48.9\pm0.2\%$	$48.2\pm0.1\%$	$49.7\pm0.1\%$
$\Delta\hat{Err}$	$4.8\pm0.0\%$	$5.4\pm0.0\%$	$5.5\pm0.2\%$
$\hat{Err}_{\mathbf{F}}$	$24.0\pm0.1\%$	$17.5\pm0.1\%$	$20.5\pm0.2\%$
$\hat{Err}_{\mathbf{M}}$	$19.2\pm0.1\%$	$12.1\pm0.1\%$	$15.0\pm0.0\%$
$\Delta\hat{FPR}$	$6.0\pm0.0\%$	$5.9\pm0.1\%$	$6.2\pm0.1\%$
$\hat{FPR}_{\mathbf{F}}$	$15.5\pm0.1\%$	$13.2\pm0.1\%$	$14.7\pm0.1\%$
$\hat{FPR}_{\mathbf{M}}$	$9.5\pm0.1\%$	$7.3\pm0.0\%$	$8.5\pm0.0\%$
$\Delta\hat{FNR}$	$1.2\pm0.0\%$	$0.4\pm0.0\%$	$0.8\pm0.0\%$
$\hat{FNR}_{\mathbf{F}}$	$8.5\pm0.1\%$	$4.3\pm0.1\%$	$5.8\pm0.1\%$
$\hat{FNR}_{\mathbf{M}}$	$9.7\pm0.1\%$	$4.7\pm0.1\%$	$6.6\pm0.1\%$
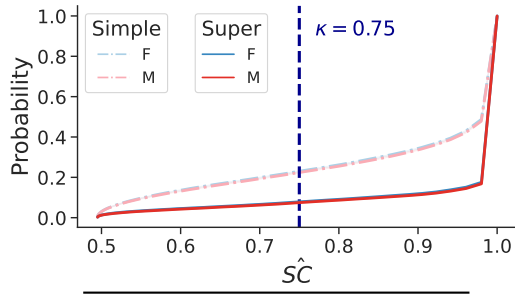
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	$0.7\pm0.0\%$	$0.4\pm0.0\%$
$\hat{AR}_{\mathbf{F}}$	$20.3\pm0.2\%$	$7.7\pm0.2\%$
$\hat{AR}_{\mathbf{M}}$	$19.6\pm0.2\%$	$7.3\pm0.2\%$

Figure 28: **Random forests** on `New Adult - CA - Employment`, by `sex`

`Employment` - **by** `race`**.**

Logistic regression prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	$1.1\pm0.2\%$	$1.2\pm0.1\%$	$1.2\pm0.1\%$
$\hat{PR}_{\mathbf{NW}}$	$55.2\pm0.3\%$	$55.3\pm0.2\%$	$55.3\pm0.2\%$
$\hat{PR}_{\mathbf{W}}$	$54.1\pm0.1\%$	$54.1\pm0.1\%$	$54.1\pm0.1\%$
$\Delta\hat{Err}$	$0.1\pm0.0\%$	$0.2\pm0.1\%$	$0.2\pm0.0\%$
$\hat{Err}_{\mathbf{NW}}$	$23.3\pm0.1\%$	$23.0\pm0.0\%$	$23.1\pm0.1\%$
$\hat{Err}_{\mathbf{W}}$	$23.4\pm0.1\%$	$23.2\pm0.1\%$	$23.3\pm0.1\%$
$\Delta\hat{FPR}$	$0.8\pm0.0\%$	$0.7\pm0.1\%$	$0.7\pm0.1\%$
$\hat{FPR}_{\mathbf{NW}}$	$16.6\pm0.1\%$	$16.5\pm0.0\%$	$16.6\pm0.0\%$
$\hat{FPR}_{\mathbf{W}}$	$15.8\pm0.1\%$	$15.8\pm0.1\%$	$15.9\pm0.1\%$
$\Delta\hat{FNR}$	$1.0\pm0.0\%$	$1.0\pm0.0\%$	$0.9\pm0.0\%$
$\hat{FNR}_{\mathbf{NW}}$	$6.6\pm0.1\%$	$6.4\pm0.1\%$	$6.5\pm0.1\%$
$\hat{FNR}_{\mathbf{W}}$	$7.6\pm0.1\%$	$7.4\pm0.1\%$	$7.4\pm0.1\%$
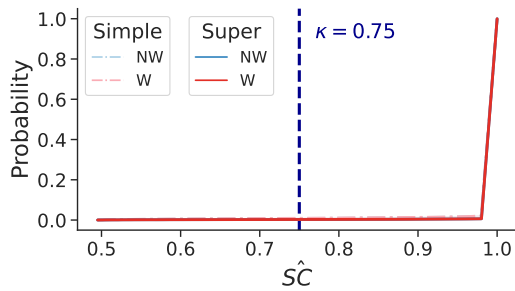
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	$0.1\pm0.0\%$	$0.0\pm0.0\%$
$\hat{AR}_{\mathbf{NW}}$	$0.7\pm0.0\%$	$0.3\pm0.0\%$
$\hat{AR}_{\mathbf{W}}$	$0.8\pm0.0\%$	$0.3\pm0.0\%$

Figure 29: **Logistic regression** on `New Adult - CA - Employment`, by `race`

Decision tree prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	0.2±0.2%	0.6±0.0%	1.2±0.1%
$\hat{PR}_{NW}$	50.2±0.3%	50.0±0.1%	51.6±0.0%
$\hat{PR}_W$	50.0±0.1%	49.4±0.1%	50.4±0.1%
$\Delta\hat{Err}$	0.6±0.0%	0.7±0.0%	0.6±0.0%
$\hat{Err}_{NW}$	22.1±0.1%	14.5±0.1%	17.7±0.1%
$\hat{Err}_W$	22.7±0.1%	15.2±0.1%	18.3±0.1%
$\Delta\hat{FPR}$	0.1±0.0%	0.6±0.1%	0.7±0.1%
$\hat{FPR}_{NW}$	13.5±0.1%	11.1±0.0%	12.5±0.0%
$\hat{FPR}_W$	13.4±0.1%	10.5±0.1%	11.8±0.1%
$\Delta\hat{FNR}$	0.6±0.0%	1.3±0.0%	1.3±0.0%
$\hat{FNR}_{NW}$	8.6±0.1%	3.4±0.1%	5.2±0.1%
$\hat{FNR}_W$	9.2±0.1%	4.7±0.1%	6.5±0.1%
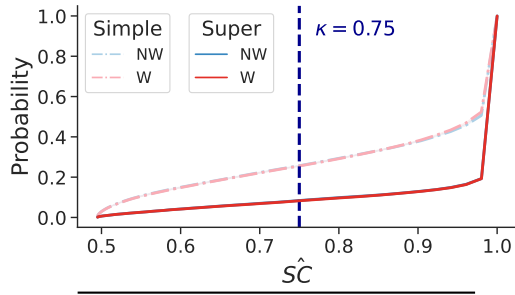
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	0.0±0.1%	0.1±0.1%
$\hat{AR}_{NW}$	22.4±0.2%	8.2±0.2%
$\hat{AR}_W$	22.4±0.1%	8.1±0.3%

Figure 30: **Decision trees** on `New Adult - CA - Employment`, by `race`

Random forest prediction set metrics

	Baseline	Simple	Super
$\Delta\hat{PR}$	0.6±0.2%	1.0±0.0%	1.4±0.1%
$\hat{PR}_{NW}$	49.4±0.3%	49.1±0.1%	50.9±0.0%
$\hat{PR}_W$	48.8±0.1%	48.1±0.1%	49.5±0.1%
$\Delta\hat{Err}$	0.5±0.0%	0.7±0.0%	0.5±0.0%
$\hat{Err}_{NW}$	21.3±0.1%	14.4±0.1%	17.5±0.1%
$\hat{Err}_W$	21.8±0.1%	15.1±0.1%	18.0±0.1%
$\Delta\hat{FPR}$	0.3±0.1%	0.6±0.1%	0.8±0.1%
$\hat{FPR}_{NW}$	12.7±0.2%	10.7±0.0%	12.1±0.0%
$\hat{FPR}_W$	12.4±0.1%	10.1±0.1%	11.3±0.1%
$\Delta\hat{FNR}$	0.8±0.0%	1.3±0.0%	1.3±0.0%
$\hat{FNR}_{NW}$	8.6±0.1%	3.7±0.1%	5.4±0.1%
$\hat{FNR}_W$	9.4±0.1%	5.0±0.1%	6.7±0.1%
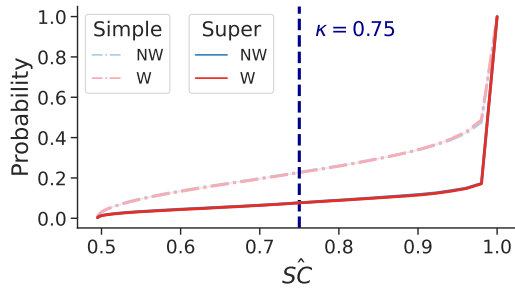
Abstention set metrics

	Simple	Super
$\Delta\hat{AR}$	0.3±0.1%	0.1±0.1%
$\hat{AR}_{NW}$	20.1±0.1%	7.6±0.1%
$\hat{AR}_W$	19.8±0.2%	7.5±0.2%

Figure 31: **Random forests** on `New Adult - CA - Employment`, by `race`

`Public Coverage` - **by** `sex`.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	2.6±0.1%	2.6±0.1%	2.6±0.1%
$\hat{PR}_F$	15.1±0.2%	14.7±0.2%	15.1±0.2%
$\hat{PR}_M$	17.7±0.3%	17.3±0.1%	17.7±0.1%
$\Delta\hat{Err}$	0.9±0.1%	0.5±0.0%	0.5±0.0%
$\hat{Err}_F$	31.2±0.3%	30.8±0.2%	31.0±0.2%
$\hat{Err}_M$	32.1±0.2%	31.3±0.2%	31.5±0.2%
$\Delta\hat{FPR}$	0.0±0.0%	0.0±0.0%	0.0±0.1%
$\hat{FPR}_F$	5.5±0.1%	5.1±0.1%	5.3±0.1%
$\hat{FPR}_M$	5.5±0.1%	5.1±0.1%	5.3±0.2%
$\Delta\hat{FNR}$	0.9±0.1%	0.5±0.0%	0.5±0.0%
$\hat{FNR}_F$	25.7±0.3%	25.7±0.2%	25.7±0.2%
$\hat{FNR}_M$	26.6±0.2%	26.2±0.2%	26.2±0.2%
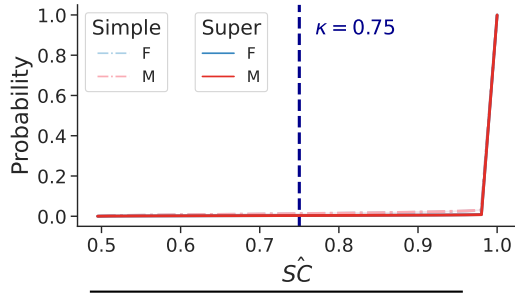
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.0±0.1%	0.0±0.0%
$\hat{AR}_F$	1.4±0.1%	0.4±0.0%
$\hat{AR}_M$	1.4±0.0%	0.4±0.0%

