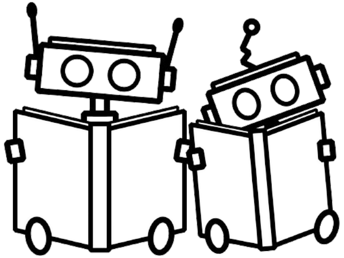
Arbitrariness and Social Prediction:

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**

Arbitrariness and fairness

Existing fairness practices…

Look at **error rates across groups**

typically, for a **single** model

Arbitrariness and fairness

Existing fairness practices…

Look at **error rates across groups** (**definite**)

typically, for a **single** model (**feasible**)

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*)

An intuition for arbitrariness

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

(Dataset used to predict prison recidivism)

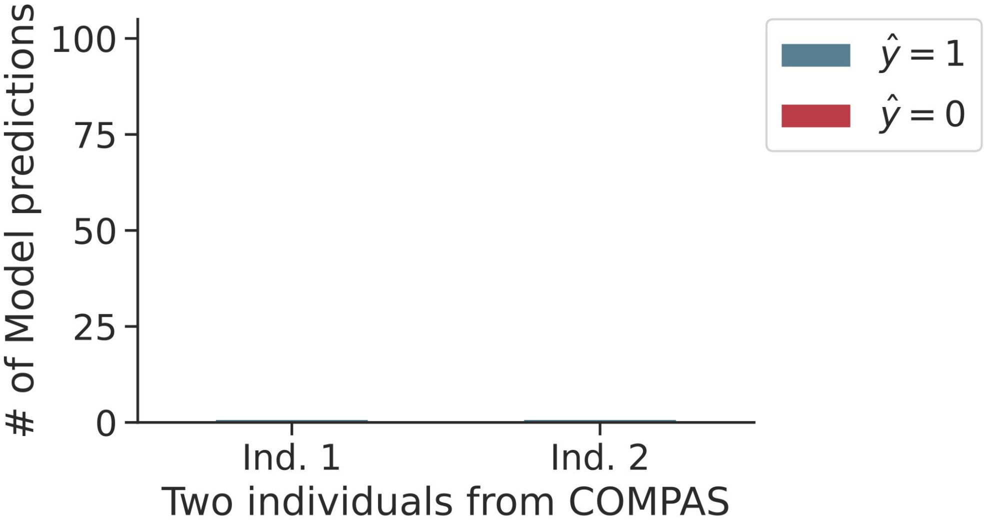
An intuition for arbitrariness

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)

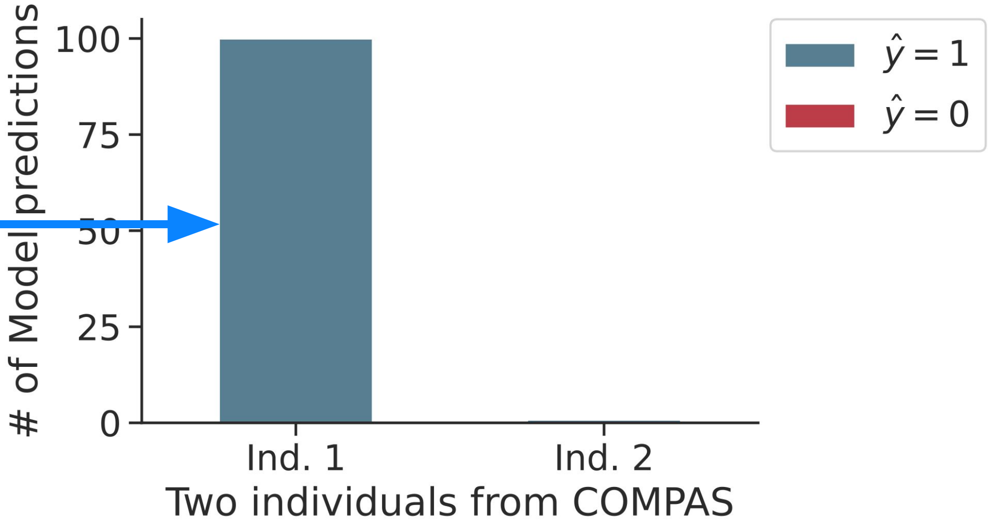
Training 100 different logistic regression models on COMPAS using bootstrapping

