Cornell Bowers C·IS
College of Computing
and Information Science

Is My Prediction Arbitrary?
Measuring Self-Consistency in Fair Classification

A. Feder Cooper
Ph.D. Candidate
Cornell University Department of Computer Science

Presenting work in collaboration with
Katherine Lee (Google DeepMind), **Solon Barocas** (Microsoft Research & Cornell),
Christopher De Sa (Cornell), **Siddhartha Sen** (Microsoft Research), **Baobao Zhang** (Syracuse)

https://afedercooper.info

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

- Does ML model accuracy differ for different subsets of individuals in the test set?
- Do ML models perform worse for marginalized demographic groups?

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

Does ML model accuracy differ for different subsets of individuals in the test set?

Do ML models perform worse for marginalized demographic groups?

Propublica (2015): Predictive accuracy can vary widely across Non-white and white individuals,

when performing "risk assessment" (for determining bail) using the **COMPAS dataset**

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

- Does ML model accuracy differ for different subsets of individuals in the test set?

- Do ML models perform worse for marginalized demographic groups?

Propublica (2015): Predictive accuracy can vary widely across Non-white and white individuals,

when performing “risk assessment” (for determining bail) using the **COMPAS dataset**

During test time **Non-white individuals denied bail more often than white individuals, especially**

when the underlying label indicated bail should be granted

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

Does ML model accuracy differ for different subsets of individuals in the test set?

Do ML models perform worse for marginalized demographic groups?

Propublica (2015): Predictive accuracy can vary widely across Non-white and white individuals,

when performing "risk assessment" (for determining bail) using the **COMPAS dataset**

During test time **Non-white individuals denied bail more often than white individuals, especially**

when the underlying label indicated bail should be granted (Unequal False Positive Rates)

Why algorithmic fairness?

Most supervised ML traditionally emphasizes (overall) ***accuracy*** as the performance metric

There may be other performance concerns in learning tasks that involve social data

Does ML model accuracy differ for different subsets of individuals in the test set?

Do ML models perform worse for marginalized demographic groups?

Propublica (2015): Predictive accuracy can vary widely across Non-white and white individuals,

when performing "risk assessment" (for determining bail) using the **COMPAS dataset**

During test time **Non-white individuals denied bail more often than white individuals, especially**

when the underlying label indicated bail should be granted (Unequal False Positive Rates)

https://afedercooper.info

Algorithmic fairness classification research

Emphasizes

group-specific accuracy

typically tries to equalize ***error rates*** across demographic groups (**false positives, false negatives, some combination**)

Algorithmic fairness classification research

Emphasizes

group-specific accuracy

typically tries to equalize ***error rates*** across demographic groups (**false positives, false negatives, some combination**)

disparities in group-specific accuracy

are treated as metrics of ***(un)fairness*** (e.g., differences in false positive rates)

Algorithmic fairness classification research

Emphasizes

group-specific accuracy

typically tries to equalize ***error rates*** across demographic groups (**false positives, false negatives, some combination**)

disparities in group-specific accuracy

are treated as metrics of ***(un)fairness*** (e.g., differences in false positive rates)

theory

writing proofs about algorithms that make claims about guarantees regarding overall **accuracy** and **algorithmic fairness**

Algorithmic fairness classification research

Emphasizes

group-specific accuracy

typically tries to equalize ***error rates*** across demographic groups (**false positives, false negatives, some combination**)

disparities in group-specific accuracy

are treated as metrics of ***(un)fairness*** (e.g., differences in false positive rates)

theory

writing proofs about algorithms that make claims about guarantees regarding overall **accuracy** and **algorithmic fairness**

Empirics tend to be secondary

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) models
2) problems the models are supposed to be predicting

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) models
2) problems the models are supposed to be predicting

Typically, researchers cross-validate a very small handful of models (e.g., 5 logistic regressions or random forests)

It turns out this gives unreliable estimates of expected error in this domain

We bootstrap

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) models
2) problems the models are supposed to be predicting

Typically, researchers cross-validate a very small handful of models (e.g., 5 logistic regressions or random forests)

It turns out this gives unreliable estimates of expected error in this domain

We bootstrap: What do empirics change about the study and use of fairness?

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) **models**
2) **problems the models are supposed to be predicting**

Typically, researchers cross-validate a very small handful of models (e.g., 5 logistic regressions or random forests)

It turns out this gives unreliable estimates of expected error in this domain

We bootstrap: What do empirics change about the study and use of fairness?

Deployed models should ***abstain from predicting*** when the predictions they produce are ***arbitrary*** (**This is exactly what our algorithm does**)

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) models
2) problems the models are supposed to be predicting

Typically, researchers cross-validate a very small handful of models (e.g., 5 logistic regressions or random forests)

It turns out this gives unreliable estimates of expected error in this domain

We bootstrap: What do empirics change about the study and use of fairness?

Deployed models should ***abstain from predicting*** when the predictions they produce are ***arbitrary*** (This is exactly what our algorithm does)

The real question: How do you know that predictions are ***arbitrary***?

An empirical emphasis on algorithmic fairness

Instead of theory, we focus on empirical methods, provide substantial empirical analysis

We want future fairness researchers to use methods like ours because they give more reliable evaluation of

1) models
2) problems the models are supposed to be predicting

Typically, researchers cross-validate a very small handful of models (e.g., 5 logistic regressions or random forests)

It turns out this gives unreliable estimates of expected error in this domain

We bootstrap: What do empirics change about the study and use of fairness?

Deployed models should ***abstain from predicting*** when the predictions they produce are ***arbitrary*** (This is exactly what our algorithm does)

The real question: How do you know that predictions are ***arbitrary***?

The answer turns out to be extremely simple

Could do more sophisticated things from the lit on model uncertainty, but in this setting (fair classification), we don't need to

An intuition for arbitrariness in empirical fairness

Training 10 different logistic regression models on **COMPAS** using bootstrapping

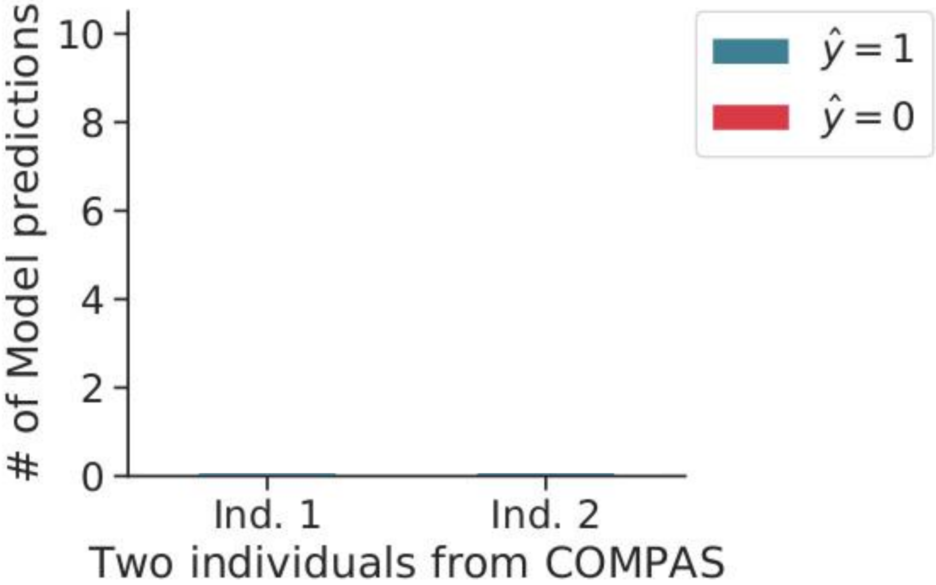
(Dataset used to predict
whether or not a person
will commit a crime;
used to determine bail)

An intuition for arbitrariness in empirical fairness

Training 10 different logistic regression models on COMPAS using **bootstrapping**

(split into train/test sets)
(resample train set)

An intuition for arbitrariness in empirical fairness



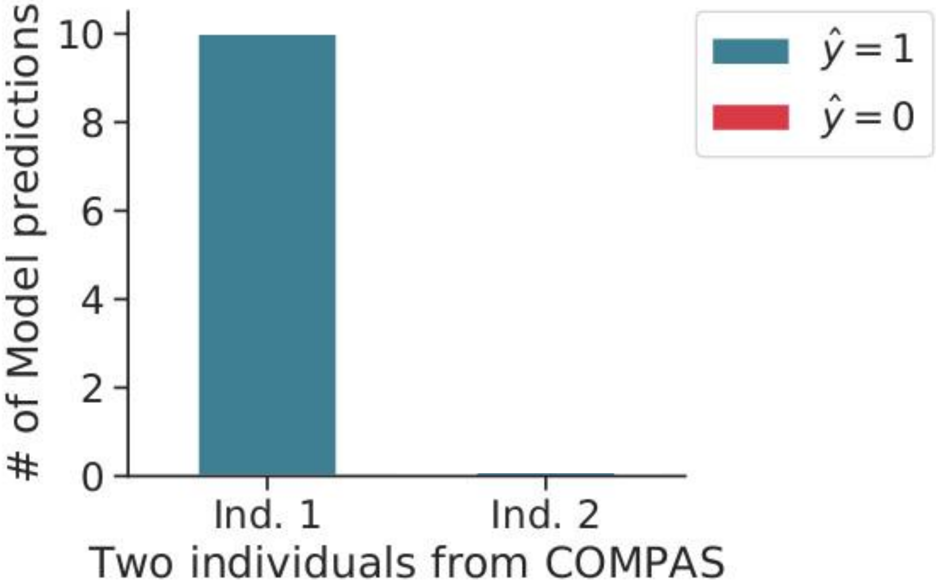
(= commits crime)

