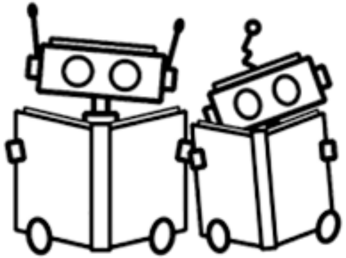
Arbitrariness and Social Prediction:

The Confounding Role of Variance in Fair Classification

A. Feder Cooper
Cornell University | The GenLaw Center



https://afedercooper.info

Arbitrariness and fairness

Existing fairness practices...

Look at **error rates across groups** (**definite**) typically, for **a single model** (**feasible**)

This can lead to **arbitrary** outcomes

(**Cooper** & Abrams, *AIES* '21 Oral; **Cooper*** et al. *ICLR* '21 Workshop Oral, **Cooper*** et al. *FAccT* '22)

Individual models → distributions over possible models

(**Cooper** et al. *CSLAW* '22)

Table 2 Definitions and classifications for popular fair machine learning classification metrics

Metric	Abbreviation	Definition
1. Predicted outcomes		
Statistical parity	SP	All groups have equal probability of being assigned to the positive class
– Treatment parity	TPar	Proportion of positive predictions of all groups must be similar
Conditional statistical parity	CSP	Requires statistics for all groups to be equal, allowing for a set of legitimate factors $L = \ell$
2. Predicted and actual outcomes		
Conditional use accuracy	CUA	Similar positive and negative predictive values across groups
Predictive parity	PP	Similar positive predictive values (or FDR) across groups
Equalized odds	EO	Similar false positive and false negative rates across groups
False positive error rate balance	FPERB	Similar false positive rates (or TNR) across groups
False negative error rate balance	FNERB	Similar false negative rates (or TPR) across groups
Treatment equality	TE	Equal ratio of false negatives and false positive between groups
Overall accuracy equality	OAE	Requires similar accuracy across groups
3. Predicted probabilities and actual outcomes		
Test fairness	TF	All groups have equal probability to belong to the positive class.
Well calibration	WC	The probability of all groups to belong to the positive class is the predicted probability score $p \in \mathcal{P}$.
Balance for positive class	BPC	Equal mean predicted probabilities for all people in the positive class, regardless of group
Balance for negative class	BNC	Equal mean predicted probabilities for all subjects in the negative class, regardless of group

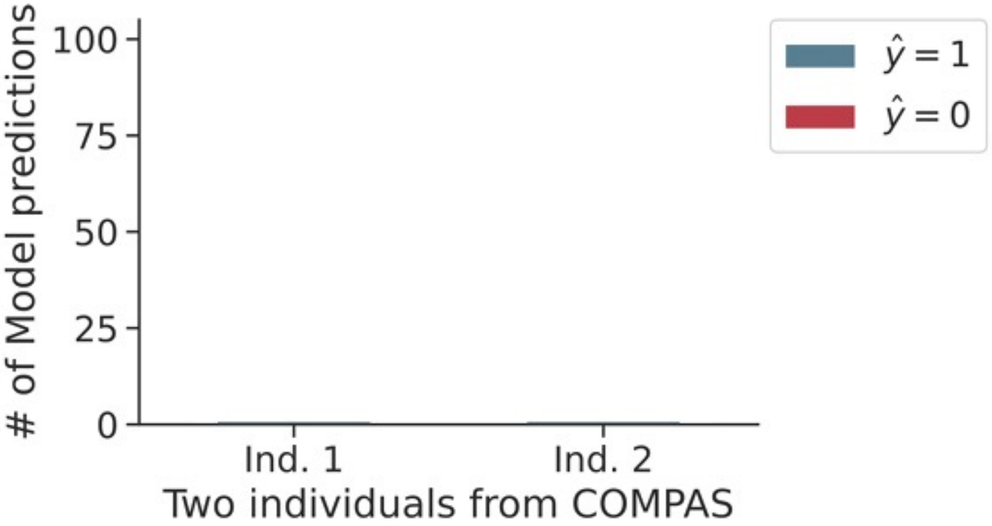
An intuition for arbitrariness

Training 100 different logistic regression models on COMPAS using **bootstrapping**