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



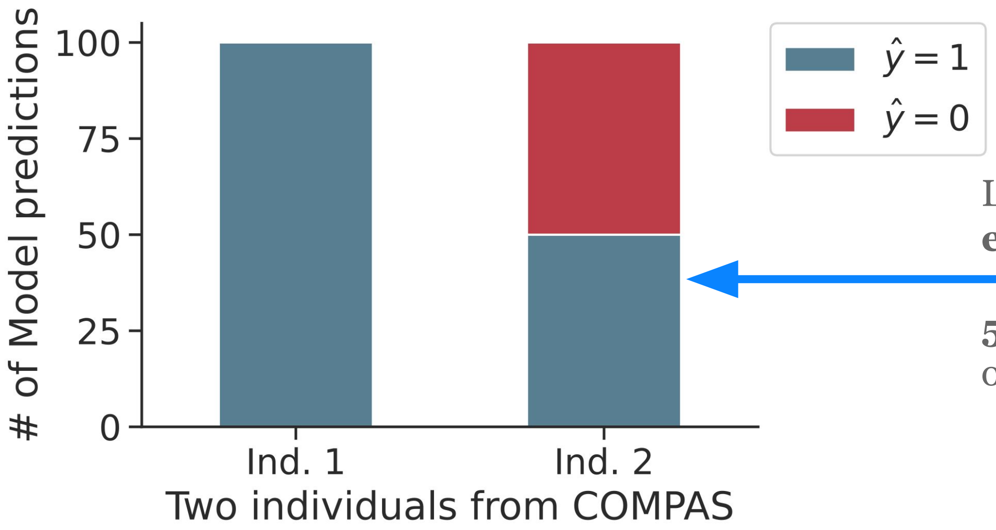
(= recidivates)

(= 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)

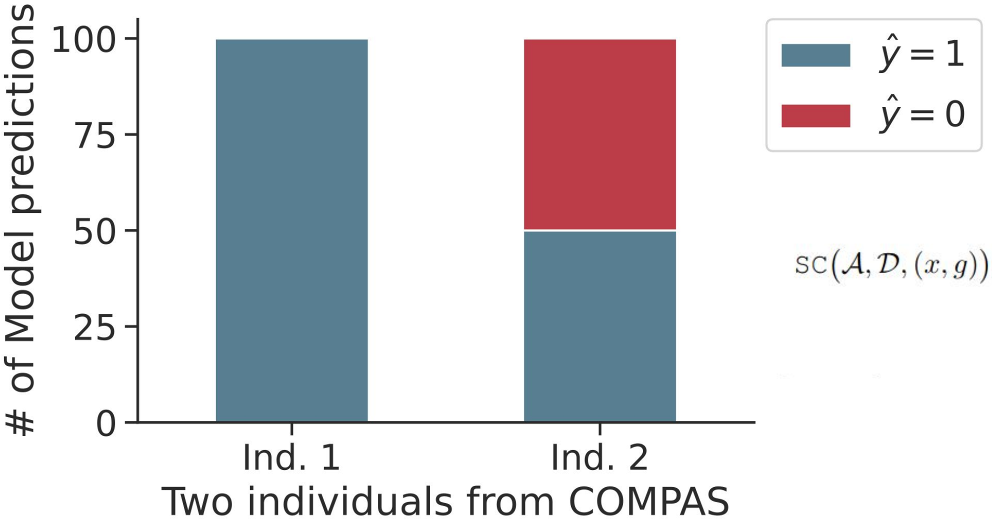
Learning process is **effectively random**

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



(= recidivates)

(= does not recidivate)

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

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)

Our contributions

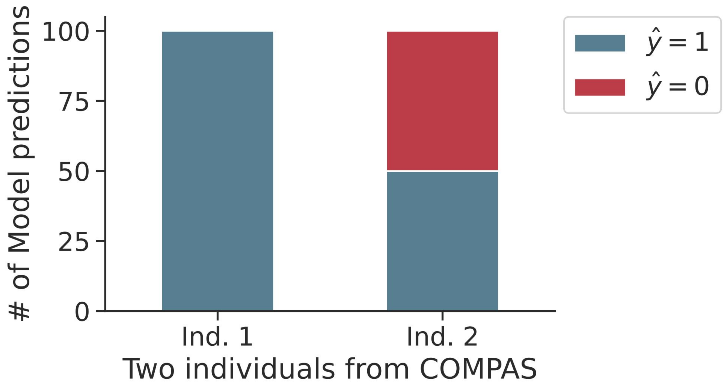
Quantifying ***arbitrariness*** via ***self-consistency***