Figure 32: **Logistic regression** on `New Adult - CA - Public Coverage`, by `sex`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	2.6±0.2%	6.8±0.1%	3.0±0.0%
$\hat{PR}_F$	35.5±0.2%	20.5±0.4%	27.5±0.3%
$\hat{PR}_M$	38.1±0.4%	27.3±0.3%	30.5±0.3%
$\Delta\hat{Err}$	0.1±0.1%	0.5±0.1%	0.2±0.1%
$\hat{Err}_F$	35.1±0.2%	18.8±0.3%	26.7±0.3%
$\hat{Err}_M$	35.2±0.1%	19.3±0.2%	26.9±0.4%
$\Delta\hat{FPR}$	0.4±0.1%	0.1±0.0%	0.3±0.1%
$\hat{FPR}_F$	17.6±0.1%	4.6±0.2%	10.0±0.3%
$\hat{FPR}_M$	17.2±0.2%	4.5±0.2%	9.7±0.2%
$\Delta\hat{FNR}$	0.6±0.0%	0.6±0.2%	0.5±0.3%
$\hat{FNR}_F$	17.4±0.2%	14.2±0.3%	16.7±0.1%
$\hat{FNR}_M$	18.0±0.2%	14.8±0.1%	17.2±0.4%
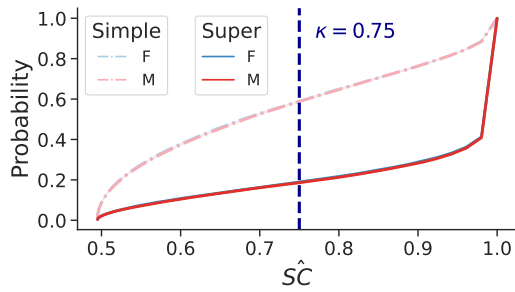
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.1±0.0%	0.3±0.1%
$\hat{AR}_F$	60.7±0.4%	18.6±0.4%
$\hat{AR}_M$	60.8±0.4%	18.3±0.3%

Figure 33: **Decision trees** on `New Adult - CA - Public Coverage`, by `sex`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	2.5±0.1%	5.3±0.0%	2.6±0.1%
$\hat{PR}_{\mathbf{F}}$	31.9±0.3%	19.5±0.3%	25.7±0.4%
$\hat{PR}_{\mathbf{M}}$	34.4±0.4%	24.8±0.3%	28.3±0.3%
$\Delta\hat{Err}$	0.4±0.1%	1.0±0.2%	0.3±0.1%
$\hat{Err}_{\mathbf{F}}$	32.3±0.2%	19.3±0.3%	26.3±0.2%
$\hat{Err}_{\mathbf{M}}$	32.7±0.1%	20.3±0.1%	26.6±0.3%
$\Delta\hat{FPR}$	0.2±0.1%	0.2±0.1%	0.3±0.1%
$\hat{FPR}_{\mathbf{F}}$	14.4±0.1%	4.1±0.2%	8.7±0.3%
$\hat{FPR}_{\mathbf{M}}$	14.2±0.2%	4.3±0.1%	8.4±0.2%
$\Delta\hat{FNR}$	0.7±0.1%	0.8±0.0%	0.7±0.1%
$\hat{FNR}_{\mathbf{F}}$	17.9±0.2%	15.2±0.2%	17.5±0.2%
$\hat{FNR}_{\mathbf{M}}$	18.6±0.3%	16.0±0.2%	18.2±0.3%
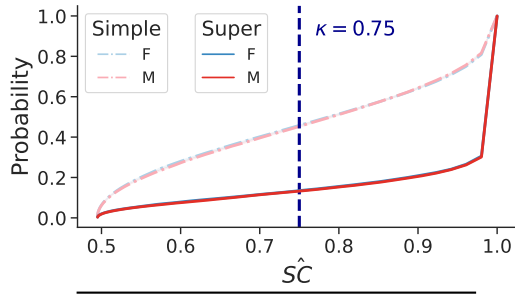
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.2±0.0%	0.2±0.2%
$\hat{AR}_{\mathbf{F}}$	48.1±0.3%	13.2±0.1%
$\hat{AR}_{\mathbf{M}}$	47.9±0.3%	13.0±0.3%

Figure 34: **Random forests** on `New Adult - CA - Public Coverage`, by `sex`

`Public Coverage` **- by** `race`.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	0.1±0.1%	0.1±0.1%	0.1±0.1%
$\hat{PR}_{\mathbf{NW}}$	16.3±0.3%	15.8±0.3%	16.2±0.3%
$\hat{PR}_{\mathbf{W}}$	16.2±0.2%	15.9±0.2%	16.3±0.2%
$\Delta\hat{Err}$	3.2±0.0%	2.7±0.1%	2.8±0.1%
$\hat{Err}_{\mathbf{NW}}$	33.4±0.3%	32.6±0.3%	32.8±0.3%
$\hat{Err}_{\mathbf{W}}$	30.2±0.3%	29.9±0.2%	30.0±0.2%
$\Delta\hat{FPR}$	0.2±0.0%	0.1±0.0%	0.1±0.1%
$\hat{FPR}_{\mathbf{NW}}$	5.6±0.1%	5.2±0.1%	5.4±0.2%
$\hat{FPR}_{\mathbf{W}}$	5.4±0.1%	5.1±0.1%	5.3±0.1%
$\Delta\hat{FNR}$	3.0±0.1%	2.6±0.1%	2.7±0.1%
$\hat{FNR}_{\mathbf{NW}}$	27.8±0.3%	27.4±0.3%	27.4±0.3%
$\hat{FNR}_{\mathbf{W}}$	24.8±0.2%	24.8±0.2%	24.7±0.2%
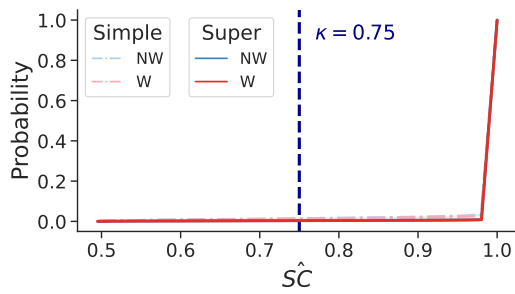
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.1±0.0%	0.1±0.0%
$\hat{AR}_{\mathbf{NW}}$	1.5±0.1%	0.4±0.0%
$\hat{AR}_{\mathbf{W}}$	1.4±0.1%	0.3±0.0%

Figure 35: **Logistic regression** on `New Adult - CA - Public Coverage`, by `race`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	1.5±0.0%	2.2±0.2%	1.4±0.1%
$\hat{\mathrm{PR}}_{\mathbf{NW}}$	37.5±0.2%	24.8±0.3%	29.6±0.4%
$\hat{\mathrm{PR}}_{\mathbf{W}}$	36.0±0.2%	22.6±0.5%	28.2±0.3%
$\Delta\hat{\mathrm{Err}}$	2.2±0.0%	3.0±0.1%	2.7±0.1%
$\hat{\mathrm{Err}}_{\mathbf{NW}}$	36.4±0.1%	20.8±0.3%	28.3±0.2%
$\hat{\mathrm{Err}}_{\mathbf{W}}$	34.2±0.1%	17.8±0.4%	25.6±0.3%
$\Delta\hat{\mathrm{FPR}}$	0.4±0.1%	0.7±0.0%	0.6±0.0%
$\hat{\mathrm{FPR}}_{\mathbf{NW}}$	17.7±0.1%	5.0±0.2%	10.2±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{W}}$	17.3±0.2%	4.3±0.2%	9.6±0.2%
$\Delta\hat{\mathrm{FNR}}$	1.8±0.0%	2.3±0.1%	2.1±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{NW}}$	18.7±0.2%	15.8±0.2%	18.1±0.3%
$\hat{\mathrm{FNR}}_{\mathbf{W}}$	16.9±0.2%	13.5±0.3%	16.0±0.2%
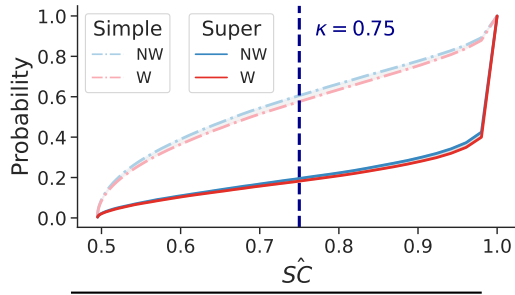
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	2.8±0.0%	1.3±0.1%
$\hat{\mathrm{AR}}_{\mathbf{NW}}$	62.3±0.3%	19.2±0.3%
$\hat{\mathrm{AR}}_{\mathbf{W}}$	59.5±0.3%	17.9±0.4%

Figure 36: **Decision trees** on `New Adult - CA - Public Coverage`, by `race`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	1.2±0.0%	1.0±0.1%	0.7±0.0%
$\hat{\mathrm{PR}}_{\mathbf{NW}}$	33.7±0.3%	22.4±0.4%	27.3±0.4%
$\hat{\mathrm{PR}}_{\mathbf{W}}$	32.5±0.3%	21.4±0.3%	26.6±0.4%
$\Delta\hat{\mathrm{Err}}$	2.7±0.1%	2.9±0.0%	2.6±0.1%
$\hat{\mathrm{Err}}_{\mathbf{NW}}$	34.0±0.2%	21.4±0.3%	27.9±0.3%
$\hat{\mathrm{Err}}_{\mathbf{W}}$	31.3±0.1%	18.5±0.3%	25.3±0.2%
$\Delta\hat{\mathrm{FPR}}$	0.5±0.0%	0.4±0.0%	0.5±0.0%
$\hat{\mathrm{FPR}}_{\mathbf{NW}}$	14.6±0.2%	4.4±0.2%	8.9±0.2%
$\hat{\mathrm{FPR}}_{\mathbf{W}}$	14.1±0.2%	4.0±0.2%	8.4±0.2%
$\Delta\hat{\mathrm{FNR}}$	2.2±0.0%	2.5±0.0%	2.3±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{NW}}$	19.4±0.2%	17.0±0.2%	19.1±0.3%
$\hat{\mathrm{FNR}}_{\mathbf{W}}$	17.2±0.2%	14.5±0.2%	16.8±0.2%
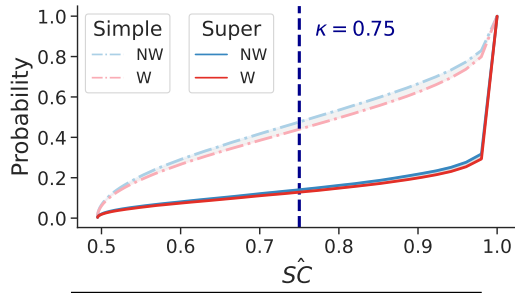
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	3.6±0.1%	1.2±0.0%
$\hat{\mathrm{AR}}_{\mathbf{NW}}$	50.0±0.2%	13.8±0.2%
$\hat{\mathrm{AR}}_{\mathbf{W}}$	46.4±0.3%	12.6±0.2%

Figure 37: **Random forests** on `New Adult - CA - Public Coverage`, by `race`

E.4.6 `HMDA`

$\hat{SC}$ CDFs for two states (`NY`, `TX`) in `HMDA` -2017, using $g=$ `ethnicity`, `race`, and `sex`, and associated error metrics on the prediction set. **Baseline** metrics computed with $B=101$ models. For **simple**, $B=101$ models; for **super**, $B=101$ ensemble models, each composed of 21 underlying models for `NY`; 15 for `TX`. We repeat for 5 test/train splits. We also report abstention rate $\hat{AR}$.