(= does not commit crime)

Training 10 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set (not used in training)

An intuition for arbitrariness in empirical fairness



(= commits crime)

(= does not commit crime)

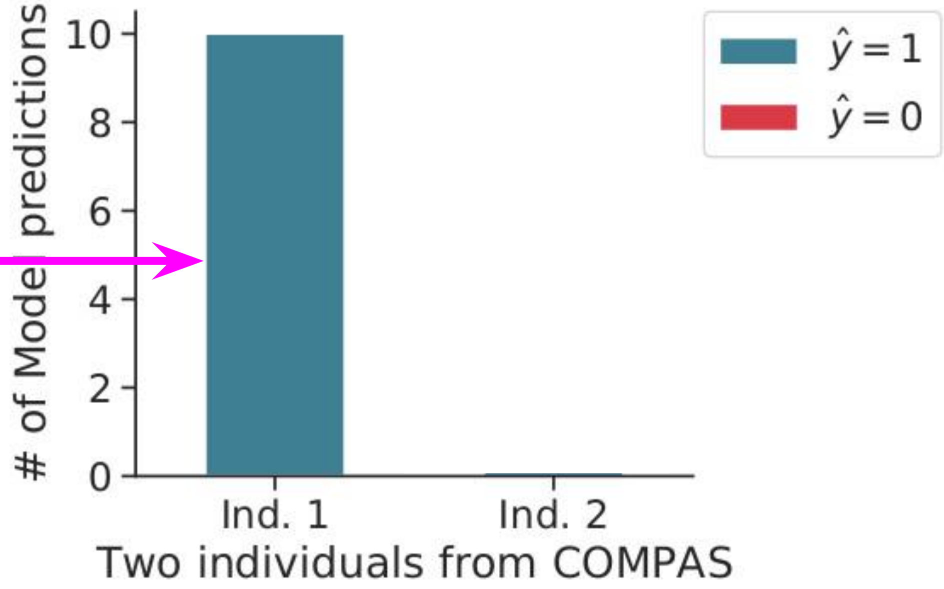
Training 10 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set (not used in training)

An intuition for arbitrariness in empirical fairness

Learning process is
really confident

All 10 models
predict the same
outcome for Ind. 1



(= commits crime)

(= does not commit crime)

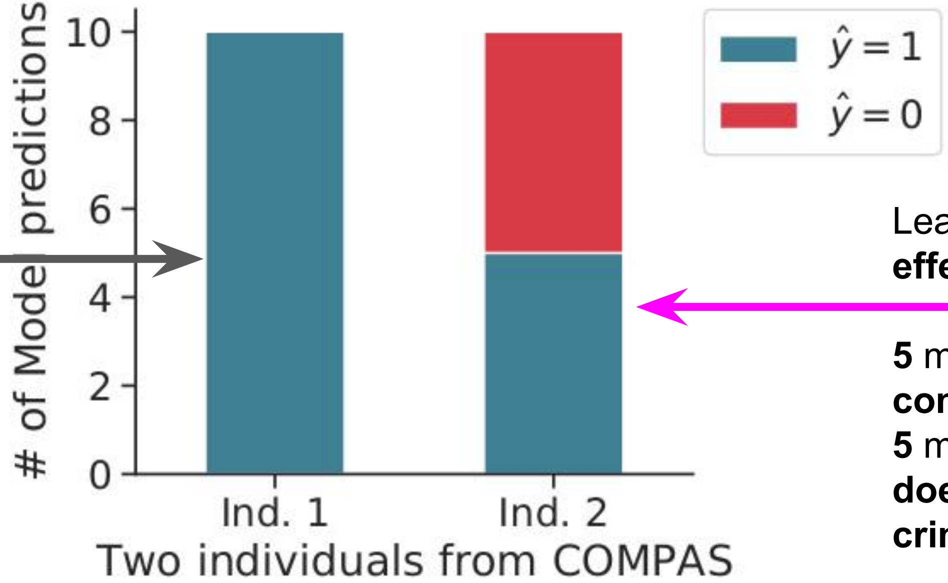
Training 10 different logistic regression models on COMPAS using bootstrapping

Looking at the resulting predictions for 2 individuals in the test set (not used in training)

An intuition for arbitrariness in empirical fairness

Learning process is **really confident**

All 10 models predict the same outcome for Ind. 1



(= commits crime)

(= does not commit crime)

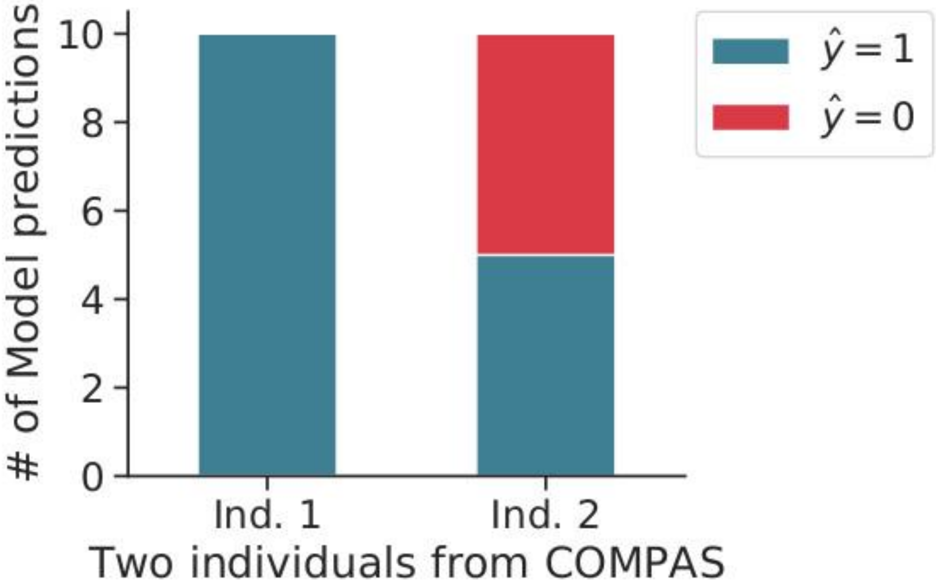
Learning process is **effectively random**

5 models predict **commits crime**, **5** models predict **does not commit crime** for Ind. 2

Training 10 different logistic regression models on COMPAS using bootstrapping

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

An intuition for arbitrariness in empirical fairness



(= commits crime)

(= does not commit crime)

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

We quantify and mitigate arbitrariness in fair classification

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)

Our contributions

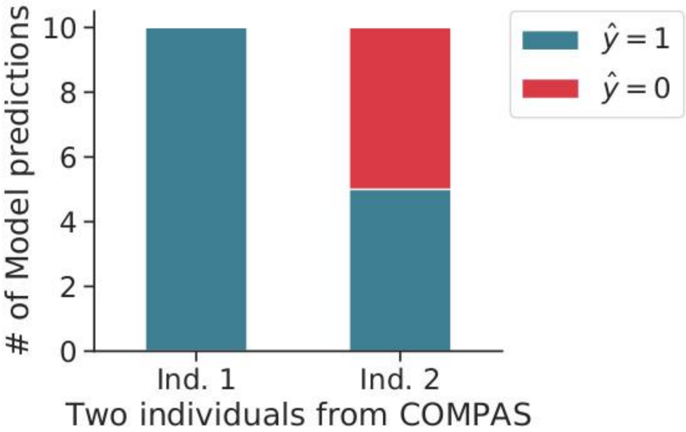
Quantifying ***arbitrariness*** via ***self-consistency***

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)

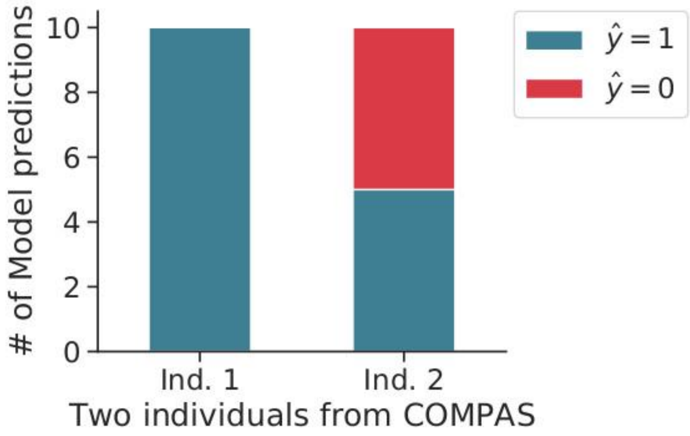
A metric: Self-consistency



A metric: Self-consistency

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

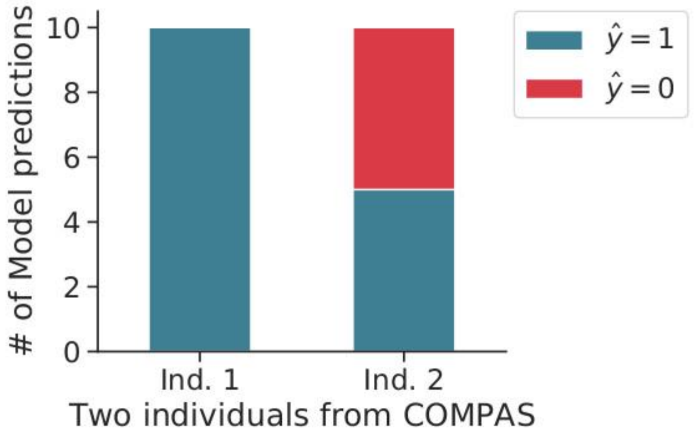
Defined in terms of # of bootstrap replicates *B*