~~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

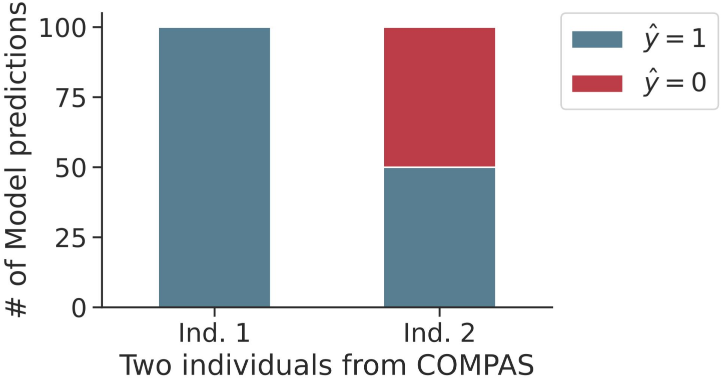


From intuition to metric

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

Defined in terms of # of bootstrap replicates B

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

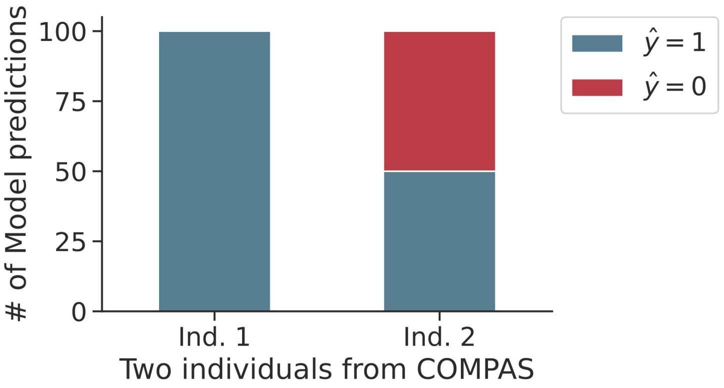


From intuition to metric

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

Defined in terms of # of bootstrap replicates B

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



*This is our **empirical approximation definition**

From intuition to metric

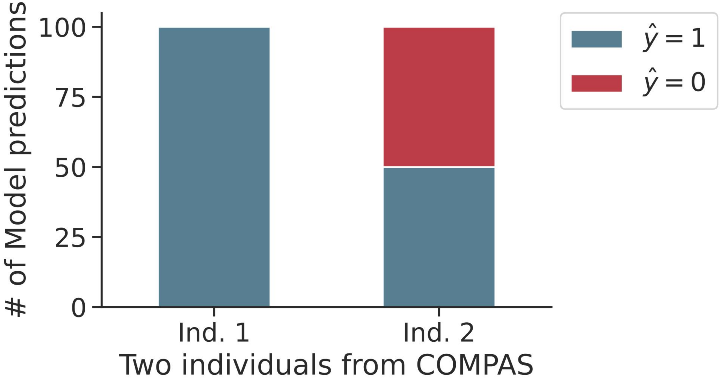
$$\textit{\textbf{self-consistency}} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

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]



From intuition to metric

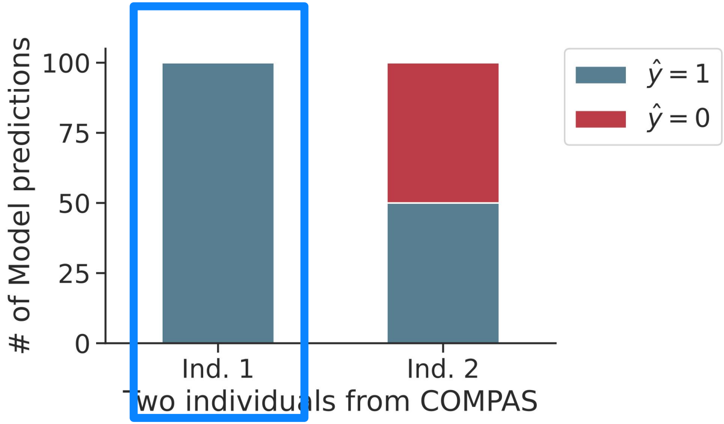
$$\textit{self-consistency} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