`NY` - 2017 - **by** `ethnicity`.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	9.7±0.2%	10.5±0.5%	10.4±0.5%
$\hat{PR}_{HL}$	73.5±0.3%	73.5±0.6%	73.1±0.6%
$\hat{PR}_{NHL}$	83.2±0.1%	84.0±0.1%	83.5±0.1%
$\Delta\hat{Err}$	1.3±0.2%	1.7±0.3%	1.8±0.4%
$\hat{Err}_{HL}$	18.7±0.3%	18.4±0.4%	18.9±0.5%
$\hat{Err}_{NHL}$	17.4±0.1%	16.7±0.1%	17.1±0.1%
$\Delta\hat{FPR}$	0.8±0.1%	1.0±0.1%	0.9±0.1%
$\hat{FPR}_{HL}$	10.7±0.2%	10.2±0.2%	10.5±0.2%
$\hat{FPR}_{NHL}$	11.5±0.1%	11.2±0.1%	11.4±0.1%
$\Delta\hat{FNR}$	2.2±0.2%	2.7±0.3%	2.7±0.4%
$\hat{FNR}_{HL}$	8.0±0.3%	8.2±0.4%	8.4±0.5%
$\hat{FNR}_{NHL}$	5.8±0.1%	5.5±0.1%	5.7±0.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.0±0.1%	0.1±0.1%
$\hat{AR}_{HL}$	1.9±0.2%	0.4±0.1%
$\hat{AR}_{NHL}$	1.9±0.1%	0.5±0.0%

Figure 38: **Logistic regression** on `HMDA` - 2017 - `NY`, by `ethnicity`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	4.0±0.2%	1.5±0.2%	8.2±0.5%
$\hat{PR}_{HL}$	73.1±0.3%	94.0±0.3%	74.1±0.6%
$\hat{PR}_{NHL}$	77.1±0.1%	95.5±0.1%	82.3±0.1%
$\Delta\hat{Err}$	0.4±0.1%	0.7±0.1%	0.6±0.2%
$\hat{Err}_{HL}$	20.6±0.2%	2.7±0.2%	12.2±0.4%
$\hat{Err}_{NHL}$	20.2±0.1%	3.4±0.1%	11.6±0.2%
$\Delta\hat{FPR}$	1.6±0.1%	1.1±0.0%	1.1±0.1%
$\hat{FPR}_{HL}$	11.5±0.2%	1.4±0.1%	5.1±0.2%
$\hat{FPR}_{NHL}$	9.9±0.1%	2.5±0.1%	6.2±0.1%
$\Delta\hat{FNR}$	1.2±0.1%	0.4±0.1%	1.6±0.3%
$\hat{FNR}_{HL}$	9.1±0.2%	1.3±0.2%	7.0±0.4%
$\hat{FNR}_{NHL}$	10.3±0.1%	0.9±0.1%	5.4±0.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	1.2±0.2%	2.3±0.1%
$\hat{AR}_{HL}$	43.3±0.4%	15.0±0.3%
$\hat{AR}_{NHL}$	42.1±0.2%	12.7±0.2%

Figure 39: **Decision trees** on `HMDA` - 2017 - `NY`, by `ethnicity`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta \hat{\text{PR}}$	6.3±0.3%	5.1±0.6%	9.4±0.4%
$\hat{\text{PR}}_{\textbf{HL}}$	71.9±0.4%	85.8±0.7%	71.6±0.5%
$\hat{\text{PR}}_{\textbf{NHL}}$	78.2±0.1%	90.9±0.1%	81.0±0.1%
$\Delta \hat{\text{Err}}$	1.1±0.1%	0.5±0.3%	1.0±0.2%
$\hat{\text{Err}}_{\textbf{HL}}$	19.1±0.2%	5.8±0.4%	13.7±0.4%
$\hat{\text{Err}}_{\textbf{NHL}}$	18.0±0.1%	6.3±0.1%	12.7±0.2%
$\Delta \hat{\text{FPR}}$	0.7±0.1%	1.5±0.1%	1.2±0.1%
$\hat{\text{FPR}}_{\textbf{HL}}$	10.1±0.2%	2.6±0.1%	5.7±0.2%
$\hat{\text{FPR}}_{\textbf{NHL}}$	9.4±0.1%	4.1±0.0%	6.9±0.1%
$\Delta \hat{\text{FNR}}$	0.4±0.2%	1.1±0.2%	2.2±0.2%
$\hat{\text{FNR}}_{\textbf{HL}}$	9.0±0.3%	3.3±0.3%	8.0±0.3%
$\hat{\text{FNR}}_{\textbf{NHL}}$	8.6±0.1%	2.2±0.1%	5.8±0.1%
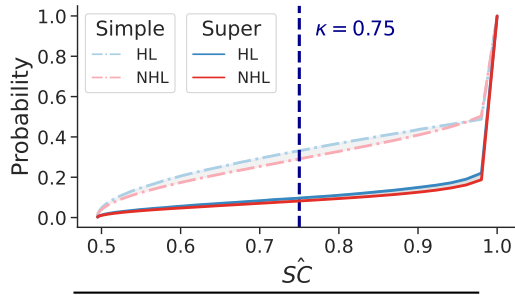
Abstention set metrics		
	Simple	**Super**
$\Delta \hat{\text{AR}}$	3.6±0.2%	1.4±0.3%
$\hat{\text{AR}}_{\textbf{HL}}$	33.9±0.3%	9.5±0.4%
$\hat{\text{AR}}_{\textbf{NHL}}$	30.3±0.1%	8.1±0.1%

Figure 40: **Random forests** on HMDA - 2017 - NY, by ethnicity

NY - 2017 **- by** race**.**

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta \hat{\text{PR}}$	11.5±0.2%	11.6±0.3%	11.5±0.3%
$\hat{\text{PR}}_{\textbf{NW}}$	73.3±0.3%	73.9±0.4%	73.5±0.4%
$\hat{\text{PR}}_{\textbf{W}}$	84.8±0.1%	85.5±0.1%	85.0±0.1%
$\Delta \hat{\text{Err}}$	2.8±0.1%	2.9±0.1%	3.0±0.0%
$\hat{\text{Err}}_{\textbf{NW}}$	19.7±0.2%	19.1±0.2%	19.6±0.1%
$\hat{\text{Err}}_{\textbf{W}}$	16.9±0.1%	16.2±0.1%	16.6±0.1%
$\Delta \hat{\text{FPR}}$	0.2±0.1%	0.2±0.1%	0.1±0.1%
$\hat{\text{FPR}}_{\textbf{NW}}$	11.3±0.2%	11.0±0.2%	11.3±0.2%
$\hat{\text{FPR}}_{\textbf{W}}$	11.5±0.1%	11.2±0.1%	11.4±0.1%
$\Delta \hat{\text{FNR}}$	3.0±0.1%	3.0±0.1%	3.0±0.1%
$\hat{\text{FNR}}_{\textbf{NW}}$	8.4±0.2%	8.1±0.2%	8.3±0.2%
$\hat{\text{FNR}}_{\textbf{W}}$	5.4±0.1%	5.1±0.1%	5.3±0.1%
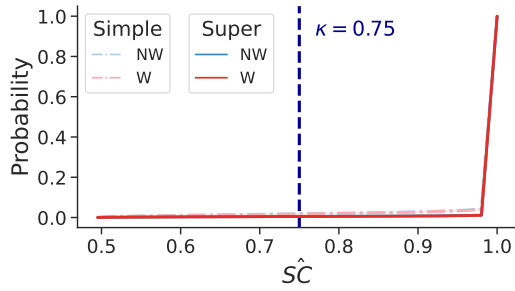
Abstention set metrics		
	Simple	**Super**
$\Delta \hat{\text{AR}}$	0.1±0.0%	0.0±0.1%
$\hat{\text{AR}}_{\textbf{NW}}$	2.0±0.1%	0.5±0.1%
$\hat{\text{AR}}_{\textbf{W}}$	1.9±0.1%	0.5±0.0%

Figure 41: **Logistic regression** on HMDA - 2017 - NY, by race

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	6.2±0.2%	3.0±0.1%	9.1±0.1%
$\hat{\text{PR}}_{\textbf{NW}}$	71.8±0.3%	93.0±0.2%	74.4±0.3%
$\hat{\text{PR}}_{\textbf{W}}$	78.0±0.1%	96.0±0.1%	83.5±0.2%
$\Delta\hat{\text{Err}}$	3.1±0.0%	0.7±0.1%	2.3±0.0%
$\hat{\text{Err}}_{\textbf{NW}}$	22.7±0.1%	3.9±0.2%	13.5±0.2%
$\hat{\text{Err}}_{\textbf{W}}$	19.6±0.1%	3.2±0.1%	11.2±0.2%
$\Delta\hat{\text{FPR}}$	2.5±0.1%	0.2±0.0%	0.6±0.1%
$\hat{\text{FPR}}_{\textbf{NW}}$	12.0±0.2%	2.5±0.1%	6.6±0.2%
$\hat{\text{FPR}}_{\textbf{W}}$	9.5±0.1%	2.3±0.1%	6.0±0.1%
$\Delta\hat{\text{FNR}}$	0.6±0.0%	0.5±0.1%	1.8±0.1%
$\hat{\text{FNR}}_{\textbf{NW}}$	10.7±0.1%	1.4±0.1%	6.9±0.2%
$\hat{\text{FNR}}_{\textbf{W}}$	10.1±0.1%	0.9±0.0%	5.1±0.1%
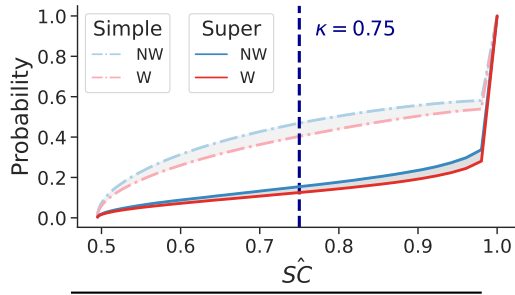
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	6.2±0.2%	2.9±0.0%
$\hat{\text{AR}}_{\textbf{NW}}$	47.1±0.4%	15.2±0.2%
$\hat{\text{AR}}_{\textbf{W}}$	40.9±0.2%	12.3±0.2%

Figure 42: **Decision trees** on HMDA - 2017 - NY, by race

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	7.9±0.2%	6.5±0.1%	9.8±0.1%
$\hat{\text{PR}}_{\textbf{NW}}$	71.5±0.3%	85.3±0.2%	72.4±0.3%
$\hat{\text{PR}}_{\textbf{W}}$	79.4±0.1%	91.8±0.1%	82.2±0.2%
$\Delta\hat{\text{Err}}$	3.3±0.0%	1.3±0.0%	2.7±0.1%
$\hat{\text{Err}}_{\textbf{NW}}$	20.7±0.1%	7.3±0.1%	14.9±0.1%
$\hat{\text{Err}}_{\textbf{W}}$	17.4±0.1%	6.0±0.1%	12.2±0.2%
$\Delta\hat{\text{FPR}}$	1.8±0.1%	0.0±0.0%	0.5±0.0%
$\hat{\text{FPR}}_{\textbf{NW}}$	10.9±0.2%	4.0±0.1%	7.2±0.1%
$\hat{\text{FPR}}_{\textbf{W}}$	9.1±0.1%	4.0±0.1%	6.7±0.1%
$\Delta\hat{\text{FNR}}$	1.4±0.0%	1.3±0.0%	2.1±0.1%
$\hat{\text{FNR}}_{\textbf{NW}}$	9.8±0.1%	3.3±0.1%	7.7±0.1%
$\hat{\text{FNR}}_{\textbf{W}}$	8.4±0.1%	2.0±0.1%	5.6±0.2%
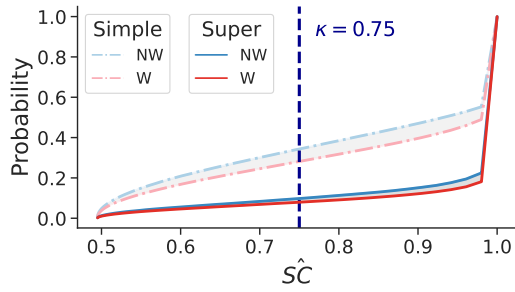
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	6.3±0.3%	1.8±0.2%
$\hat{\text{AR}}_{\textbf{NW}}$	35.6±0.4%	9.6±0.3%
$\hat{\text{AR}}_{\textbf{W}}$	29.3±0.1%	7.8±0.1%

Figure 43: **Random forests** on HMDA - 2017 - NY, by race

NY - 2017 - **by** sex.