A metric: Self-consistency

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

Defined in terms of # of bootstrap replicates *B*



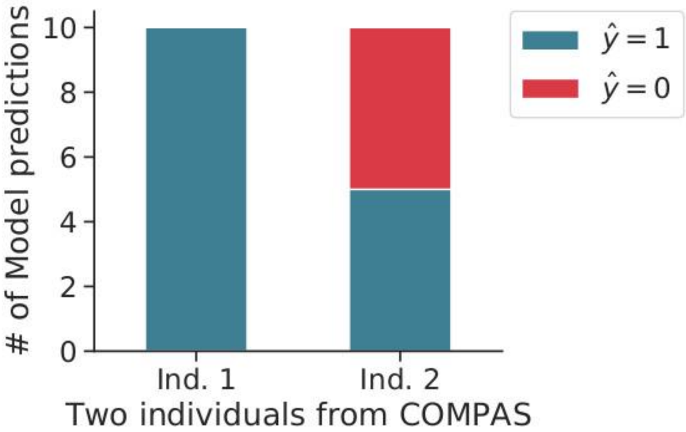
***B* = 10 logistic regression models**

A metric: Self-consistency

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



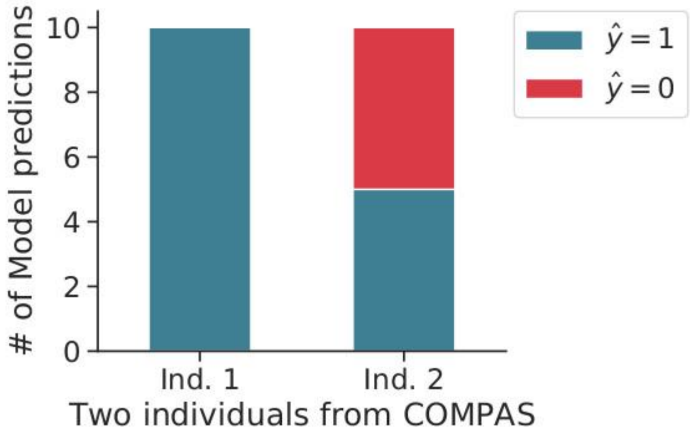
B = 10 logistic regression models

A metric: Self-consistency

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



B = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

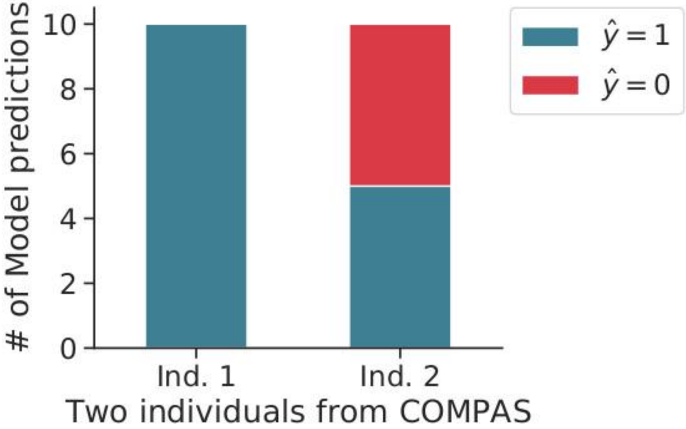
A metric: Self-consistency

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

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** …



B = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

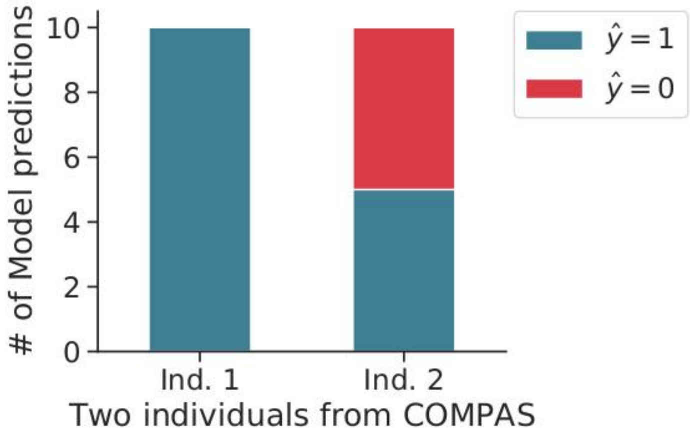
A metric: Self-consistency

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

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** …
… which is derived from a formal definition …



B = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

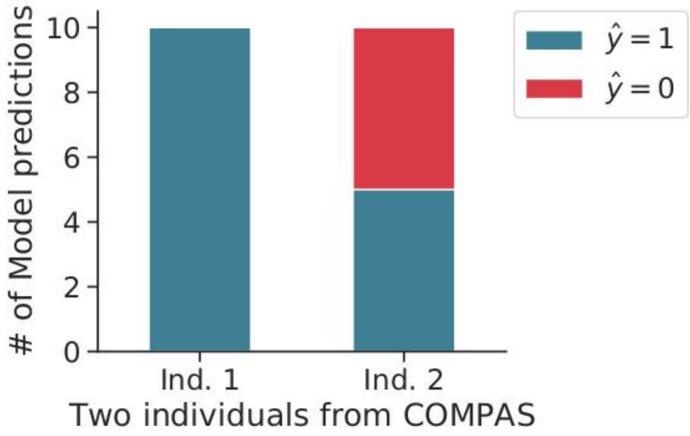
A metric: Self-consistency

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

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** …
… which is derived from a formal definition …
… which has nice properties because it is (further) derived from a formal definition of variance



B = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

A metric: Self-consistency

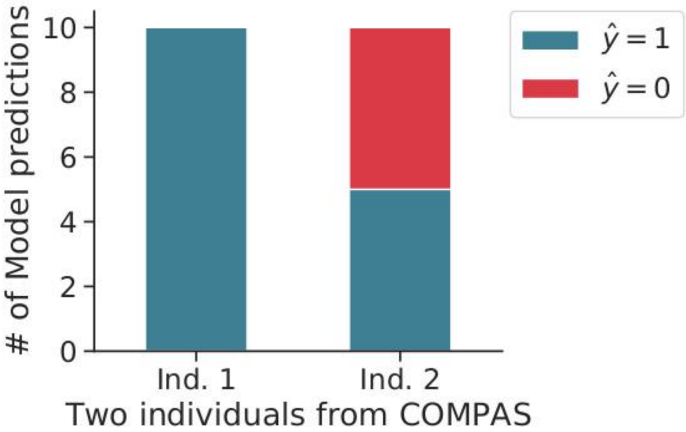
self-consistency* $= 1 - \frac{2B_0 B_1}{B(B-1)}.$

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 = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

A metric: Self-consistency

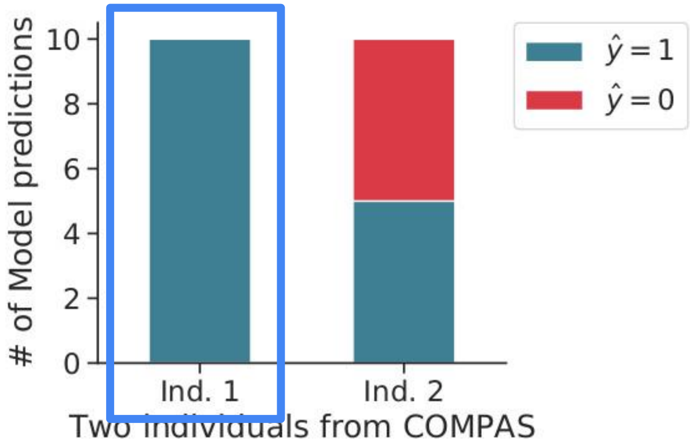
self-consistency* $= 1 - \frac{2B_0B_1}{B(B-1)}.$

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 = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

A metric: Self-consistency

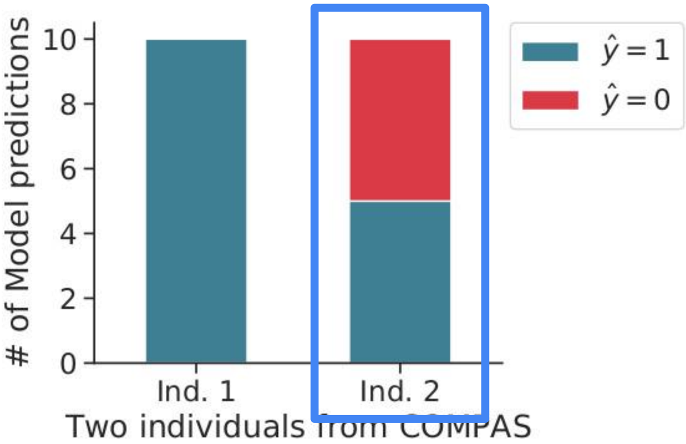
$$\textit{self-consistency}^{*} = 1 - \frac{2B_0B_1}{B(B-1)}.$$