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: $B_0 = 50$, $B_1 = 50$

From intuition to metric

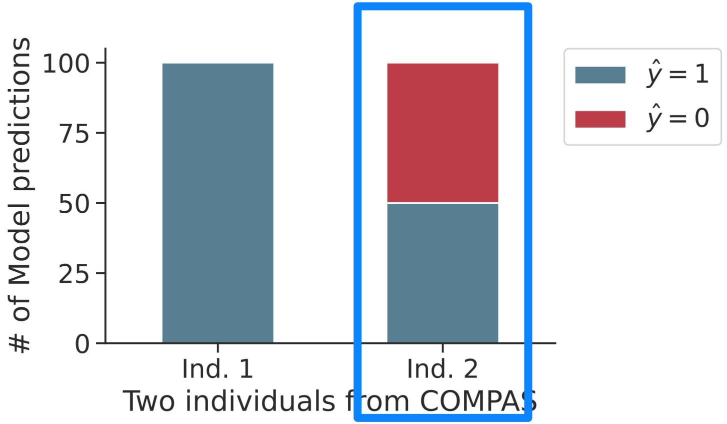
$$\textit{self-consistency} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

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: $B_0 = 50$, $B_1 = 50$

From intuition to metric

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

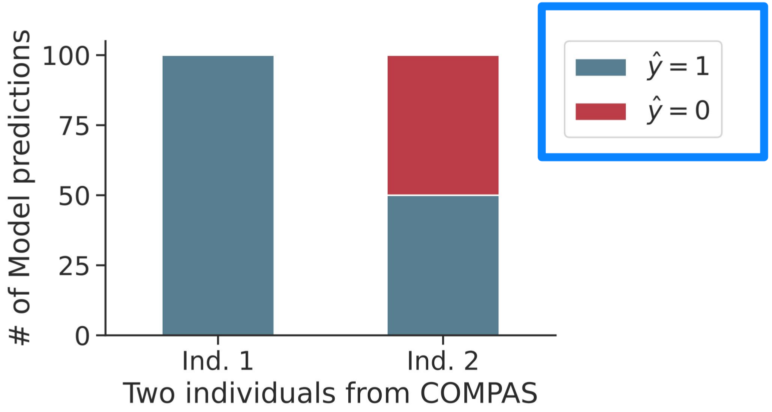
Defined in terms of # of bootstrap replicates B

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

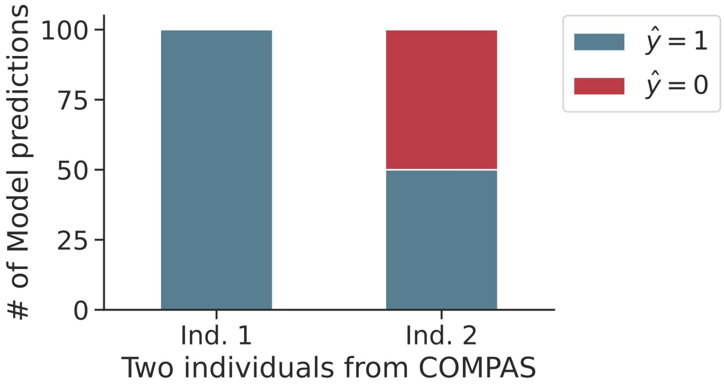
Interpretation

a value on [~0.5, 1]

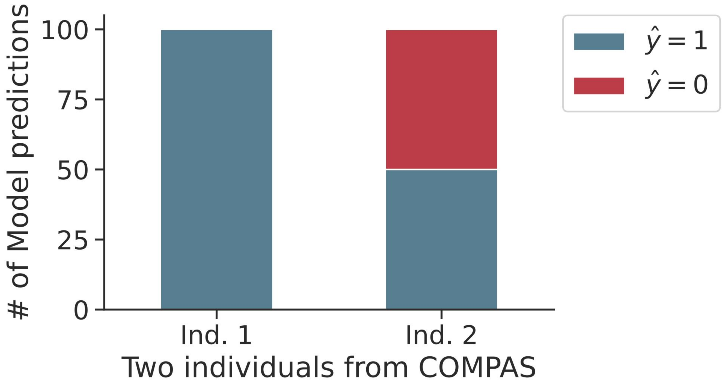
does ***not*** depend on dataset labels y