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	1.5±0.0%	1.5±0.2%	1.5±0.2%
$\hat{\text{PR}}_{\mathbf{F}}$	81.5±0.1%	82.2±0.3%	81.7±0.3%
$\hat{\text{PR}}_{\mathbf{M}}$	83.0±0.1%	83.7±0.1%	83.2±0.1%
$\Delta\hat{\text{Err}}$	0.5±0.1%	0.6±0.1%	0.6±0.1%
$\hat{\text{Err}}_{\mathbf{F}}$	17.8±0.2%	17.2±0.2%	17.6±0.2%
$\hat{\text{Err}}_{\mathbf{M}}$	17.3±0.1%	16.6±0.1%	17.0±0.1%
$\Delta\hat{\text{FPR}}$	0.2±0.1%	0.3±0.1%	0.2±0.1%
$\hat{\text{FPR}}_{\mathbf{F}}$	11.6±0.2%	11.3±0.2%	11.5±0.2%
$\hat{\text{FPR}}_{\mathbf{M}}$	11.4±0.1%	11.0±0.1%	11.3±0.1%
$\Delta\hat{\text{FNR}}$	0.3±0.0%	0.3±0.0%	0.3±0.1%
$\hat{\text{FNR}}_{\mathbf{F}}$	6.2±0.1%	5.9±0.1%	6.1±0.2%
$\hat{\text{FNR}}_{\mathbf{M}}$	5.9±0.1%	5.6±0.1%	5.8±0.1%
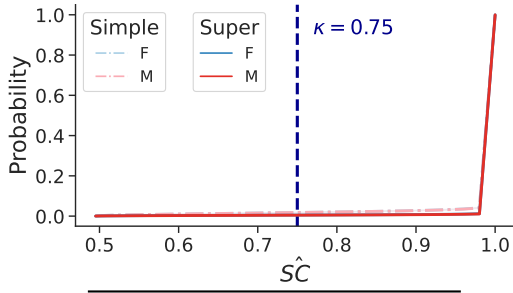
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	0.1±0.0%	0.0±0.0%
$\hat{\text{AR}}_{\mathbf{F}}$	2.0±0.1%	0.5±0.0%
$\hat{\text{AR}}_{\mathbf{M}}$	1.9±0.1%	0.5±0.0%

Figure 44: **Logistic regression** on HMDA - 2017 - NY, by sex

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	1.5±0.0%	0.9±0.0%	1.9±0.1%
$\hat{\text{PR}}_{\mathbf{F}}$	75.8±0.1%	94.8±0.1%	80.5±0.2%
$\hat{\text{PR}}_{\mathbf{M}}$	77.3±0.1%	95.7±0.1%	82.4±0.1%
$\Delta\hat{\text{Err}}$	0.4±0.1%	0.1±0.0%	0.3±0.0%
$\hat{\text{Err}}_{\mathbf{F}}$	20.5±0.2%	3.3±0.1%	11.8±0.2%
$\hat{\text{Err}}_{\mathbf{M}}$	20.1±0.1%	3.4±0.1%	11.5±0.2%
$\Delta\hat{\text{FPR}}$	0.2±0.1%	0.2±0.0%	0.1±0.0%
$\hat{\text{FPR}}_{\mathbf{F}}$	10.1±0.2%	2.2±0.1%	6.1±0.1%
$\hat{\text{FPR}}_{\mathbf{M}}$	9.9±0.1%	2.4±0.1%	6.2±0.1%
$\Delta\hat{\text{FNR}}$	0.3±0.0%	0.2±0.0%	0.4±0.1%
$\hat{\text{FNR}}_{\mathbf{F}}$	10.4±0.1%	1.1±0.1%	5.7±0.2%
$\hat{\text{FNR}}_{\mathbf{M}}$	10.1±0.1%	0.9±0.1%	5.3±0.1%
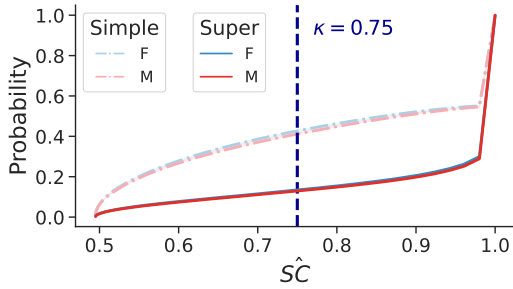
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	1.4±0.0%	0.5±0.0%
$\hat{\text{AR}}_{\mathbf{F}}$	43.1±0.2%	13.2±0.2%
$\hat{\text{AR}}_{\mathbf{M}}$	41.7±0.2%	12.7±0.2%

Figure 45: **Decision trees** on HMDA - 2017 - NY, by sex

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	1.8±0.0%	1.7±0.1%	2.0±0.1%
$\hat{\text{PR}}_{\mathbf{F}}$	76.6±0.1%	89.4±0.2%	79.0±0.2%
$\hat{\text{PR}}_{\mathbf{M}}$	78.4±0.1%	91.1±0.1%	81.0±0.1%
$\Delta\hat{\text{Err}}$	0.4±0.1%	0.1±0.1%	0.4±0.0%
$\hat{\text{Err}}_{\mathbf{F}}$	18.4±0.2%	6.3±0.2%	13.0±0.2%
$\hat{\text{Err}}_{\mathbf{M}}$	18.0±0.1%	6.2±0.1%	12.6±0.2%
$\Delta\hat{\text{FPR}}$	0.0±0.1%	0.3±0.0%	0.1±0.1%
$\hat{\text{FPR}}_{\mathbf{F}}$	9.4±0.2%	3.8±0.1%	6.7±0.2%
$\hat{\text{FPR}}_{\mathbf{M}}$	9.4±0.1%	4.1±0.1%	6.8±0.1%
$\Delta\hat{\text{FNR}}$	0.4±0.0%	0.4±0.1%	0.5±0.1%
$\hat{\text{FNR}}_{\mathbf{F}}$	8.9±0.1%	2.5±0.1%	6.3±0.2%
$\hat{\text{FNR}}_{\mathbf{M}}$	8.5±0.1%	2.1±0.0%	5.8±0.1%
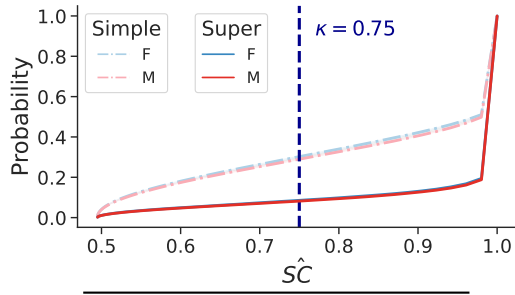
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	1.2±0.0%	0.3±0.0%
$\hat{\text{AR}}_{\mathbf{F}}$	31.4±0.1%	8.4±0.1%
$\hat{\text{AR}}_{\mathbf{M}}$	30.2±0.1%	8.1±0.1%

Figure 46: **Random forests** on `HMDA - 2017 - NY`, by `sex`

`TX - 2017` **- by** `ethnicity`**.**

Logistic regression prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	13.3±0.1%	13.1±0.0%	13.0±0.0%
$\hat{\text{PR}}_{\mathbf{HL}}$	64.7±0.2%	65.1±0.1%	65.0±0.1%
$\hat{\text{PR}}_{\mathbf{NHL}}$	78.0±0.1%	78.2±0.1%	78.0±0.1%
$\Delta\hat{\text{Err}}$	3.4±0.1%	3.3±0.1%	3.3±0.2%
$\hat{\text{Err}}_{\mathbf{HL}}$	17.5±0.2%	17.1±0.2%	17.3±0.2%
$\hat{\text{Err}}_{\mathbf{NHL}}$	14.1±0.1%	13.8±0.1%	14.0±0.0%
$\Delta\hat{\text{FPR}}$	0.6±0.0%	0.6±0.0%	0.6±0.0%
$\hat{\text{FPR}}_{\mathbf{HL}}$	5.6±0.1%	5.4±0.0%	5.6±0.0%
$\hat{\text{FPR}}_{\mathbf{NHL}}$	6.2±0.1%	6.0±0.0%	6.2±0.0%
$\Delta\hat{\text{FNR}}$	4.0±0.1%	3.9±0.1%	3.9±0.2%
$\hat{\text{FNR}}_{\mathbf{HL}}$	11.9±0.2%	11.6±0.2%	11.7±0.2%
$\hat{\text{FNR}}_{\mathbf{NHL}}$	7.9±0.1%	7.7±0.1%	7.8±0.0%
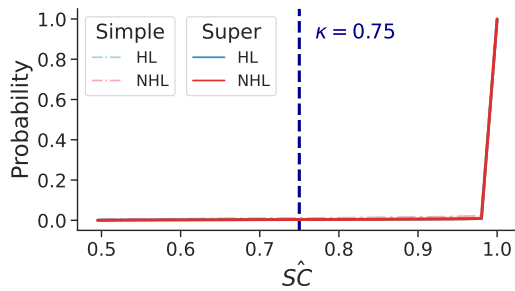
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\text{AR}}$	0.1±0.0%	0.0±0.0%
$\hat{\text{AR}}_{\mathbf{HL}}$	1.1±0.0%	0.4±0.0%
$\hat{\text{AR}}_{\mathbf{NHL}}$	1.0±0.0%	0.4±0.0%