(split into train/test sets)

(resample train set)

Arbitrariness and fairness



(= recidivates)

(= does not recidivate)

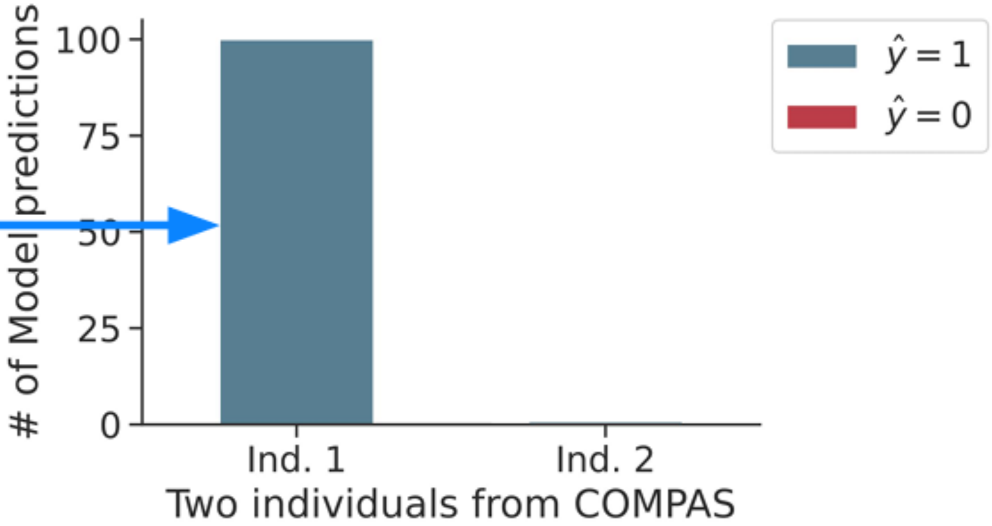
Training 100 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set

Arbitrariness and fairness

Learning process is **really confident**

All 100 models predict the same outcome for Ind. 1



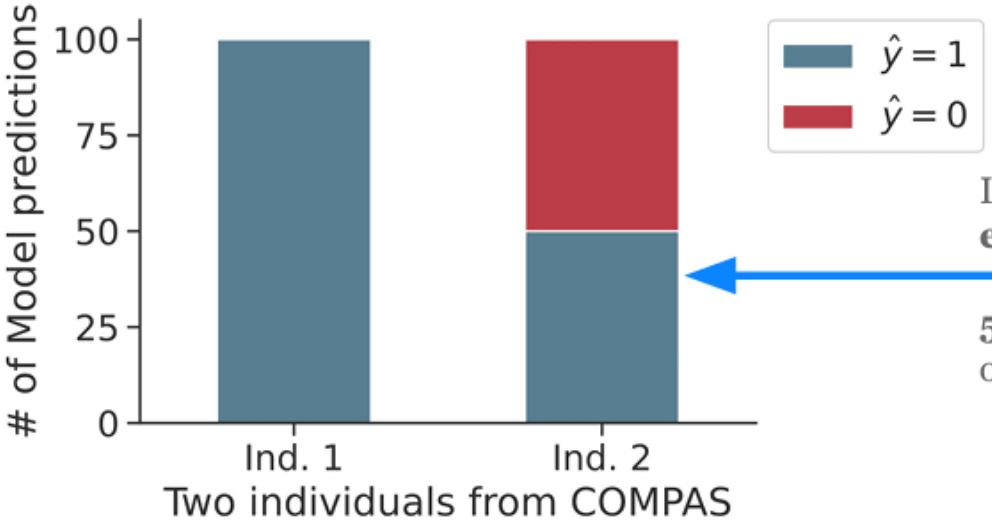
(= recidivates)

(= does not recidivate)

Training 100 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set

Arbitrariness and fairness



(= recidivates)