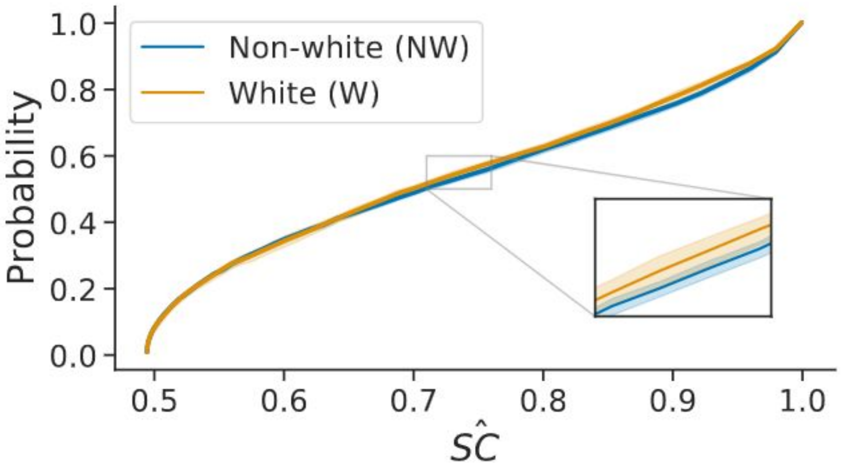


Illustrating self-consistency



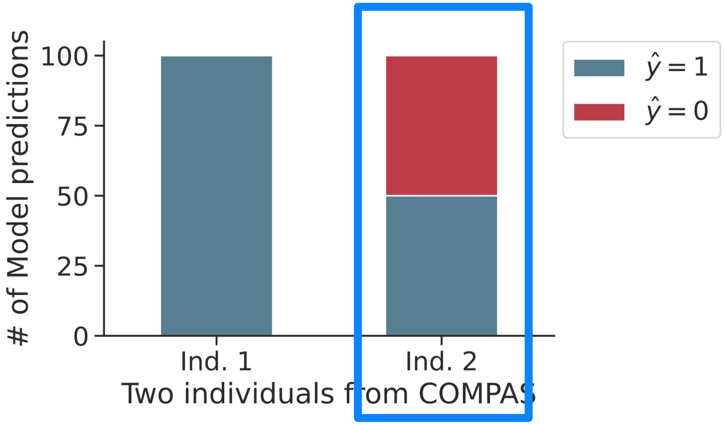
Illustrating self-consistency





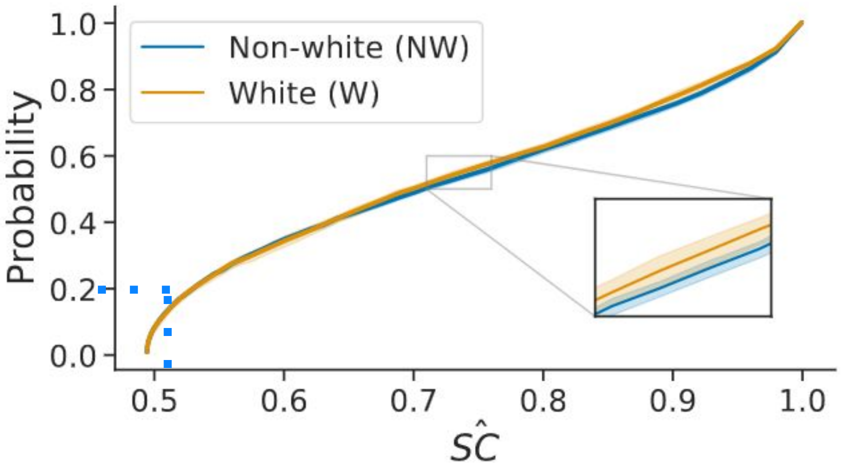
COMPAS, random forests, B=101
(mean +/- STD over 10 trials)