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 = 10 logistic regression models

Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

A metric: Self-consistency

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

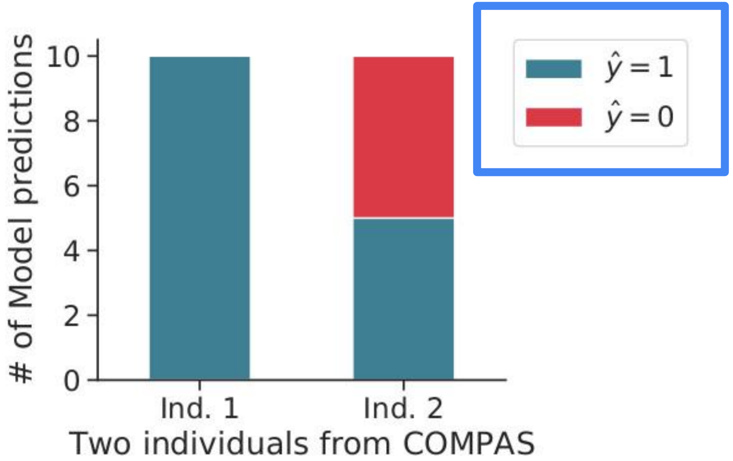
Defined in terms of # of bootstrap replicates *B*

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

Interpretation

a value on [~0.5, 1]

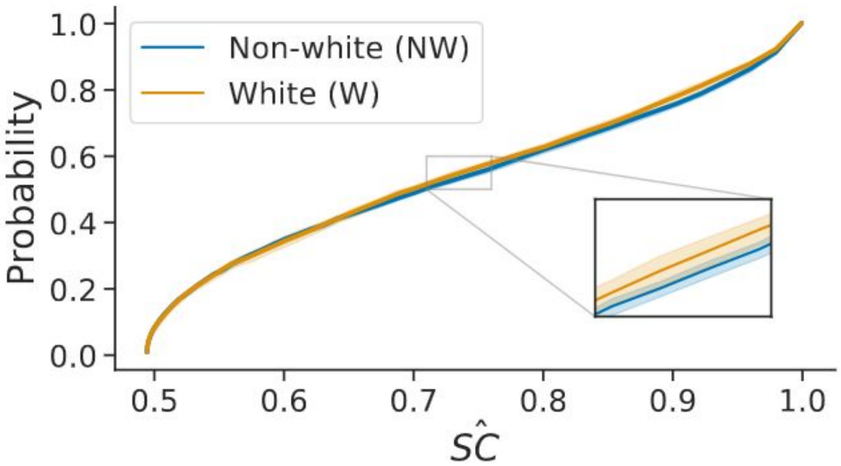
does ***not*** depend on dataset labels *y* (traditional fairness metrics do)



B = 10 logistic regression models

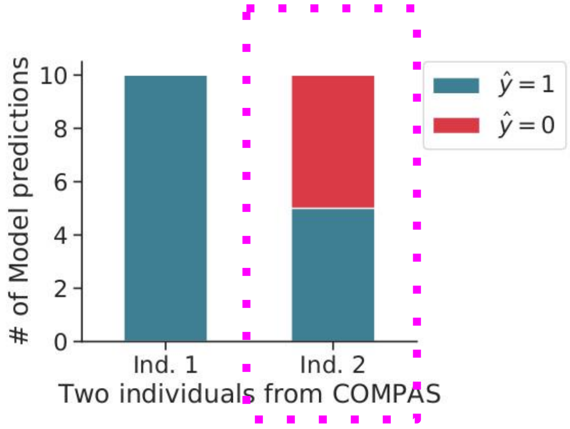
Ind. 1: $B_0 = 0$, $B_1 = 10$
Ind. 2: $B_0 = 5$, $B_1 = 5$