(= does not recidivate)

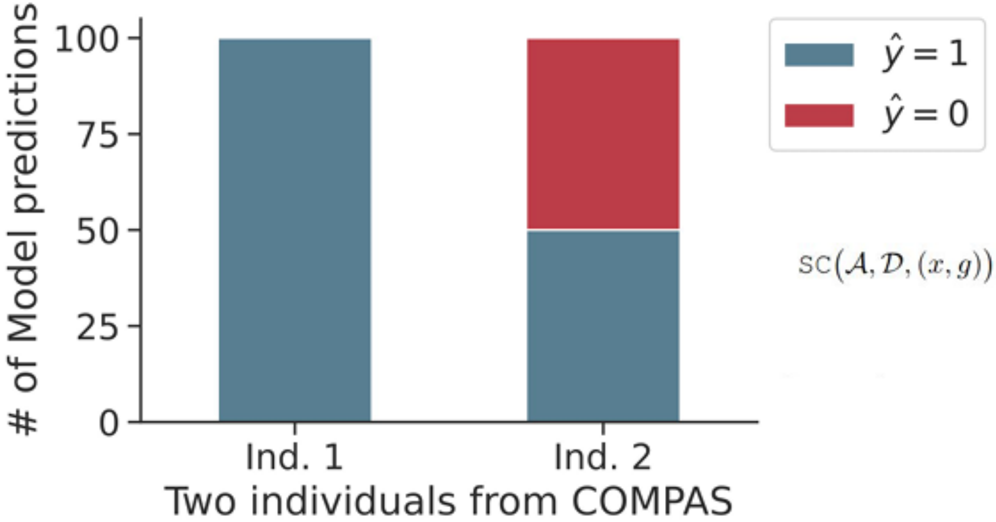
Learning process is **effectively random**

50 models predict each outcome for Ind. 2

Training 100 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set

Arbitrariness and fairness



$\hat{y} = 1$ (= recidivates)

$\hat{y} = 0$ (= does not recidivate)

$$\text{SC}(\mathcal{A}, \mathcal{D}, (x, g)) \triangleq \mathbb{P}_{h_{D_i}\sim\mu, h_{D_j}\sim\mu}\{h_{D_i}(x) = h_{D_j}(x)\}$$
$$= \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}[h_{D_i}(x) = h_{D_j}(x)].$$

We turn this picture into a metric (***self-consistency***) to capture ***arbitrariness***

Our contributions

Quantifying ***arbitrariness*** via ***self-consistency***

Developing an algorithm that ***abstains*** from making arbitrary predictions

Running a large-scale empirical study on the ***role of arbitrariness in fair classification***

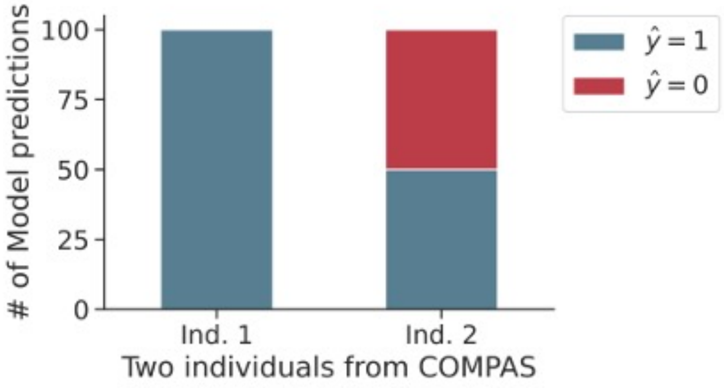
Packaging a large-scale dataset (won't get into this, but at the end will explain why)

From intuition to metric

$$\textit{self-consistency}^{*} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