Illustrating self-consistency



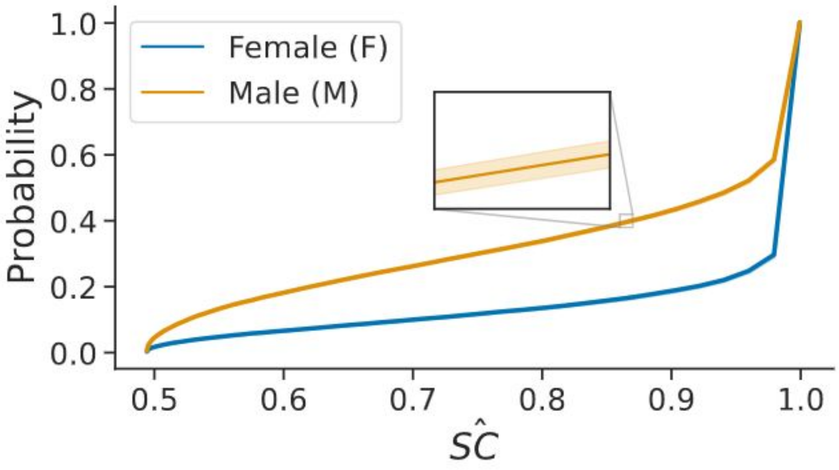
About 20% of COMPAS looks like Ind. 2

Their predictions are ***arbitrary***

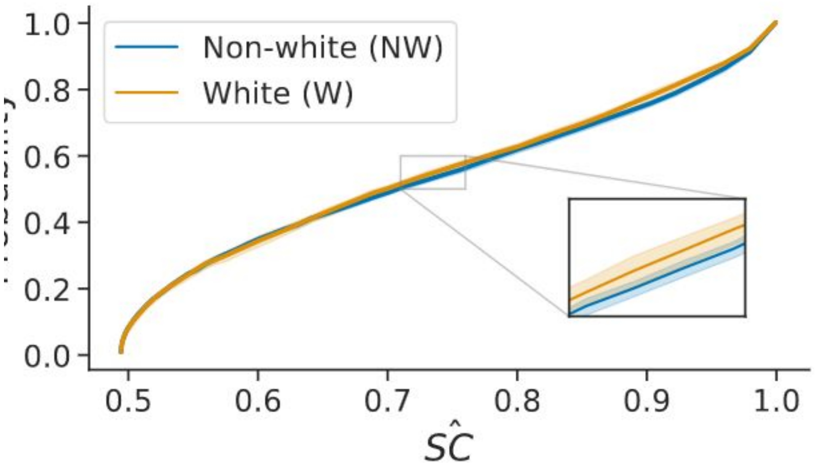


COMPAS, random forests, *B*=101
(mean +/- STD over 10 trials)

Illustrating self-consistency



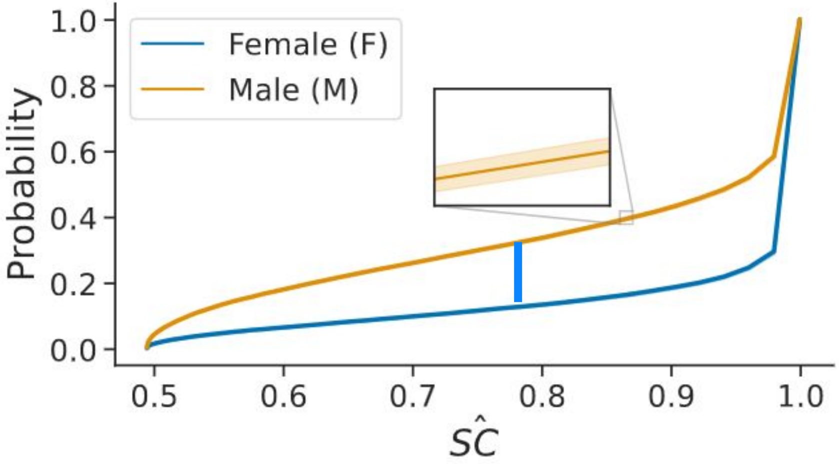
Old Adult, random forests, B=101
(mean +/- STD over 10 trials)



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

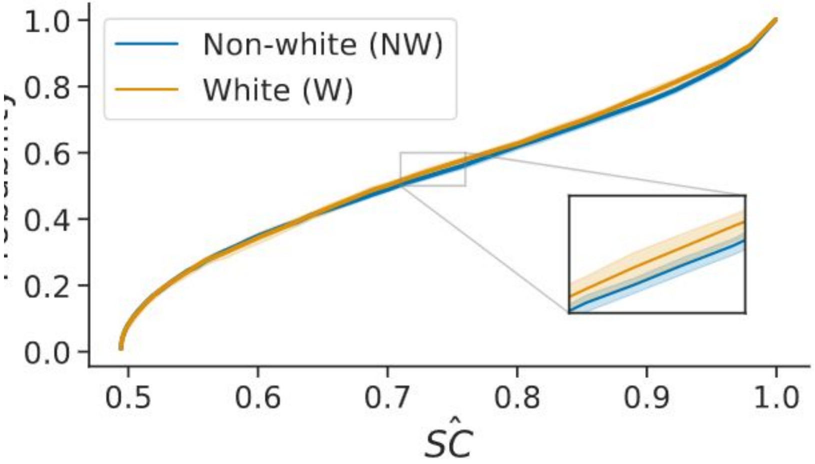
https://afedercooper.info

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 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)

Our algorithm (really really quickly)