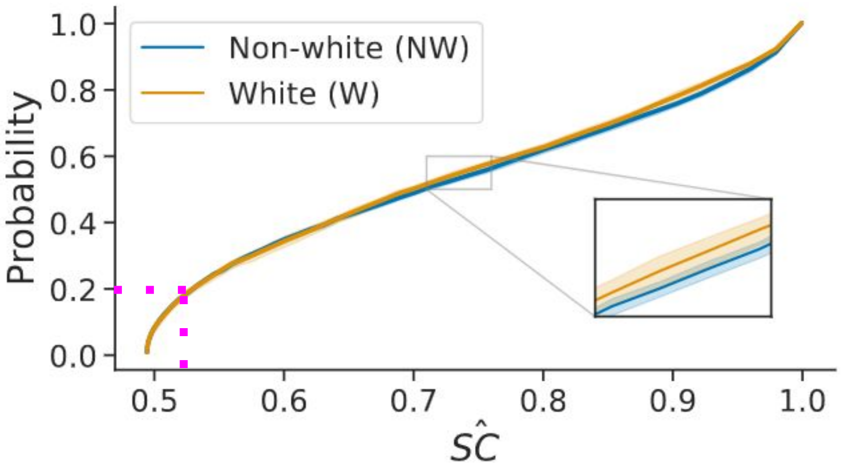
Illustrating Self-consistency



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

Illustrating Self-consistency

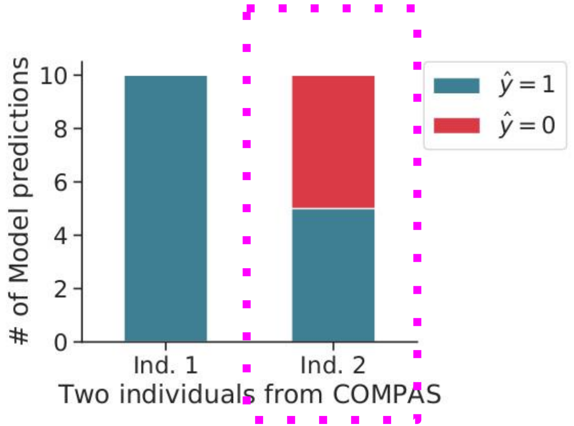


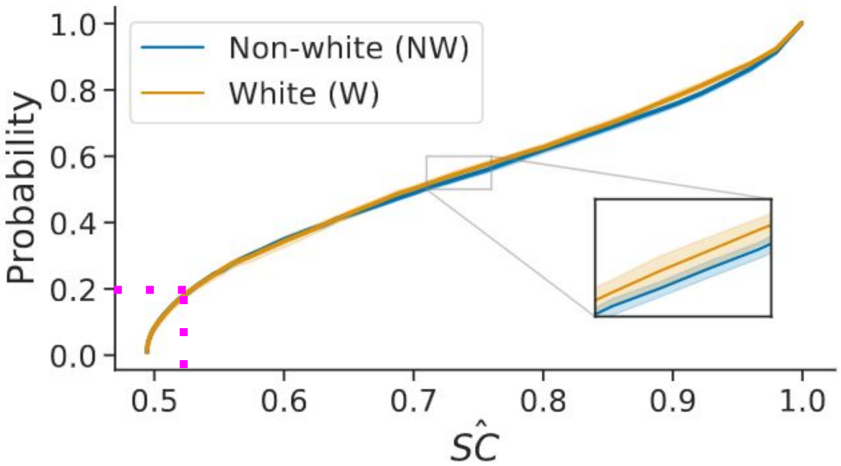


About 20% of the COMPAS test set looks approximately like Ind. 2

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

Illustrating Self-consistency



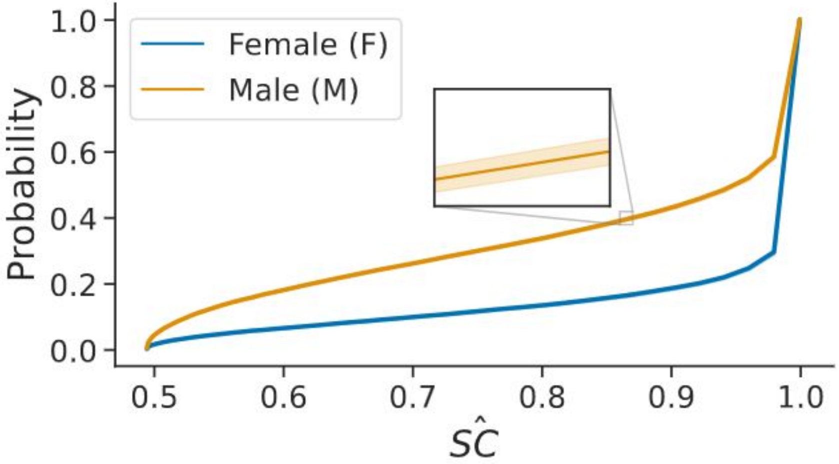


About 20% of the COMPAS test set
looks approximately like Ind. 2

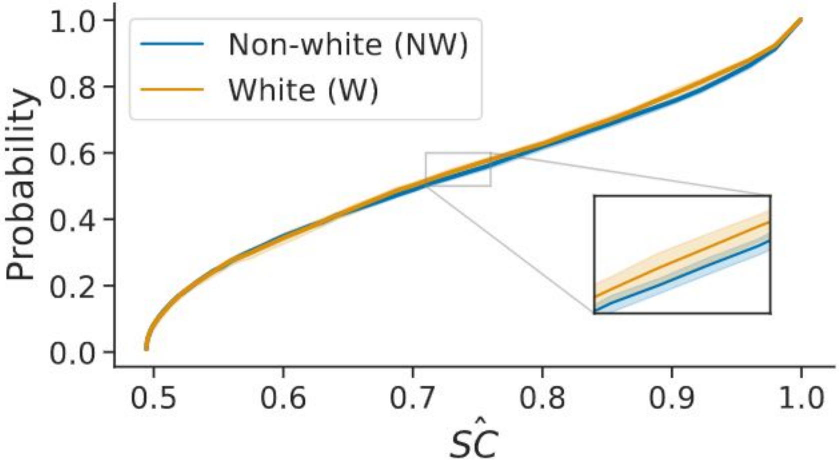
Their predictions are ***arbitrary***

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

Illustrating Self-consistency

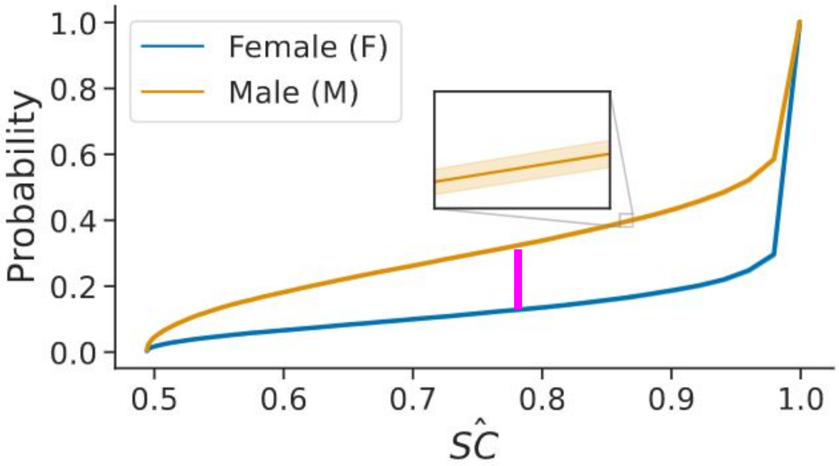


Old Adult, random forests, *B*=101
(mean +/- STD over 10 trials)

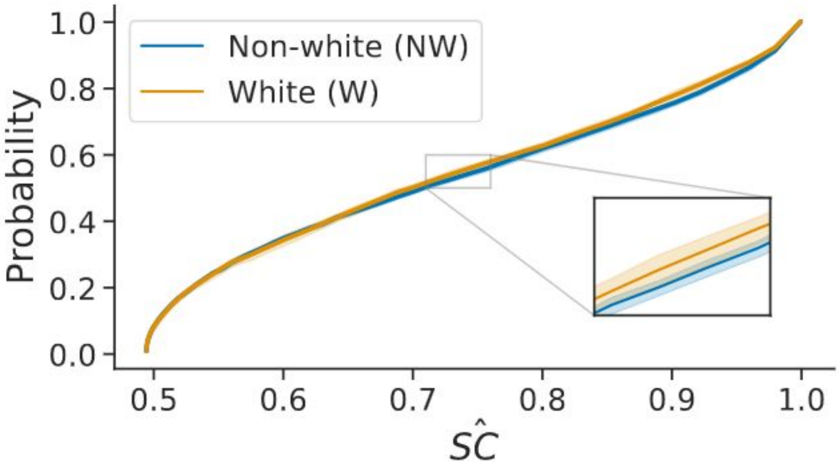


COMPAS, random forests, *B*=101
(mean +/- STD over 10 trials)

Illustrating Self-consistency



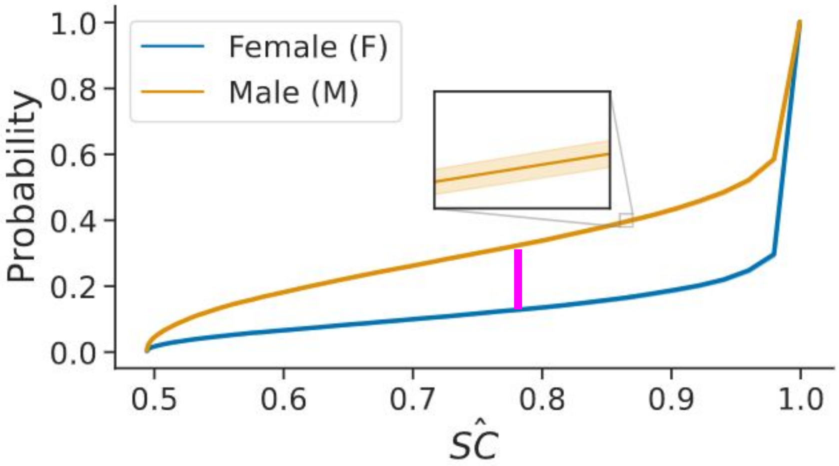
Old Adult, random forests, *B*=101
(mean +/- STD over 10 trials)



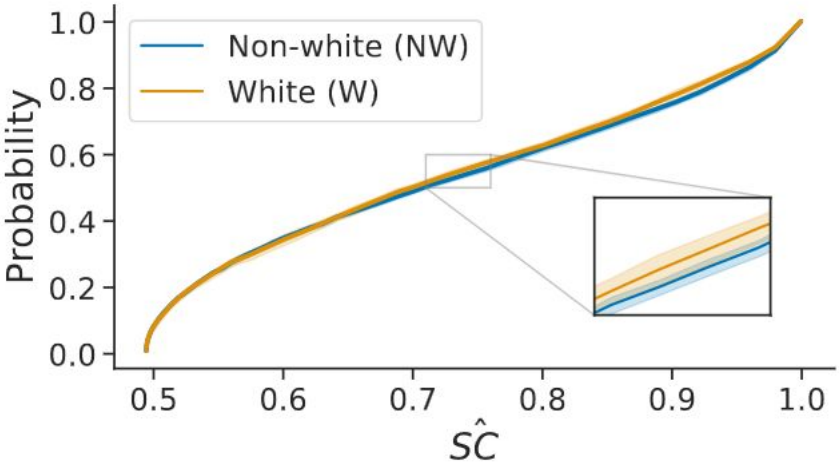
COMPAS, random forests, *B*=101
(mean +/- STD over 10 trials)

systematic arbitrariness

Illustrating Self-consistency



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



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

systematic arbitrariness
(actually happens rarely in practice)

Our contributions

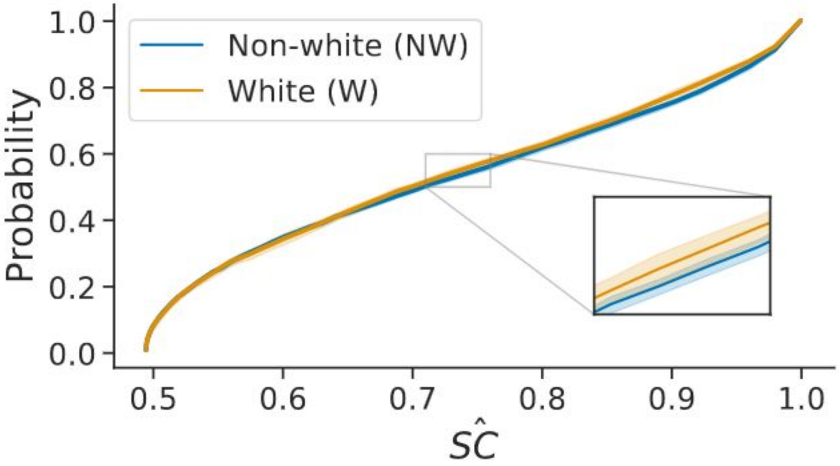
Quantifying ***arbitrariness*** via ***self-consistency***

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

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

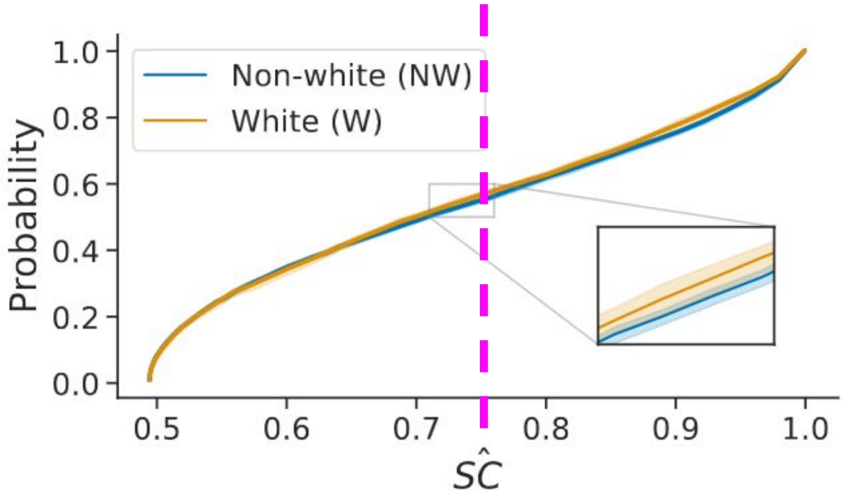
Packaging a large-scale dataset (won't get into this, but at the end will explain why)