Defined in terms of # of bootstrap replicates B

B_0 = the number of 0 predictions
B_1 = the number of 1 predictions



Definition 3. For all pairs of possible models $h_{\boldsymbol{D}_i}, h_{\boldsymbol{D}_j} \sim \mu$ $(i \neq j)$, the **self-consistency of the learning process** for a test $(\boldsymbol{x}, \boldsymbol{g})$ is

$$\mathsf{SC}(\mathcal{A}, \mathbb{D}, (\boldsymbol{x}, \boldsymbol{g})) \triangleq \mathbb{E}_{h_{\boldsymbol{D}_i} \sim \mu, h_{\boldsymbol{D}_j} \sim \mu}\left[h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})\right] = p_{h_{\boldsymbol{D}_i} \sim \mu, h_{\boldsymbol{D}_j} \sim \mu}\left(h_{\boldsymbol{D}_i}(\boldsymbol{x}) = h_{\boldsymbol{D}_j}(\boldsymbol{x})\right). \quad (2)$$

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 $B = B_0 + B_1 > 1$,

$$\hat{\mathsf{SC}}(\mathcal{A}, \hat{\mathbb{D}}, (\boldsymbol{x}, \boldsymbol{g})) := \frac{1}{B(B-1)} \sum_{i \neq j} \mathbf{1}\left[\hat{h}_{\hat{\boldsymbol{D}}_i}(\boldsymbol{x}) = \hat{h}_{\hat{\boldsymbol{D}}_j}(\boldsymbol{x})\right] = 1 - \frac{2B_0B_1}{B(B-1)}. \quad (3)$$

https://afedercooper.info

From intuition to metric

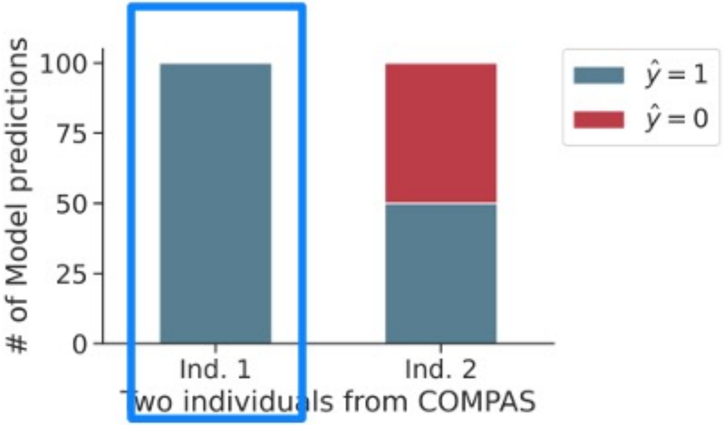
$$\textit{self-consistency} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

Defined in terms of # of bootstrap replicates B

B_0 = the number of 0 predictions
B_1 = the number of 1 predictions

Interpretation

a value on [~0.5, **1**]



***B* = 100 logistic regression models**

Ind. 1: $B_0 = 0$, $B_1 = 100$

Ind. 2: $B_0 = 50$, $B_1 = 50$

https://afedercooper.info

From intuition to metric

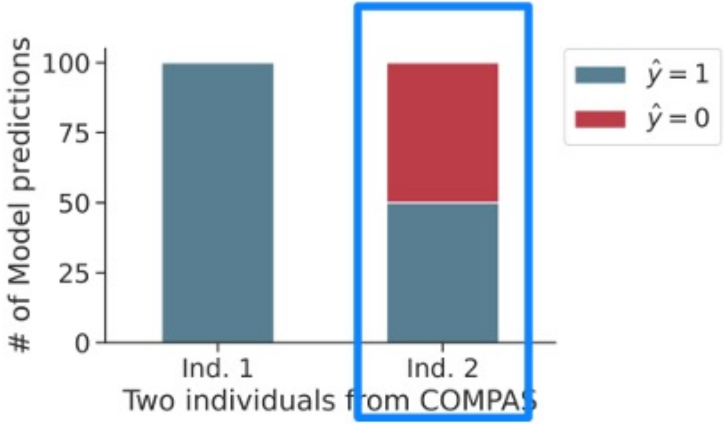
$$\textit{self-consistency} = 1 - \frac{2B_0 B_1}{B(B-1)}.$$