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

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

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

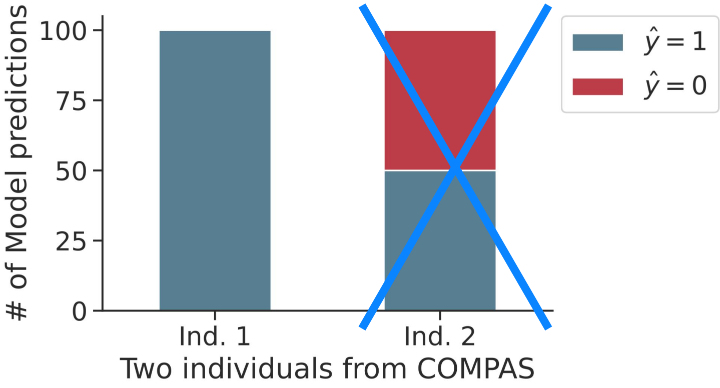
Our algorithm (really really quickly)

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



Our contributions

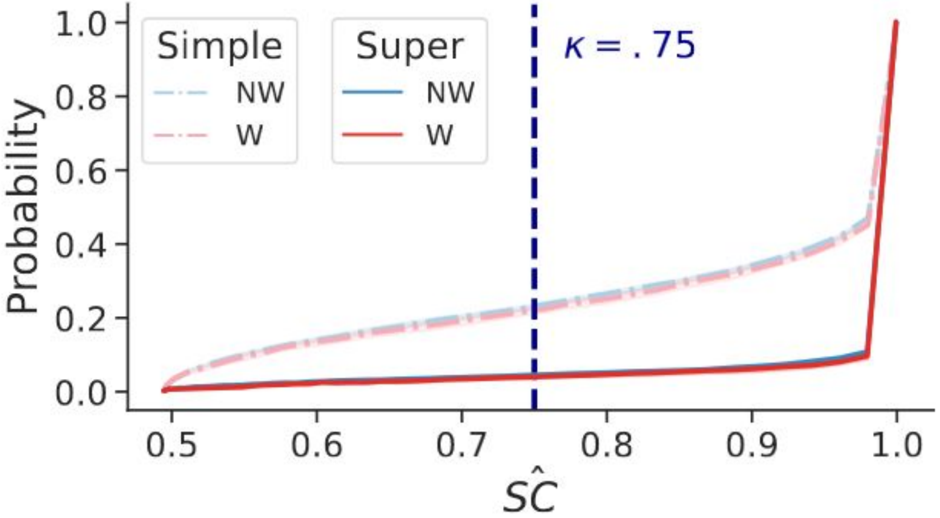
Quantifying ***arbitrariness*** via ***self-consistency***

~~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)

An example from our results: COMPAS

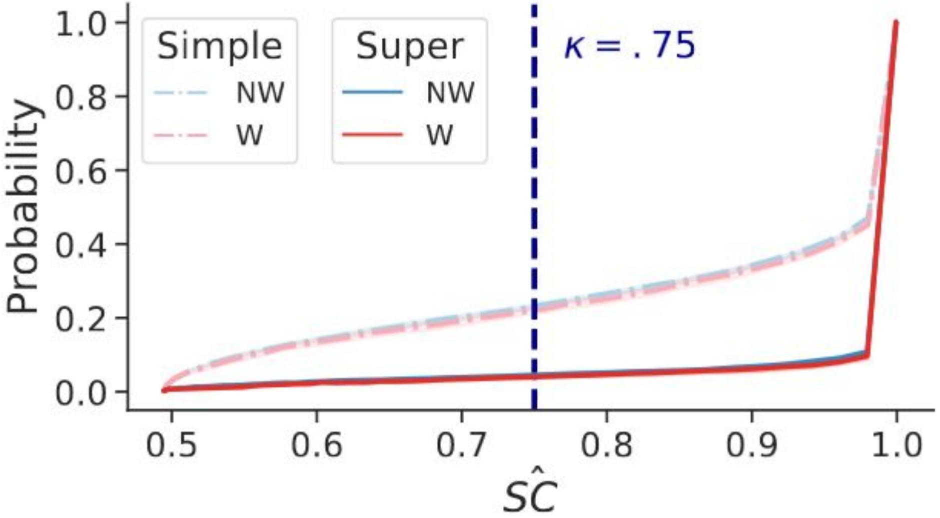


COMPAS, logistic regression, B=101
(mean +/- STD over 10 trials)