An algorithm: Improving self-consistency



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

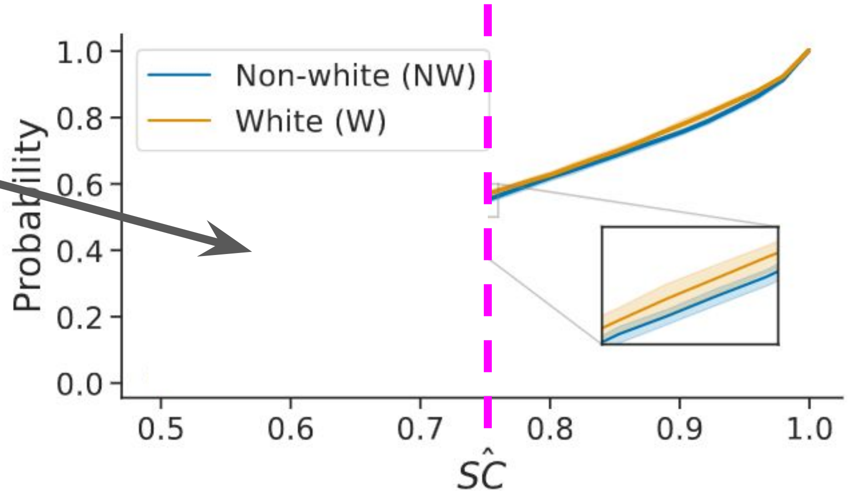
An algorithm: Improving self-consistency



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

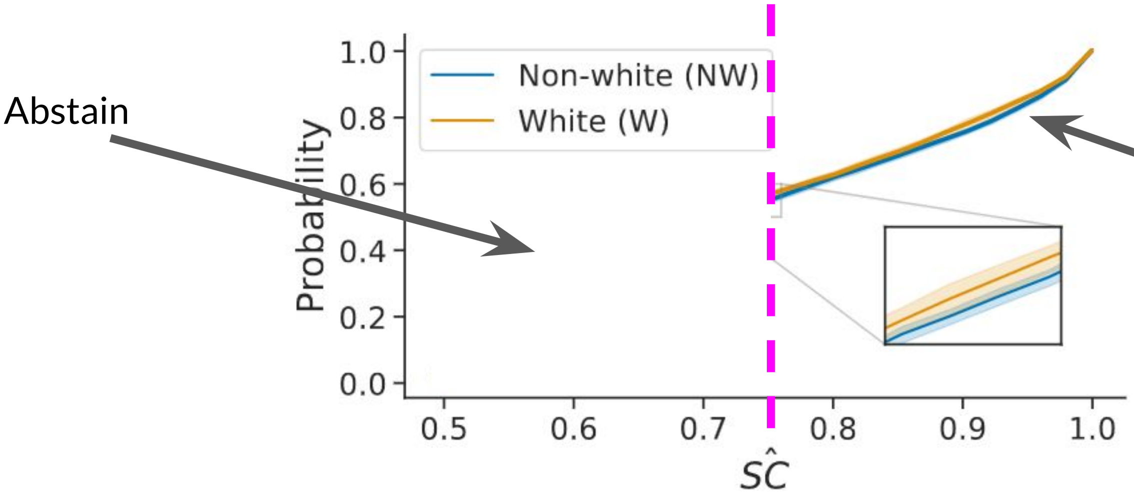
An algorithm: Improving self-consistency

Abstain



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

An algorithm: Improving self-consistency



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

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates
- For a particular data instance, the ensemble produces *B* predictions

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of B models from bootstrap replicates
- For a particular data instance, the ensemble produces B predictions
- We use these B predictions to compute self-consistency

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of B models from bootstrap replicates
- For a particular data instance, the ensemble produces B predictions
- We use these B predictions to compute self-consistency

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates
- For a particular data instance, the ensemble produces *B* predictions
- We use these *B* predictions to compute self-consistency
- A user selects a minimally-acceptable level of self-consistency (anything below this chosen level is deemed too arbitrary)

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates
- For a particular data instance, the ensemble produces *B* predictions
- We use these *B* predictions to compute self-consistency
- A user selects a minimally-acceptable level of self-consistency (anything below this chosen level is deemed too arbitrary)
- If self-consistency for a data instance is below the threshold, the algorithm **abstains from prediction** (otherwise, it predicts the majority vote label)

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates
- For a particular data instance, the ensemble produces *B* predictions
- We use these *B* predictions to compute self-consistency
- A user selects a minimally-acceptable level of self-consistency (anything below this chosen level is deemed too arbitrary)
- If self-consistency for a data instance is below the threshold, the algorithm **abstains from prediction** (otherwise, it predicts the majority vote label)

* We run two versions of this algorithm:
    simple ensembling (ensembles common model types in fair classification)

An algorithm: Improving self-consistency

Going to skip over the details but, in short:

- Our algorithm builds an ensemble of *B* models from bootstrap replicates
- For a particular data instance, the ensemble produces *B* predictions
- We use these *B* predictions to compute self-consistency
- A user selects a minimally-acceptable level of self-consistency (anything below this chosen level is deemed too arbitrary)
- If self-consistency for a data instance is below the threshold, the algorithm **abstains from prediction** (otherwise, it predicts the majority vote label)

* We run two versions of this algorithm:
  simple ensembling (ensembles common model types in fair classification)
  super ensembling (ensembles simple ensemble models, i.e., nested ensembles)

An algorithm: Improving self-consistency

This approach is **really simple**, and yet it yields

An algorithm: Improving self-consistency

This approach is **really simple**, and yet it yields

Really effective empirical results

Pretty shocking insights about the current state of fair classification research

An algorithm: Improving self-consistency

This approach is **really simple**, and yet it yields

Really effective empirical results

Pretty shocking insights about the current state of fair classification research

We are going to go through one example, looking at

How self-consistency changes

The effects on common error-based fairness metrics (since these are standard measurements in the field)

Our contributions

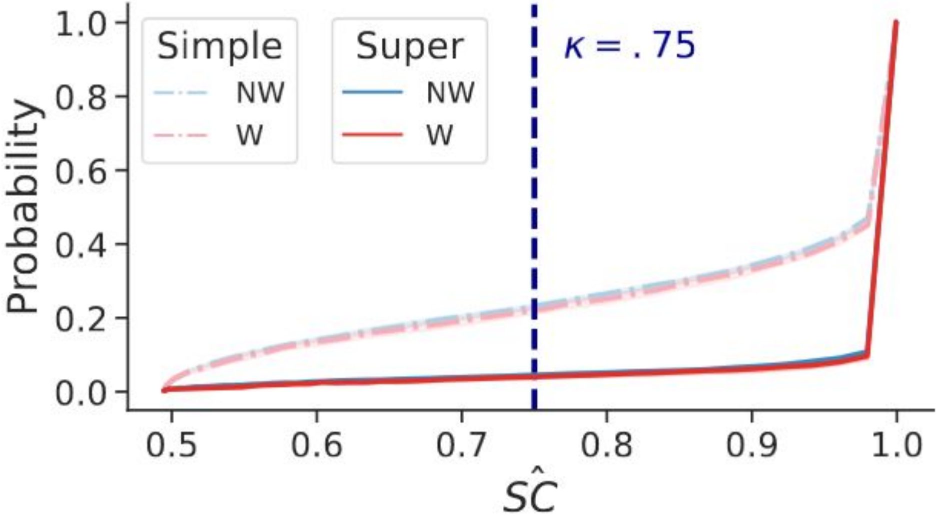
Quantifying ***arbitrariness*** via ***self-consistency***

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)

Evaluating our algorithm

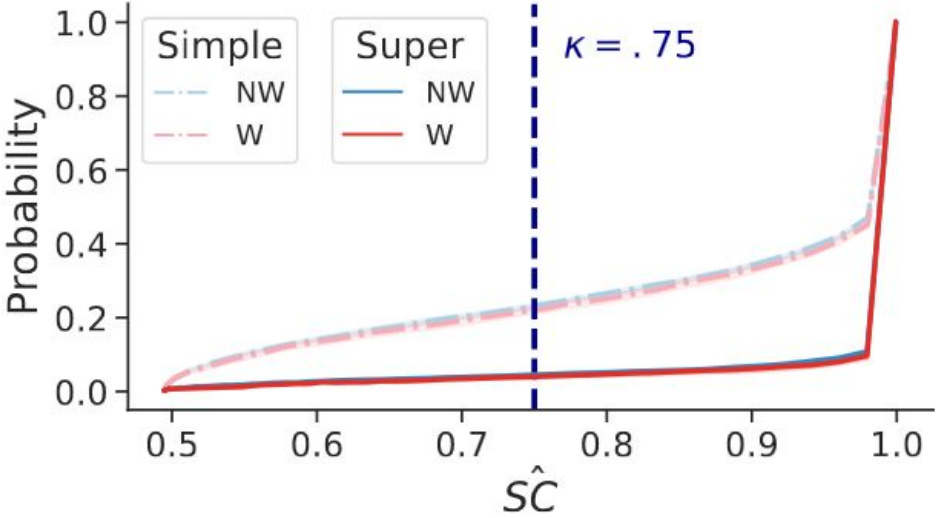


COMPAS, logistic regression, *B*=101
(mean +/- STD over 10 trials)