Defined in terms of # of bootstrap replicates B

B_0 = the number of 0 predictions
B_1 = the number of 1 predictions

Interpretation

a value on [**~0.5**, 1]



B **= 100 logistic regression models**

Ind. 1: $B_0 = 0$, $B_1 = 100$

Ind. 2: $B_0 = 50$, $B_1 = 50$

https://afedercooper.info

From intuition to metric

$$\textit{self-consistency} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

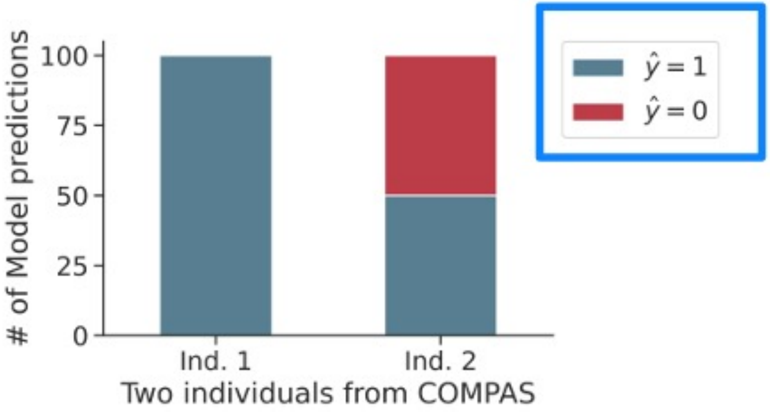
Defined in terms of # of bootstrap replicates B

B_0 = the number of 0 predictions
B_1 = the number of 1 predictions

Interpretation

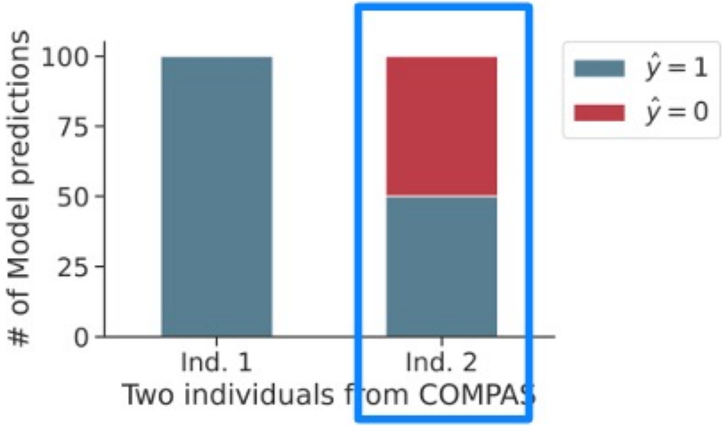
a value on [~0.5, 1]