An example from our results: COMPAS

Fairness metrics

Examine false positive rate disparities



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

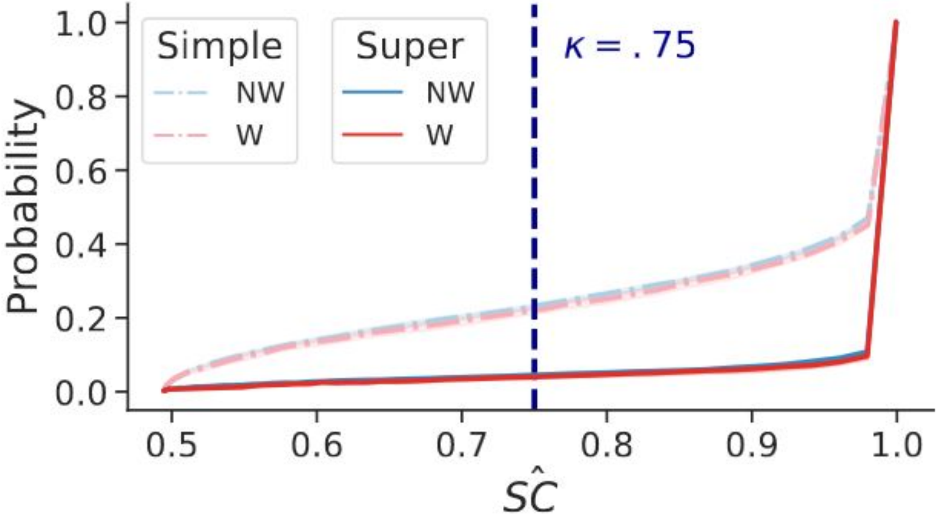
An example from our results: COMPAS

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%)

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



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

An example from our results: COMPAS

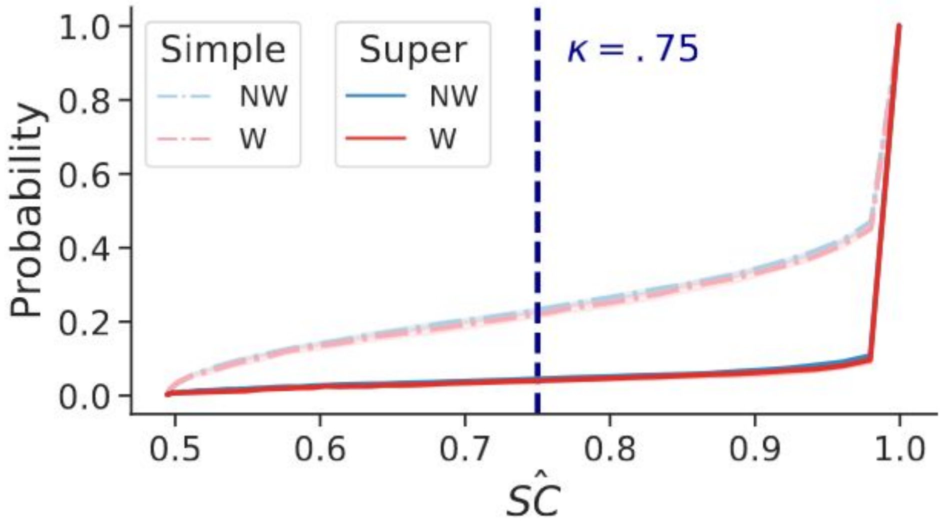
Fairness metrics

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)

Summarizing our experiments

Summarizing our experiments

Datasets:

- **(South) German Credit**
- **COMPAS**
- **Old Adult**
- **Taiwan Credit**

Summarizing our experiments

Datasets:

- (South) German Credit
- COMPAS
- Old Adult
- Taiwan Credit
- **New Adult (race, sex)**
    - **Income**
    - **Public Coverage**
    - **Employment**

Summarizing our experiments

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**

Summarizing our experiments

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

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

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** …

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

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** …

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**)

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**)

There are huge takeaways here

(Please ask me about the details)

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?

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?

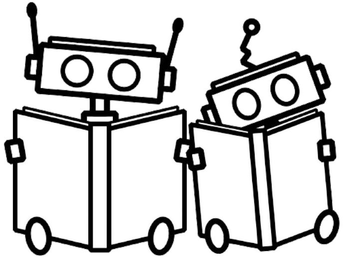
Arbitrariness is rampant when predicting on social data.

How practically useful are prior theoretical formulation choices?

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

A. Feder Cooper
Cornell University | The GenLaw Center

Thank you!



https://afedercooper.info