Figure 47: **Logistic regression** on `HMDA - 2017 - TX`, by `ethnicity`

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	7.4±0.2%	4.1±0.1%	8.5±0.0%
$\hat{\mathrm{PR}}_{\mathbf{HL}}$	72.0±0.2%	90.2±0.1%	76.7±0.1%
$\hat{\mathrm{PR}}_{\mathbf{NHL}}$	79.4±0.0%	94.3±0.0%	85.2±0.1%
$\Delta\hat{\mathrm{Err}}$	3.7±0.0%	0.7±0.1%	2.0±0.1%
$\hat{\mathrm{Err}}_{\mathbf{HL}}$	19.2±0.1%	3.0±0.1%	8.4±0.1%
$\hat{\mathrm{Err}}_{\mathbf{NHL}}$	15.5±0.1%	2.3±0.0%	6.4±0.0%
$\Delta\hat{\mathrm{FPR}}$	2.4±0.1%	0.1±0.0%	0.4±0.0%
$\hat{\mathrm{FPR}}_{\mathbf{HL}}$	10.1±0.1%	1.2±0.0%	3.3±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{NHL}}$	7.7±0.0%	1.1±0.0%	2.9±0.1%
$\Delta\hat{\mathrm{FNR}}$	1.2±0.1%	0.6±0.0%	1.5±0.0%
$\hat{\mathrm{FNR}}_{\mathbf{HL}}$	9.1±0.1%	1.8±0.0%	5.0±0.0%
$\hat{\mathrm{FNR}}_{\mathbf{NHL}}$	7.9±0.0%	1.2±0.0%	3.5±0.0%
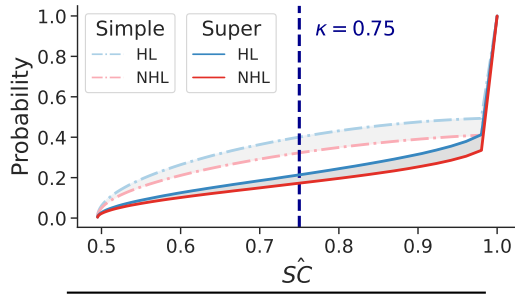
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	7.3±0.0%	4.1±0.1%
$\hat{\mathrm{AR}}_{\mathbf{HL}}$	40.3±0.1%	21.1±0.0%
$\hat{\mathrm{AR}}_{\mathbf{NHL}}$	33.0±0.1%	17.0±0.1%

Figure 48: **Decision trees** on HMDA - 2017 - TX, by ethnicity

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{\mathrm{PR}}$	8.7±0.1%	6.9±0.1%	9.6±0.1%
$\hat{\mathrm{PR}}_{\mathbf{HL}}$	70.6±0.2%	83.0±0.2%	72.7±0.2%
$\hat{\mathrm{PR}}_{\mathbf{NHL}}$	79.3±0.1%	89.9±0.1%	82.3±0.1%
$\Delta\hat{\mathrm{Err}}$	3.3±0.0%	1.2±0.1%	2.2±0.1%
$\hat{\mathrm{Err}}_{\mathbf{HL}}$	17.3±0.1%	5.3±0.1%	10.5±0.1%
$\hat{\mathrm{Err}}_{\mathbf{NHL}}$	14.0±0.1%	4.1±0.0%	8.3±0.0%
$\Delta\hat{\mathrm{FPR}}$	1.7±0.1%	0.1±0.0%	0.4±0.1%
$\hat{\mathrm{FPR}}_{\mathbf{HL}}$	8.5±0.1%	2.0±0.0%	4.1±0.0%
$\hat{\mathrm{FPR}}_{\mathbf{NHL}}$	6.8±0.0%	1.9±0.0%	3.7±0.1%
$\Delta\hat{\mathrm{FNR}}$	1.7±0.1%	1.1±0.1%	1.8±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{HL}}$	8.9±0.1%	3.3±0.1%	6.4±0.1%
$\hat{\mathrm{FNR}}_{\mathbf{NHL}}$	7.2±0.0%	2.2±0.0%	4.6±0.0%
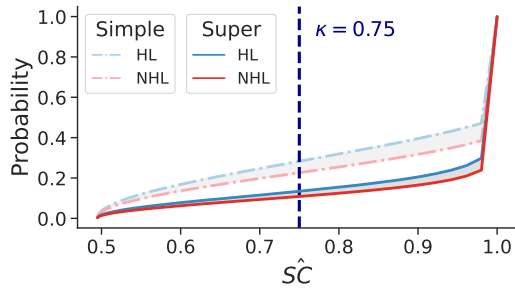
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{\mathrm{AR}}$	6.2±0.1%	2.5±0.1%
$\hat{\mathrm{AR}}_{\mathbf{HL}}$	31.9±0.0%	13.2±0.1%
$\hat{\mathrm{AR}}_{\mathbf{NHL}}$	25.7±0.1%	10.7±0.0%

Figure 49: **Random forests** on HMDA - 2017 - TX, by ethnicity

TX - 2017 - **by** `race`.

	Logistic regression prediction set metrics		
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	2.6±0.1%	2.7±0.1%	2.7±0.1%
$\hat{PR}_{NW}$	72.6±0.2%	72.8±0.0%	72.6±0.0%
$\hat{PR}_{W}$	75.2±0.1%	75.5±0.1%	75.3±0.1%
$\Delta\hat{Err}$	0.0±0.1%	0.2±0.0%	0.1±0.0%
$\hat{Err}_{NW}$	14.9±0.2%	14.4±0.1%	14.7±0.1%
$\hat{Err}_{W}$	14.9±0.1%	14.6±0.1%	14.8±0.1%
$\Delta\hat{FPR}$	1.1±0.1%	0.9±0.1%	1.0±0.1%
$\hat{FPR}_{NW}$	7.0±0.2%	6.6±0.1%	6.8±0.1%
$\hat{FPR}_{W}$	5.9±0.1%	5.7±0.0%	5.8±0.0%
$\Delta\hat{FNR}$	1.2±0.0%	1.1±0.1%	1.1±0.0%
$\hat{FNR}_{NW}$	7.9±0.1%	7.8±0.0%	7.9±0.1%
$\hat{FNR}_{W}$	9.1±0.1%	8.9±0.1%	9.0±0.1%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.0±0.0%	0.0±0.0%
$\hat{AR}_{NW}$	1.0±0.0%	0.4±0.0%
$\hat{AR}_{W}$	1.0±0.0%	0.4±0.0%

Figure 50: **Logistic regression** on `HMDA - 2017 - TX`, by `race`

	Decision tree prediction set metrics		
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	2.5±0.0%	1.8±0.0%	4.1±0.2%
$\hat{PR}_{NW}$	75.6±0.1%	91.9±0.1%	79.8±0.3%
$\hat{PR}_{W}$	78.1±0.1%	93.7±0.1%	83.9±0.1%
$\Delta\hat{Err}$	0.0±0.1%	0.0±0.0%	0.1±0.0%
$\hat{Err}_{NW}$	16.4±0.1%	2.4±0.0%	6.8±0.1%
$\hat{Err}_{W}$	16.4±0.0%	2.4±0.0%	6.9±0.1%
$\Delta\hat{FPR}$	1.2±0.0%	0.0±0.0%	0.1±0.0%
$\hat{FPR}_{NW}$	9.2±0.1%	1.1±0.0%	2.9±0.1%
$\hat{FPR}_{W}$	8.0±0.1%	1.1±0.0%	3.0±0.1%
$\Delta\hat{FNR}$	1.2±0.0%	0.1±0.0%	0.1±0.1%
$\hat{FNR}_{NW}$	7.2±0.1%	1.4±0.0%	3.9±0.1%
$\hat{FNR}_{W}$	8.4±0.1%	1.3±0.0%	3.8±0.0%

Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	0.1±0.1%	0.0±0.3%
$\hat{AR}_{NW}$	34.7±0.2%	18.0±0.3%
$\hat{AR}_{W}$	34.8±0.1%	18.0±0.0%

Figure 51: **Decision trees** on `HMDA - 2017 - TX`, by `race`
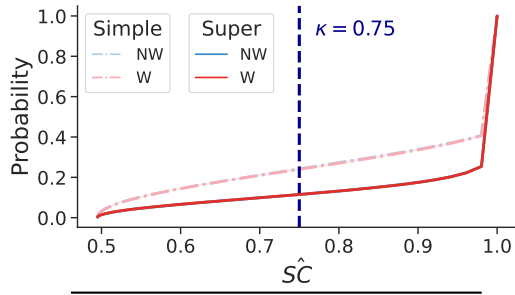
Abstention set metrics

	Simple	**Super**
$\Delta\hat{\text{AR}}$	$0.1\pm0.1\%$	$0.0\pm0.0\%$
$\hat{\text{AR}}_{\mathbf{NW}}$	$27.3\pm0.1\%$	$11.3\pm0.1\%$
$\hat{\text{AR}}_{\mathbf{W}}$	$27.2\pm0.0\%$	$11.3\pm0.1\%$

Random forest prediction set metrics

	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	$3.4\pm0.1\%$	$3.3\pm0.0\%$	$4.5\pm0.0\%$
$\hat{\text{PR}}_{\mathbf{NW}}$	$74.4\pm0.2\%$	$85.6\pm0.1\%$	$76.3\pm0.1\%$
$\hat{\text{PR}}_{\mathbf{W}}$	$77.8\pm0.1\%$	$88.9\pm0.1\%$	$80.8\pm0.1\%$
$\Delta\hat{\text{Err}}$	$0.0\pm0.0\%$	$0.1\pm0.0\%$	$0.1\pm0.0\%$
$\hat{\text{Err}}_{\mathbf{NW}}$	$14.8\pm0.1\%$	$4.3\pm0.0\%$	$8.7\pm0.1\%$
$\hat{\text{Err}}_{\mathbf{W}}$	$14.8\pm0.1\%$	$4.4\pm0.0\%$	$8.8\pm0.1\%$
$\Delta\hat{\text{FPR}}$	$0.7\pm0.0\%$	$0.1\pm0.1\%$	$0.1\pm0.1\%$
$\hat{\text{FPR}}_{\mathbf{NW}}$	$7.8\pm0.1\%$	$1.8\pm0.1\%$	$3.7\pm0.1\%$
$\hat{\text{FPR}}_{\mathbf{W}}$	$7.1\pm0.1\%$	$1.9\pm0.0\%$	$3.8\pm0.0\%$
$\Delta\hat{\text{FNR}}$	$0.8\pm0.0\%$	$0.2\pm0.1\%$	$0.1\pm0.0\%$
$\hat{\text{FNR}}_{\mathbf{NW}}$	$6.9\pm0.1\%$	$2.6\pm0.1\%$	$5.1\pm0.0\%$
$\hat{\text{FNR}}_{\mathbf{W}}$	$7.7\pm0.1\%$	$2.4\pm0.0\%$	$5.0\pm0.0\%$

Figure 52: **Random forests** on HMDA - 2017 - TX, by race

TX - 2017 - **by** sex.
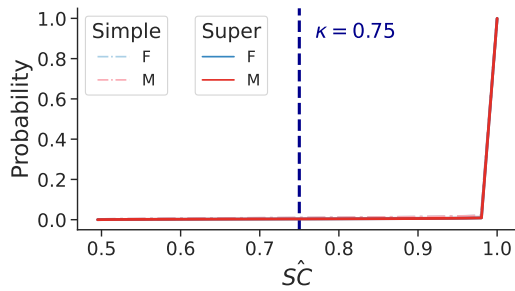
Abstention set metrics

	Simple	**Super**
$\Delta\hat{\text{AR}}$	$0.0\pm0.0\%$	$0.0\pm0.0\%$
$\hat{\text{AR}}_{\mathbf{F}}$	$1.0\pm0.0\%$	$0.4\pm0.0\%$
$\hat{\text{AR}}_{\mathbf{M}}$	$1.0\pm0.0\%$	$0.4\pm0.0\%$