Evaluating our algorithm

Simple ensembling

Can abstain **a lot**



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

Evaluating our algorithm

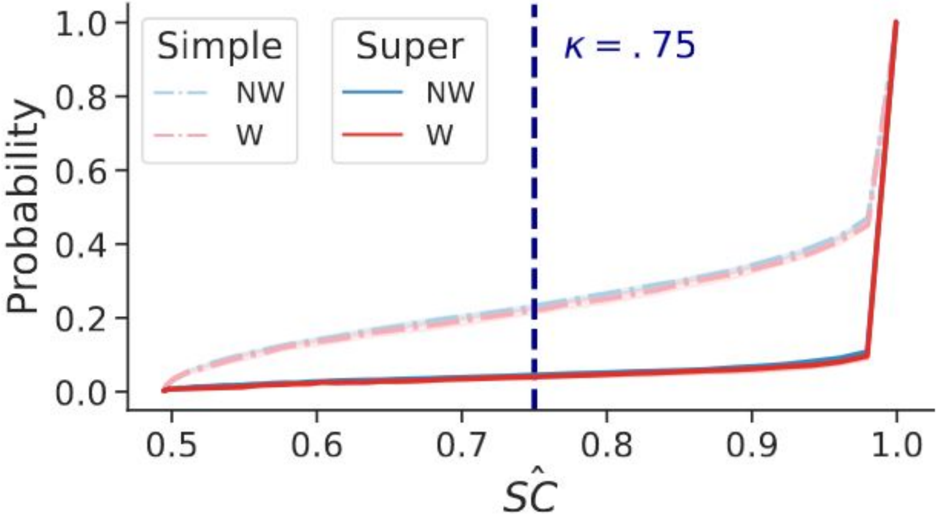
Simple ensembling

Can abstain **a lot**

Super ensembling

Brings down the curve → has higher self-consistency

Abstains **less**



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

Evaluating our algorithm

Simple ensembling

Can abstain **a lot**

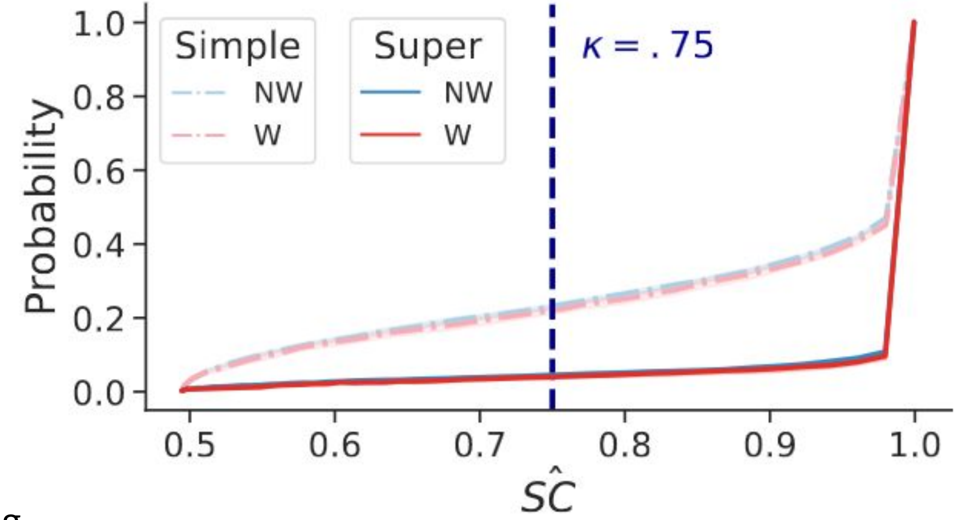
Super ensembling

Brings down the curve → has higher self-consistency

Abstains **less**

Both

Improve overall self-consistency by abstaining



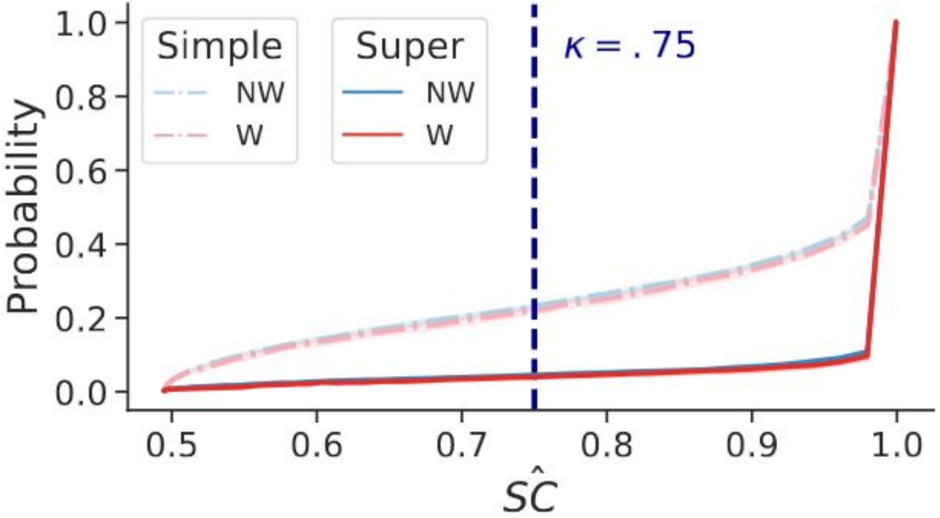
COMPAS, logistic regression, B=101
(mean +/- STD over 10 trials)

Evaluating our algorithm

Fairness metrics

Examine false positive rate disparities

	Baseline
$\Delta\hat{\text{FPR}}$	$2.1 \pm 0.0\%$
$\hat{\text{FPR}}_{\text{NW}}$	$14.7 \pm 1.3\%$
$\hat{\text{FPR}}_{\text{W}}$	$12.6 \pm 1.3\%$

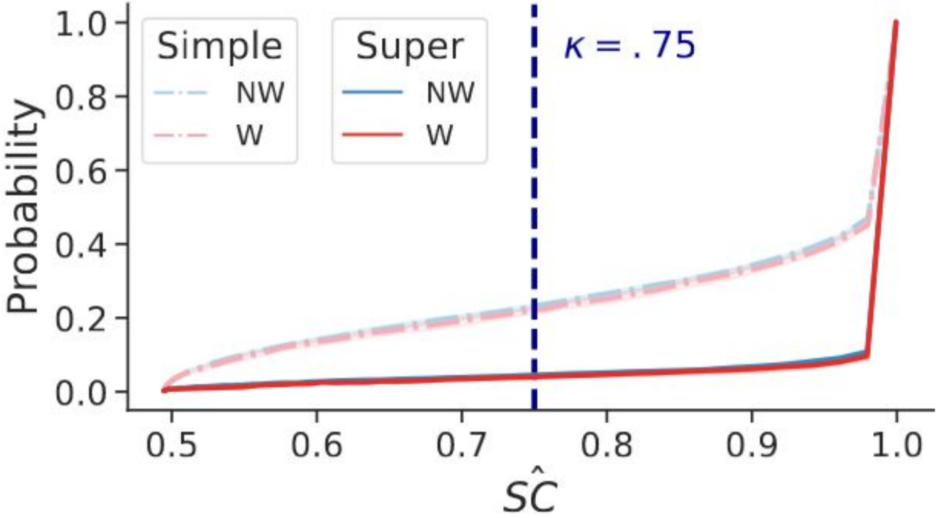


COMPAS, logistic regression, *B*=101
(mean +/- STD over 10 trials)

Evaluating our algorithm

Fairness metrics

Examine false positive rate disparities



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

	Baseline
$\Delta\hat{\mathbf{FPR}}$	$2.1 \pm 0.0\%$
$\hat{\mathbf{FPR}}_{\mathbf{NW}}$	$14.7 \pm 1.3\%$
$\hat{\mathbf{FPR}}_{\mathbf{W}}$	$12.6 \pm 1.3\%$

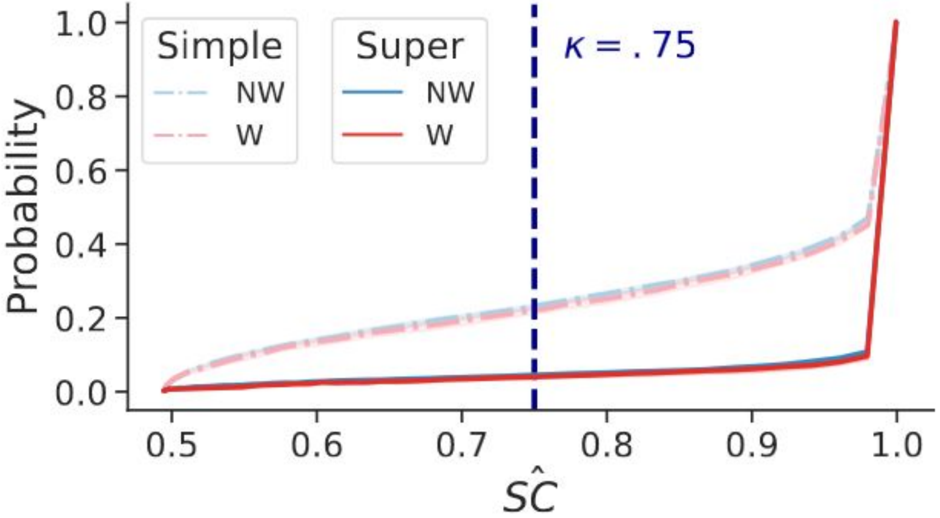
Expected error, which is not alone attainable by a single model (averages computed over underlying 1010 models)