does ***not*** depend on dataset labels y



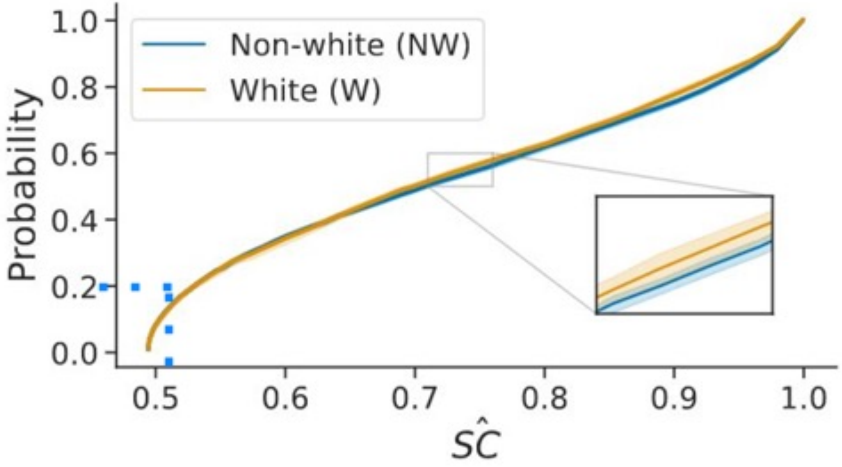
https://afedercooper.info

Illustrating self-consistency



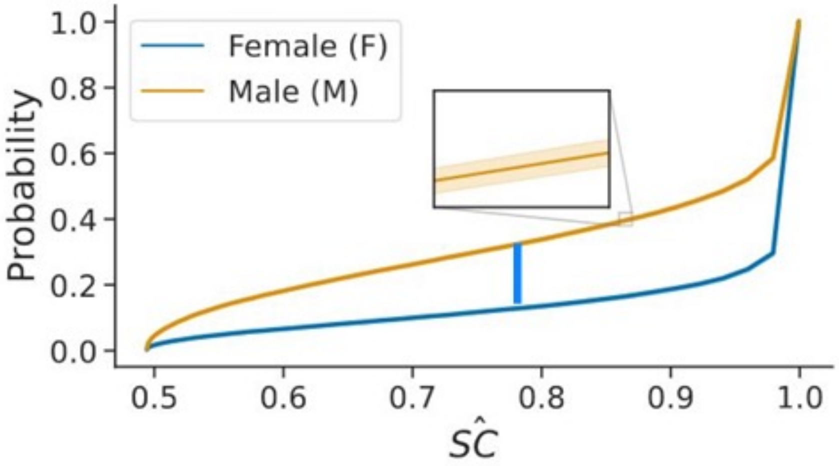
About 20% of COMPAS looks like Ind. 2

Their predictions are ***arbitrary***



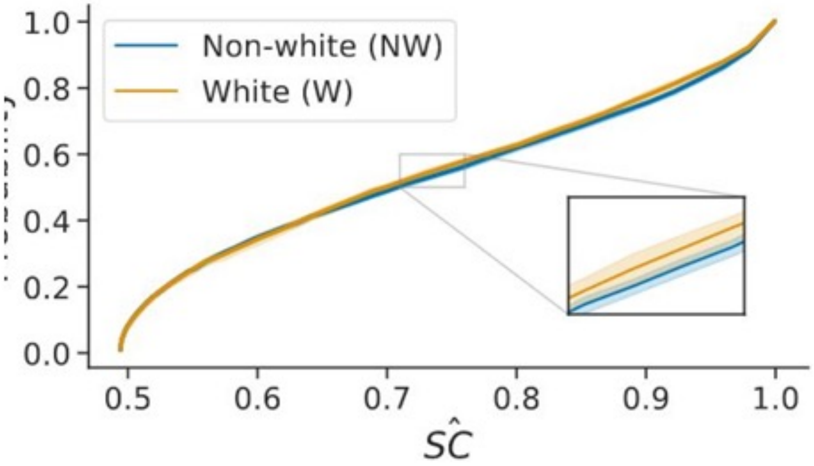
COMPAS, random forests, B=101
(mean +/- STD over 10 trials)

Illustrating self-consistency



Old Adult, random forests, B=101
(mean +/- STD over 10 trials)

systematic arbitrariness
(actually happens rarely in practice)



COMPAS, random forests, B=101
(mean +/- STD over 10 trials)

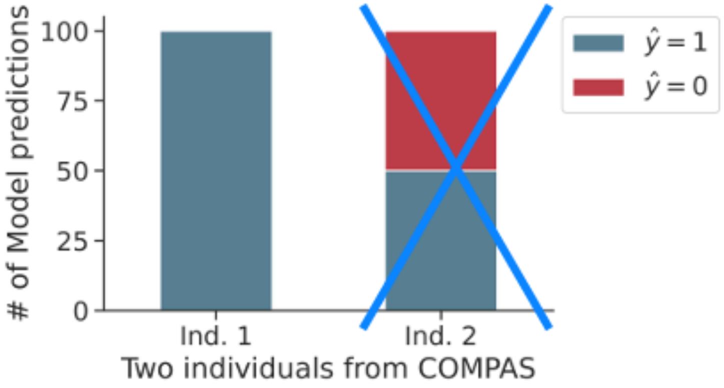
Our algorithm (really really quickly)

Self-consistency is derived from variance (High self-consistency → low variance)...