Logistic regression prediction set metrics

	Baseline	**Simple**	**Super**
$\Delta\hat{\text{PR}}$	$5.7\pm0.1\%$	$5.6\pm0.0\%$	$5.6\pm0.0\%$
$\hat{\text{PR}}_{\mathbf{F}}$	$70.8\pm0.2\%$	$71.1\pm0.2\%$	$70.9\pm0.2\%$
$\hat{\text{PR}}_{\mathbf{M}}$	$76.5\pm0.1\%$	$76.7\pm0.2\%$	$76.5\pm0.2\%$
$\Delta\hat{\text{Err}}$	$1.1\pm0.2\%$	$1.0\pm0.1\%$	$1.0\pm0.1\%$
$\hat{\text{Err}}_{\mathbf{F}}$	$15.7\pm0.2\%$	$15.3\pm0.0\%$	$15.5\pm0.0\%$
$\hat{\text{Err}}_{\mathbf{M}}$	$14.6\pm0.0\%$	$14.3\pm0.1\%$	$14.5\pm0.1\%$
$\Delta\hat{\text{FPR}}$	$0.4\pm0.0\%$	$0.4\pm0.1\%$	$0.4\pm0.0\%$
$\hat{\text{FPR}}_{\mathbf{F}}$	$5.8\pm0.1\%$	$5.6\pm0.1\%$	$5.7\pm0.0\%$
$\hat{\text{FPR}}_{\mathbf{M}}$	$6.2\pm0.1\%$	$6.0\pm0.0\%$	$6.1\pm0.0\%$
$\Delta\hat{\text{FNR}}$	$1.4\pm0.1\%$	$1.3\pm0.1\%$	$1.3\pm0.1\%$
$\hat{\text{FNR}}_{\mathbf{F}}$	$9.8\pm0.2\%$	$9.6\pm0.0\%$	$9.7\pm0.0\%$
$\hat{\text{FNR}}_{\mathbf{M}}$	$8.4\pm0.1\%$	$8.3\pm0.1\%$	$8.4\pm0.1\%$

Figure 53: **Logistic regression** on HMDA - 2017 - TX, by sex

Decision tree prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	3.5±0.0%	2.2±0.1%	4.8±0.0%
$\hat{PR}_\mathbf{F}$	75.2±0.1%	91.8±0.0%	79.8±0.1%
$\hat{PR}_\mathbf{M}$	78.7±0.1%	94.0±0.1%	84.6±0.1%
$\Delta\hat{Err}$	1.1±0.1%	0.2±0.0%	0.5±0.0%
$\hat{Err}_\mathbf{F}$	17.2±0.1%	2.6±0.0%	7.2±0.1%
$\hat{Err}_\mathbf{M}$	16.1±0.0%	2.4±0.0%	6.7±0.1%
$\Delta\hat{FPR}$	0.8±0.1%	0.2±0.0%	0.3±0.1%
$\hat{FPR}_\mathbf{F}$	8.8±0.1%	1.0±0.0%	2.8±0.0%
$\hat{FPR}_\mathbf{M}$	8.0±0.0%	1.2±0.0%	3.1±0.1%
$\Delta\hat{FNR}$	0.3±0.1%	0.4±0.0%	0.9±0.1%
$\hat{FNR}_\mathbf{F}$	8.4±0.1%	1.6±0.0%	4.5±0.1%
$\hat{FNR}_\mathbf{M}$	8.1±0.0%	1.2±0.0%	3.6±0.0%
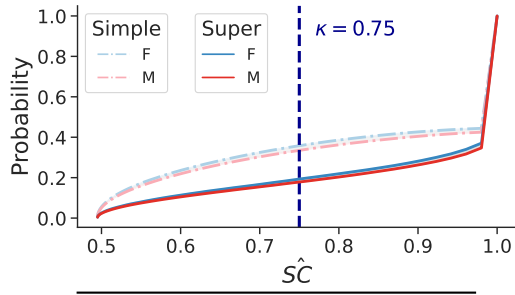
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	2.0±0.1%	1.4±0.0%
$\hat{AR}_\mathbf{F}$	36.2±0.2%	19.0±0.1%
$\hat{AR}_\mathbf{M}$	34.2±0.1%	17.6±0.1%

Figure 54: **Decision trees** on `HMDA - 2017 - TX`, by `sex`

Random forest prediction set metrics			
	Baseline	**Simple**	**Super**
$\Delta\hat{PR}$	4.4±0.0%	4.0±0.0%	5.2±0.1%
$\hat{PR}_\mathbf{F}$	74.1±0.1%	85.5±0.1%	76.3±0.2%
$\hat{PR}_\mathbf{M}$	78.5±0.1%	89.5±0.1%	81.5±0.1%
$\Delta\hat{Err}$	1.0±0.1%	0.4±0.0%	0.7±0.1%
$\hat{Err}_\mathbf{F}$	15.5±0.1%	4.6±0.0%	9.3±0.0%
$\hat{Err}_\mathbf{M}$	14.5±0.0%	4.2±0.0%	8.6±0.1%
$\Delta\hat{FPR}$	0.3±0.1%	0.3±0.0%	0.4±0.1%
$\hat{FPR}_\mathbf{F}$	7.4±0.1%	1.7±0.0%	3.5±0.1%
$\hat{FPR}_\mathbf{M}$	7.1±0.0%	2.0±0.0%	3.9±0.0%
$\Delta\hat{FNR}$	0.7±0.0%	0.8±0.0%	1.1±0.1%
$\hat{FNR}_\mathbf{F}$	8.1±0.1%	3.0±0.0%	5.8±0.1%
$\hat{FNR}_\mathbf{M}$	7.4±0.1%	2.2±0.0%	4.7±0.0%
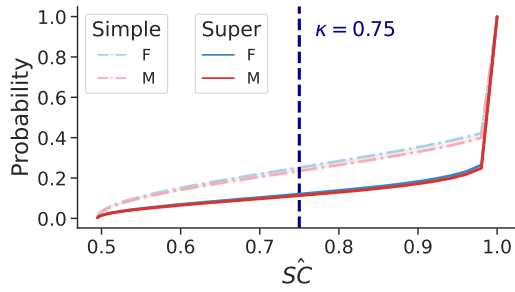
Abstention set metrics		
	Simple	**Super**
$\Delta\hat{AR}$	1.7±0.0%	0.8±0.2%
$\hat{AR}_\mathbf{F}$	28.4±0.1%	11.9±0.2%
$\hat{AR}_\mathbf{M}$	26.7±0.1%	11.1±0.0%

Figure 55: **Random forests** on `HMDA - 2017 - TX`, by `sex`

E.4.7 Discussion of extended results for Algorithm 1

Overall, our results support that examining self-consistency and error together provide a much richer picture of model behavior, both with respect to arbitrariness and fairness metric disparities. Particularly in smaller datasets, the learning process produces models with a large degree of variance. As a result, ensembling with confidence can lead to huge abstention rates.

Improving self-consistency by doing a round of variance reduction first and then ensembling with confidence (i.e., super-ensembling) can lead to improvements in error over baselines while having a lower abstention rate. These improvements are typically shared across subgroups, but may not be symmetric; some subgroups may benefit more than others. As a result, even though accuracy increases absolutely for both groups, relative fairness metrics can decrease. This is a different instantiation of the fairness-accuracy trade-off than is often written about, which posits a necessary decrease in accuracy for one subgroup to improve fairness between binarized groups. Our results suggest that it is worth first tuning for accuracy, and then seeing how fairness interventions can balance the benefits across subgroups. Of course, it is possible that doing this could lead to injecting variance back into the model outputs, thereby reducing self-consistency and inducing arbitrariness. We leave this investigation to future work.

Additionally, our results reify that choice of model matters a lot. While overall error rates across model types may be similar, the sources of that error are not necessarily the same. This is an obvious point, relating to bias and variance. However, a lot of fair classification work describes similar performance across logistic regression, decision trees, random forests, SVMs, and MLPs (e.g., Chen et al. [11]). Looking at self-consistency confirms that this is not the case, with decision trees and random forests in particular exhibiting higher variance, and thus being more amenable to variance reduction and improvements in overall error. As fair classification research transitions to larger benchmarks, it will likely be fruitful to investigate more complex model classes.

We provide run times on our cluster environment in Table 2. We did not select for CPUs with any particular features, and thus the run times are quite variable.

We also provide details on systematic arbitrariness in Tables 3 and 4, which we measure using the $\hat{\mathcal{W}}_1$. As noted in Appendix B.3, since this metric is an average, its measure necessarily changes if we compute it over a different set of levels $\hat{\mathbb{K}}$. To make distances across interventions comparable, we treat CDF values below κ as 0, so that all of the probability mass is on $\hat{\mathrm{SC}} \geq \kappa$. We therefore report two versions of these results, those for no abstention and those that account for abstention at values $< \kappa$. Δ is the difference between $\hat{\mathcal{W}}_{1\,\mathrm{Simple}} - \hat{\mathcal{W}}_{1\,\mathrm{Super}}$. Positive differences indicate cases for which the super-ensembling method decreases the Wasserstein-1 distance between subgroup CDFs; negative differences indicate increases. While in some cases there is an increase, it is worth noting that this aligns with cases for which the $\hat{\mathcal{W}}_1$ distance is very close to 0. Old Adult, highlighted below, is the only dataset that exhibts large amounts of systematic arbitrariness (for decision tress and random forests, in particular; it exhibits the highest amount for logistic regression, but is overall low). Old Adult and New Adult – Employment (by sex) are two of the only tasks that show any fairness disparities that are $> 3\%$.

Table 2: These times are recorded for our cluster environment (hh:mm:ss), described in Appendix E.2 for our Algorithm 1 experiments. At the time of running, due to time constraints, the authors had not yet parallelized this part of the code.