...so let's try to do variance reduction to improve self-consistency

→ Leo Breiman's 1996 bagging algorithm (with a twist)

Abstain if too self-inconsistent



An example from our results: COMPAS

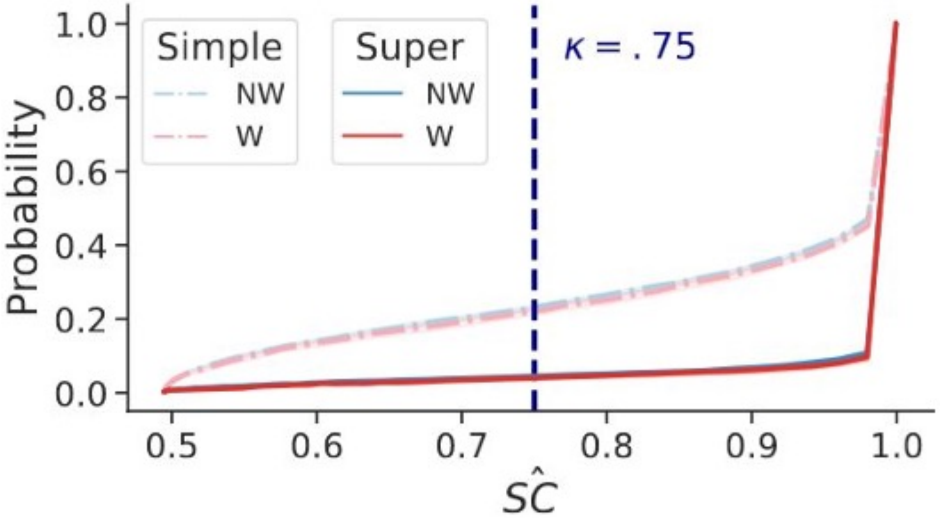
Fairness metrics

Examine false positive rate disparities

We yield results that are very close-to-fair (<2% disparity in FPR) (and **super** variant abstains <5%)

And we haven't run any algorithmic fairness method!

	Simple	Super
$\Delta\hat{\text{FPR}}$	$3.0 \pm 1.4\%$	$1.8 \pm 1.0\%$
$\hat{\text{FPR}}_{\text{NW}}$	$11.4 \pm 1.0\%$	$12.9 \pm .8\%$
$\hat{\text{FPR}}_{\text{W}}$	$8.4 \pm 1.0\%$	$11.1 \pm .6\%$



COMPAS, logistic regression, B=101
(mean +/- STD over 10 trials)

https://afedercooper.info

Summarizing our experiments

Overall, these patterns hold (and more)

Datasets:

- (South) German Credit
- COMPAS
- Old Adult
- Taiwan Credit
- New Adult (race, sex)
  - Income
  - Public Coverage
  - Employment
- Home Mortgage Disclosure Act (race, ethnicity, sex)
  - NY - 2017
  - TX - 2017

We improve self-consistency, attain accuracy, *and* (in almost every single case) **achieve close-to-fairness** ...

***... without* using a single field-standard theory-backed technique that aims to improve fairness**

We packaged this because we struggled to find algorithmic unfairness above

Models: logistic regression, decision trees, random forests, MLPs, SVMs (**most common fair classification models**)

Is this notion of “fairness” easier for lawmakers to understand and implement?

Takeaways

This finding is **really shocking**

What does it mean for empirical rigor and reproducibility of existing approaches?

Do fairness interventions actually improve fairness in practice?

Are conclusions from prior empirical work confounded by a more general problem of arbitrariness in predictions?

Arbitrariness is rampant when predicting on social data.

How practically useful are prior theoretical formulation choices?

What happens when what we take as given in research, turns out to not be the case? What would have happened if we did this simple bagging approach years ago?

What about mutli-class classification? Do you think it extends?