Dataset	g	Logistic regression	Decision trees	Random forests
South German Credit	sex	00:42:50	00:25:28	00:34:42
COMPAS	race	00:57:05	00:39:24	00:31:47
Old Adult	sex	01:08:37	01:23:39	00:57:11
Taiwan Credit	sex	00:31:35	01:34:57	01:53:33
New Adult – CA				
Income	sex,race	01:39:53	02:51:13	04:59:07
Employment	sex,race	02:20:15	02:18:16	03:00:15
Public Coverage	sex,race	01:13:33	02:02:57	02:24:08
HMDA – 2017				
NY	sex, race, ethnicity	03:50:52	05:00:19	05:39:44
TX	sex, race, ethnicity	05:18:59	04:10:34	04:18:59

Table 3: Empirical Wasserstein-1 ($\hat{\mathcal{W}}_1$) measurements without abstention

Dataset	g	Logistic regression			Decision trees			Random forests		
		Simple	Super	Δ	Simple	Super	Δ	Simple	Super	Δ
German Credit	sex	0.0181	0.0101	0.0079	0.0162	0.0244	-0.0082	0.0181	0.0175	0.0006
COMPAS	race	0.0073	0.0043	0.0030	0.0189	0.0170	0.0019	0.0073	0.0031	0.0043
Old Adult	sex	0.0206	0.0033	0.0173	0.1386	0.0273	0.1112	0.1266	0.0255	0.1011
Taiwan Credit	sex	0.0028	0.0007	0.0020	0.0223	0.0108	0.0115	0.0240	0.0065	0.0175
New Adult - CA Income	sex	0.0009	0.0003	0.0006	0.0138	0.0055	0.0083	0.0089	0.0018	0.0071
	race	0.0003	0.0001	0.0002	0.0170	0.0073	0.0096	0.0163	0.0055	0.0108
Employment	sex	0.0004	0.0002	0.0003	0.0010	0.0011	-0.0001	0.0043	0.0031	0.0013
	race	0.0004	0.0001	0.0004	0.0020	0.0007	0.0013	0.0021	0.0015	0.0006
Public Coverage	sex	0.0004	0.0001	0.0003	0.0029	0.0024	0.0005	0.0045	0.0023	0.0024
	race	0.0010	0.0003	0.0007	0.0200	0.0089	0.0113	0.0235	0.0089	0.0147
HMDA - 2017 NY	sex	0.0002	0.0004	-0.0002	0.0096	0.0039	0.0056	0.0080	0.0023	0.0056
	race	0.0009	0.0005	0.0005	0.0433	0.0203	0.0231	0.0409	0.0133	0.0276
	ethnicity	0.0005	0.0005	0.0000	0.0229	0.0156	0.0073	0.0248	0.0108	0.0140
TX	sex	0.0001	0.0001	0.0000	0.0153	0.0097	0.0055	0.0113	0.0054	0.0058
	race	0.0001	0.0001	0.0000	0.0010	0.0012	-0.0002	0.0013	0.0007	0.0006
	ethnicity	0.0007	0.0002	0.0004	0.0509	0.0291	0.0219	0.0379	0.0190	0.0188

Table 4: Empirical Wasserstein-1 ($\hat{\mathcal{W}}_1$) measurements with abstention using $\kappa \geq .75$

Dataset	g	Logistic regression			Decision trees			Random forests		
		Simple	Super	Δ	Simple	Super	Δ	Simple	Super	Δ
German Credit	sex	0.0113	0.0080	0.0034	0.0090	0.0094	-0.0004	0.0084	0.0132	-0.0048
COMPAS	race	0.0035	0.0019	0.0017	0.0039	0.0060	-0.0021	0.0041	0.0019	0.0022
Old Adult	sex	0.0110	0.0020	0.0090	0.0654	0.0155	0.0500	0.0634	0.0139	0.0494
Taiwan Credit	sex	0.0014	0.0005	0.0009	0.0057	0.0059	-0.0002	0.0107	0.0040	0.0067
New Adult - CA Income	sex	0.0005	0.0002	0.0004	0.0051	0.0032	0.0019	0.0047	0.0012	0.0035
	race	0.0002	0.0000	0.0002	0.0073	0.0040	0.0033	0.0082	0.0028	0.0053
Employment	sex	0.0002	0.0001	0.0001	0.0005	0.0005	0.0001	0.0020	0.0014	0.0006
	race	0.0002	0.0000	0.0002	0.0012	0.0003	0.0008	0.0008	0.0005	0.0003
Public Coverage	sex	0.0002	0.0001	0.0001	0.0006	0.0012	-0.0006	0.0011	0.0009	0.0002
	race	0.0006	0.0001	0.0005	0.0068	0.0049	0.0019	0.0106	0.0047	0.0059
HMDA - 2017 NY	sex	0.0001	0.0002	-0.0001	0.0033	0.0020	0.0012	0.0040	0.0013	0.0028
	race	0.0004	0.0002	0.0002	0.0155	0.0111	0.0044	0.0190	0.0076	0.0114
	ethnicity	0.0002	0.0002	0.0001	0.0055	0.0083	-0.0028	0.0081	0.0059	0.0022
TX	sex	0.0000	0.0000	0.0000	0.0061	0.0050	0.0011	0.0058	0.0029	0.0028
	race	0.0000	0.0001	0.0000	0.0004	0.0005	-0.0002	0.0007	0.0004	0.0003
	ethnicity	0.0003	0.0001	0.0003	0.0229	0.0159	0.0070	0.01200	0.0104	0.0095

E.5 Reliability and fairness metrics in `COMPAS` and `South German Credit`

Even before we apply our intervention to improve self-consistency, our results in Section 3 show close-to-parity $\hat{\text{Err}}$, $\hat{\text{FPR}}$, and $\hat{\text{FNR}}$ across subgroups in `COMPAS` (and similarly for `South German Credit`, below). These results are surprising. We run $B = 101$ models to produce estimates of variance and self-consistency, but of course doing this also has the effect of estimating the expected error more generally (with variance representing a portion of that error). Our estimates of expected error for these tasks indicate that the average model produced training on `COMPAS` and `South German Credit`, with respect to popular fairness definitions like Equality of Opportunity and Equalized Odds [4, 34] *are in fact baseline close to parity, with **no** fairness intervention applied*. We found this across model types for both datasets, though the story becomes more complicated when we apply techniques to improve self-consistency (see discussion at the end of Appendix E.4).

Of course, we did not expect this result, as these are two of the *de facto* standard benchmark datasets in algorithmic fairness. They are used in countless other studies to probe and verify algorithmic fairness interventions [26]. As a result, we initially thought that our results must be incorrect. We therefore looked at the underlying models in our bootstrap runs to see the error of the underlying models.

We re-ran our baseline experiments with $B = 1001$ and for 100 test/train splits for logistic regression. In Figure 56, we plot the $(100,100)$ bootstrap models that went into these results. For another view on analogous information, in Table 5, we provide an excerpt of the results for `COMPAS` regarding the underlying 1010 random forest classifiers used to produce Figure 2a.

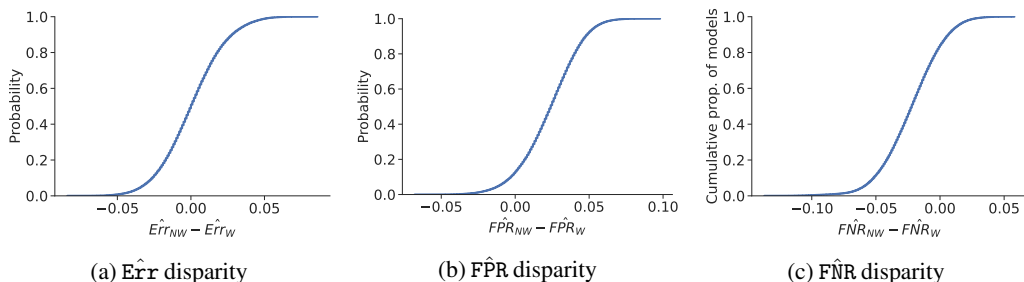| (a) $\hat{\text{Err}}$ disparity | (b) $\hat{\text{FPR}}$ disparity | (c) $\hat{\text{FNR}}$ disparity |

Figure 56: Cumulative distribution of error disparity across 100,100 logistic regression models trained on `COMPAS`.

Overall, we can see that there is a wide range of error disparities that trend in both directions, with a skew toward higher $\hat{\text{FPR}}$ for $g = \text{NW}$. These results support our claim that training many models is necessary to get an accurate picture of expected error, with implications both for reproducibility of experiments that just train and analyze a small handful of models and for generalizability. There are models that exhibit worse degrees of unfairness in both directions, but they are more unlikely than models that exhibit smaller disparities.

We subset the above results to the 100 models that produce the lowest $\hat{\text{Err}}$, as this is often the selection criteria for picking models to post-process. We plot these results below. These top-performing models in fact exhibit (on average) closer-to-parity for $\hat{\text{FPR}}$ and $\hat{\text{FNR}}$.
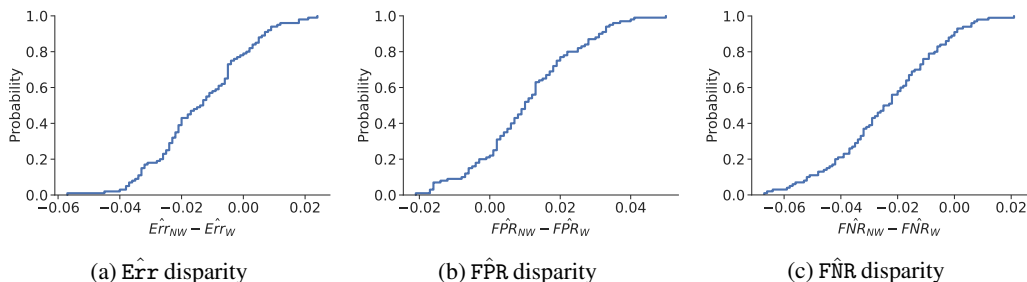

| (a) $\hat{\text{Err}}$ disparity | (b) $\hat{\text{FPR}}$ disparity | (c) $\hat{\text{FNR}}$ disparity |

Figure 57: CDF of error disparity across the top 100 logistic regression models (of the 100,100 models) trained on `COMPAS`.

Table 5: Comparing subgroup error rates in COMPAS for different random forest classifiers trained to produce Figure 2a. Each table looks at the top-3 highest differences between subgroups for the specified metric: (a) $\hat{\text{Err}}_{\text{NW}} - \hat{\text{Err}}_{\text{W}}$, when $\hat{\text{Err}}_{\text{NW}} > \hat{\text{Err}}_{\text{W}}$; (b) $\hat{\text{Err}}_{\text{W}} - \hat{\text{Err}}_{\text{NW}}$, when $\hat{\text{Err}}_{\text{W}} > \hat{\text{Err}}_{\text{NW}}$; (c) $\hat{\text{FPR}}_{\text{NW}} - \hat{\text{FPR}}_{\text{W}}$, when $\hat{\text{FPR}}_{\text{NW}} > \hat{\text{FPR}}_{\text{W}}$; (d) $\hat{\text{FPR}}_{\text{W}} - \hat{\text{FPR}}_{\text{NW}}$, when $\hat{\text{FPR}}_{\text{W}} > \hat{\text{FPR}}_{\text{NW}}$; (e) $\hat{\text{FNR}}_{\text{NW}} - \hat{\text{FNR}}_{\text{W}}$, when $\hat{\text{FNR}}_{\text{NW}} > \hat{\text{FNR}}_{\text{W}}$; and, (f) (e) $\hat{\text{FNR}}_{\text{W}} - \hat{\text{FNR}}_{\text{NW}}$, when $\hat{\text{FNR}}_{\text{W}} > \hat{\text{FNR}}_{\text{NW}}$. We highlight the overall error metric in gray, the larger metric (being subtracted from) in blue, the smaller metric (being subtracted) in red, and the difference in the metric between subgroups in purple. Note that run 757 appears twice, which we mark in orange.

(a) The top-3 most unfair models by subgroup-specific $\hat{\text{Err}}$, when $\hat{\text{Err}}_{\text{NW}} > \hat{\text{Err}}_{\text{W}}$ (i.e., unfair toward NW).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{Err}}_{\text{NW}} - \hat{\text{Err}}_{\text{W}}$
762	8	504	0.374	0.179	0.196	0.405	0.204	0.201	0.315	0.13	0.186	0.09
757	8	464	0.369	0.167	0.202	0.395	0.201	0.193	0.318	0.101	0.218	0.077
328	4	116	0.371	0.165	0.206	0.395	0.181	0.214	0.323	0.134	0.189	0.072

(b) The top-3 most unfair models by subgroup-specific $\hat{\text{Err}}$, when $\hat{\text{Err}}_{\text{W}} > \hat{\text{Err}}_{\text{NW}}$ (i.e., unfair toward W).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{Err}}_{\text{W}} - \hat{\text{Err}}_{\text{NW}}$
414	5	75	0.376	0.167	0.209	0.352	0.158	0.194	0.422	0.186	0.236	0.07
435	5	180	0.376	0.199	0.177	0.355	0.189	0.166	0.416	0.217	0.198	0.061
413	5	70	0.378	0.189	0.189	0.359	0.188	0.171	0.413	0.191	0.222	0.054

(c) The top-3 most unfair models by subgroup-specific $\hat{\text{FPR}}$, when $\hat{\text{FPR}}_{\text{NW}} > \hat{\text{FPR}}_{\text{W}}$ (i.e., unfair toward NW).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{NW}} - \hat{\text{FPR}}_{\text{W}}$
757	8	464	0.369	0.167	0.202	0.395	0.201	0.193	0.318	0.101	0.218	0.1
729	8	240	0.358	0.162	0.197	0.376	0.189	0.187	0.323	0.107	0.216	0.082
791	8	736	0.377	0.171	0.205	0.395	0.198	0.197	0.341	0.118	0.222	0.08

(d) The top-3 most unfair models by subgroup-specific $\hat{\text{FPR}}$, when $\hat{\text{FPR}}_{\text{W}} > \hat{\text{FPR}}_{\text{NW}}$ (i.e., unfair toward W).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}} - \hat{\text{FPR}}_{\text{NW}}$
639	7	280	0.36	0.187	0.173	0.352	0.174	0.178	0.376	0.212	0.164	0.038
807	9	72	0.381	0.191	0.19	0.372	0.179	0.192	0.398	0.214	0.184	0.035
543	6	264	0.358	0.155	0.203	0.351	0.144	0.206	0.37	0.175	0.196	0.031

(e) The top-3 most unfair models by subgroup-specific $\hat{\text{FNR}}$, when $\hat{\text{FNR}}_{\text{NW}} > \hat{\text{FNR}}_{\text{W}}$ (i.e., unfair toward NW).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{NW}} - \hat{\text{FNR}}_{\text{W}}$
246	3	141	0.379	0.166	0.213	0.398	0.169	0.229	0.345	0.161	0.184	0.045
506	6	42	0.367	0.17	0.197	0.386	0.175	0.211	0.332	0.161	0.171	0.04
204	3	15	0.384	0.185	0.199	0.394	0.181	0.213	0.365	0.192	0.173	0.04

(f) The top-3 most unfair models by subgroup-specific $\hat{\text{FNR}}$, when $\hat{\text{FNR}}_{\text{W}} > \hat{\text{FNR}}_{\text{NW}}$ (i.e., unfair toward W).

Run #	s	b	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{Err}}_{\text{NW}}$	$\hat{\text{FPR}}_{\text{NW}}$	$\hat{\text{FNR}}_{\text{NW}}$	$\hat{\text{Err}}_{\text{W}}$	$\hat{\text{FPR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}}$	$\hat{\text{FNR}}_{\text{W}} - \hat{\text{FNR}}_{\text{NW}}$
474	5	375	0.373	0.175	0.199	0.356	0.183	0.174	0.406	0.159	0.247	0.073
401	5	10	0.378	0.189	0.19	0.363	0.197	0.167	0.406	0.173	0.233	0.066
52	1	53	0.367	0.172	0.196	0.351	0.178	0.173	0.397	0.16	0.238	0.065

This detailed view provides insight into how such a result is possible. Broadly speaking, individual runs have roughly similar error;[22] yet, the subgroup-specific error rates that compose the overall error can nevertheless vary widely depending on the underlying training data. This observation aligns with current interest in *model multiplicity* in the algorithmic fairness community [7, 58], which imports the idea from Breiman [10]. In this case, as suggested by Table 5, there are models that demonstrate unfairness toward both subgroups with respect to each error rate metric $\hat{\text{Err}}$, $\hat{\text{FPR}}$, and $\hat{\text{FNR}}$. When we move away from attempting to find a *single* model that performs well (accurately or fairly) on COMPAS, and instead consider the information contained across different possible models, we yield the result that the average, expected behavior smooths over the variance in underlying models such that the result is close to fair: ***The average of unfair models with high variance in subgroup error rates is essentially fair.***

Stability analysis. To verify the stability of this result, we re-execute our experiments for increasing numbers of train/test splits S and replicates B. While our results for COMPAS are generally tight for small S (e.g., Figures 2a and 5), this was not the case for German Credit, for which it was difficult to estimate self-consistency consistently. As a result, for COMPAS, we did not expect markedly different results for increased S. Our results for $S = 100, B = 1001$ using logistic regression (Figure 58, Table 6) confirm this intuition.
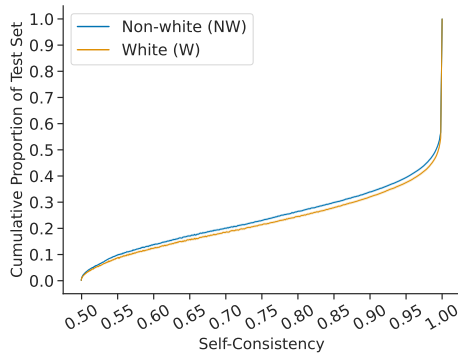

Table 6: Mean $\pm$ STD across $S = 100$ train/test splits $\times B = 1001$ runs.

	COMPAS			
	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{SC}}$
Total	.333±.008	.14±.009	.192±.01	.883±.004
$g = $NW	.333±.01	.148±.011	.185±.012	.88±.005
$g = $W	.332±.014	.125±.013	.207±.016	.888±.006

Figure 58: COMPAS split by $g = $ race, $B = 1001, S = 100$

We provide analogous results for German Credit, with $S = 1000, B = 1001$ using random forests (Figure 59, Table 7). It takes an enormous number of runs to produce stable estimates of error and $\hat{\text{SC}}$ for German Credit, which indicate statistical equality across groups. Arguably, our results below for 1,001,000 models still are very high variance (certainly with respect to error metrics). This task really has too few data points (≈ 600) to generalize reliably.
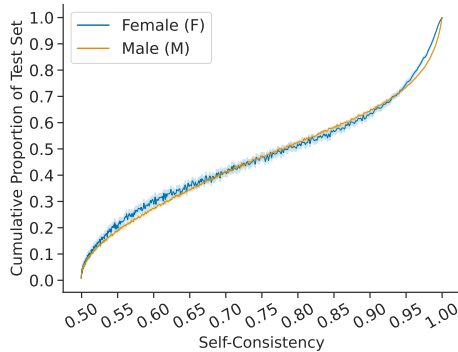

Table 7: Mean $\pm$ STD across $S = 1000$ train/test splits $\times B = 1001$ runs.

	South German Credit			
	$\hat{\text{Err}}$	$\hat{\text{FPR}}$	$\hat{\text{FNR}}$	$\hat{\text{SC}}$
Total	.28±.021	.173±.028	.107±.017	.769±.015
$g = $F	.288±.064	.183±.072	.105±.037	.766±.04
$g = $M	.279±.023	.171±.029	.108±.018	.769±.016

Figure 59: German Credit split by $g = $ sex, $S = 1000, B = 100$

---

[22]This should be taken relatively. In general, COMPAS demonstrates high error; the error is relatively tight given just how much error there is. The error fluctuates depending on the training data, but the average error rate across train/test splits is rather tight, despite the fluctuations in error within the B runs of each split.

F   Brief notes on future research

There are many interesting directions for future work that are out of scope for the present project. We address some topics below.

Novel theory.   We do not include extensive novel theory in this project. Nevertheless, our project raises interesting questions for theory in future work. Notably, we could compose our methodology with post-processing [34] for cases in which there is observed empirical unfairness. We could then investigate picking group-specific thresholds that take variance into account. We could reconfigure the formulations in [34] and related work, with respect to the fairness-accuracy trade-off, as actually representing multiple such trade-off curves (that are a function of different models under consideration). There may be interesting directions for mathematical analysis in this direction.

We could also extend traditional results on bagging and variance reduction for classifiers. While bagging has guarantees for variance reduction for regression, it does not have the same guarantees for classification [8, 9]. It generally is observed to work well in practice for variance reduction if the underlying classifiers are high variance — which is indeed the regime we are in for this paper. However, there are interesting theory questions regarding abstention that we could investigate with theoretical tools, which could let us come up with other ways of reasoning about bagging and variance reduction.

Both of these directions are out of scope for the present paper. They are interesting, but do not have to do with our main experimental aims and contributions, and thus do not make it into a conference-length submission. We are not interested in novel theory in the present study. If anything, our work highlights how over-attention to theory can (directly or indirectly) bring about serious problems of mismeasurement in practice. That is a main takeaway for our work, which by nature does not involve novel theory.

Arbitrariness beyond algorithmic fairness.   Our framework for reasoning about self-consistency and arbitrariness does not inherently have to do with algorithmic fairness. We could apply it to other domains. For example, it would be interesting to ask similar questions in deep learning and generative AI. We think that such work would be interesting, but is again out of scope for the present study. The first author of this project is in fact working on such questions as separate work. However, this project's research aims are inherently focused on fairness; the project was designed in response to observations in experimental practices in the fairness community, fairness definitions, and fairness theory.

Experiments on synthetic data.   Our results indicate that unfairness (as defined with respect to model error rates) is not frequently observed on common benchmark tasks in fair classification. Of course, there could be other datasets in fairness domains that are not currently used as benchmarks that more clearly demonstrate unfairness in practice. Hypothetically, there could be datasets for which we use Algorithm 1 to reduce arbitrariness, and yet we still see significant systematic arbitrariness or differences in error rates (and thus unfairness) due to noise or bias. We just did not really see this for almost all of the tasks we investigate in this paper, which happen to be the ones that the fairness community uses for experiments.

To study Algorithm 1 in light of these other possibilities, we could develop synthetic datasets that retain unfairness after dealing with arbitrariness. We did not do this in the present study for two reasons. First, our focus was the practice of fairness research, as it currently stands, with a data-centric approach on the datasets people actually use for their research. We are not interested in synthetic data for this project.

However, future theory results that extend our work could be vetted experimentally with synthetic data. The work we mention above regarding composition with post-processing, as well as revisting impossibility results from a distributional approach over possible models, may be very interesting to examine under data settings that we can control.

How to deal with abstention.   Future work could also perform a deeper exploration of the trade-off between abstention rate and error. We could characterize a Pareto-optimal trade-off that is a function of the choice of self-consistency level κ, and also examine in experiments and analytically how absention leads to improvements in accuracy. Future work could also identify patterns in abstention sets beyond low self-consistency. In this, looking to metrics from model multiplicity may be helpful. Further, future work could combine human decision-making or other automated elements to see how we can root out arbitrariness.

Reproducibility.   As mentioned in our Ethics Statement, we made attempts to reproduce prior work in fair classification, and often could not. We ultimately made reproducibility of specific papers out of scope for the present project, as we could make our contributions about arbitrariness and

variance without such work. It would nevertheless be useful to focus future work on reproducing prior algorithmic fairness studies, and seeing if conclusions change in those works as a function of using Algorithm 1 prior to introducing the proposed fairness intervention.

Law and policy. As mentioned in our Ethics Statement, our work regarding arbitrariness raises concrete questions for the law around due process and automated decision-making. Such contributions are also out of the scope of the present work, but we are currently developing them for future submission to a law review journal.