Evaluating our algorithm

Fairness metrics

Examine false positive rate disparities

	Baseline	Simple
$\Delta \hat{\mathbf{FPR}}$	$2.1 \pm 0.0\%$	$3.0 \pm 0.0\%$
$\hat{\mathbf{FPR}}_{\mathrm{NW}}$	$14.7 \pm 1.3\%$	$11.4 \pm 1.0\%$
$\hat{\mathbf{FPR}}_{\mathrm{W}}$	$12.6 \pm 1.3\%$	$8.4 \pm 1.0\%$



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

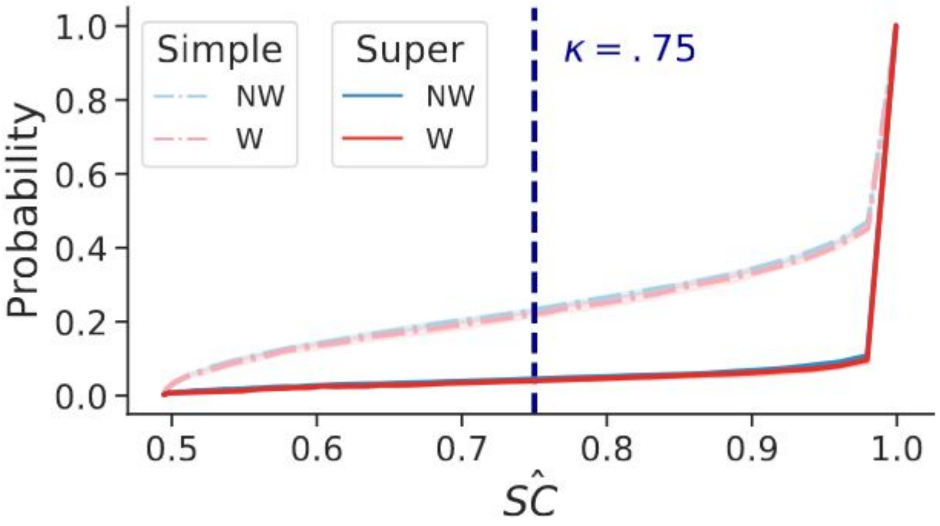
We are able to obtain this
result with simple ensemble models

Evaluating our algorithm

Fairness metrics

Examine false positive rate disparities

	Baseline	Simple	Super
$\Delta\hat{\mathbf{FPR}}$	$2.1 \pm 0.0\%$	$3.0 \pm 0.0\%$	$1.8 \pm .2\%$
$\hat{\mathbf{FPR}}_{\mathbf{NW}}$	$14.7 \pm 1.3\%$	$11.4 \pm 1.0\%$	$12.9 \pm .8\%$
$\hat{\mathbf{FPR}}_{\mathbf{W}}$	$12.6 \pm 1.3\%$	$8.4 \pm 1.0\%$	$11.1 \pm .6\%$



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

We are able to obtain this result with super ensemble models

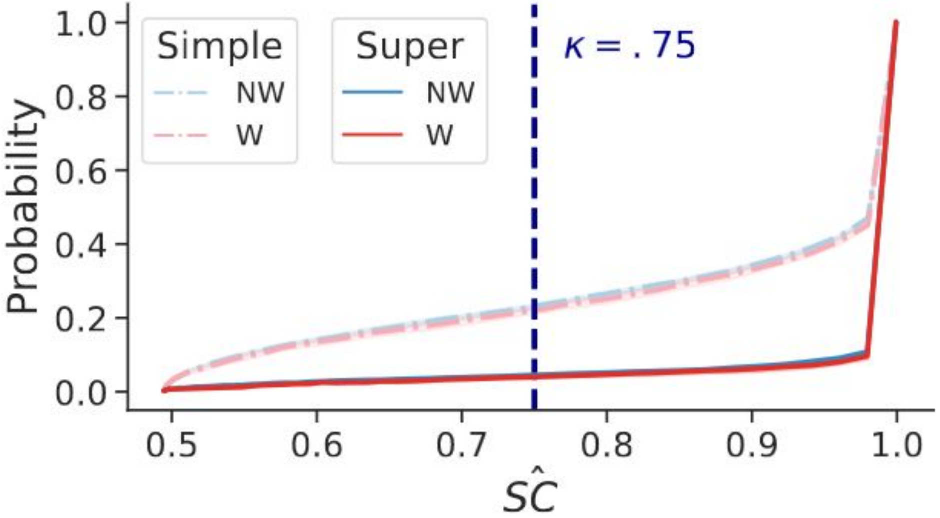
Evaluating our algorithm

Fairness metrics

Examine false positive rate disparities

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

	Baseline	Simple	Super
$\Delta\hat{\mathbf{FPR}}$	$2.1 \pm 0.0\%$	$3.0 \pm 0.0\%$	$1.8 \pm .2\%$
$\hat{\mathbf{FPR}}_{\mathbf{NW}}$	$14.7 \pm 1.3\%$	$11.4 \pm 1.0\%$	$12.9 \pm .8\%$
$\hat{\mathbf{FPR}}_{\mathbf{W}}$	$12.6 \pm 1.3\%$	$8.4 \pm 1.0\%$	$11.1 \pm .6\%$



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

Evaluating our algorithm

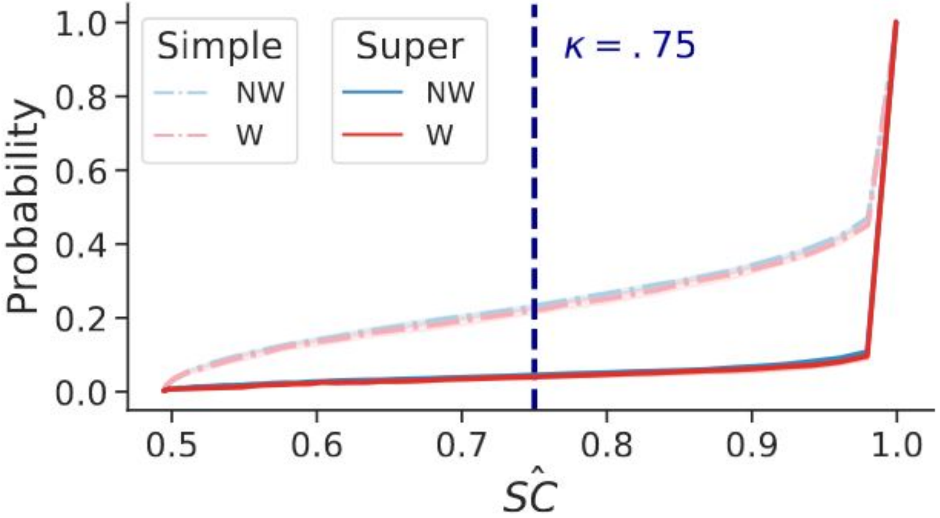
Fairness metrics

Examine false positive rate disparities

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

And we haven't run any algorithmic fairness method!

	Baseline	Simple	Super
$\Delta\hat{\text{FPR}}$	$2.1 \pm 0.0\%$	$3.0 \pm 0.0\%$	$1.8 \pm .2\%$
$\hat{\text{FPR}}_{\text{NW}}$	$14.7 \pm 1.3\%$	$11.4 \pm 1.0\%$	$12.9 \pm .8\%$
$\hat{\text{FPR}}_{\text{W}}$	$12.6 \pm 1.3\%$	$8.4 \pm 1.0\%$	$11.1 \pm .6\%$



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

Summarizing our study

Datasets:

- (South) German Credit
- COMPAS
- Old Adult

Summarizing our study

Datasets:

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

Summarizing our study

Datasets:

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

Summarizing our study

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 packaged this because we struggled to find algorithmic unfairness above

Summarizing our study

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 packaged this because we struggled to find algorithmic unfairness above

Models:

- Logistic regression
- Decision trees
- Random forests
- MLPs
- SVMs

These are the most common fair classification models

Summarizing our study

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 packaged this because we struggled to find algorithmic unfairness above

Models:
- Logistic regression
- Decision trees
- Random forests
- MLPs
- SVMs

These are the most common fair classification models

Summarizing our study

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 packaged this because we struggled to find algorithmic unfairness above

Models:
- Logistic regression
- Decision trees
- Random forests
- MLPs
- SVMs

These are the most common fair classification models

Overall, these patterns hold (and more)

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

Summarizing our study

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 packaged this because we struggled to find algorithmic unfairness above

Models:

- Logistic regression
- Decision trees
- Random forests
- MLPs
- SVMs

These are the most common fair classification models

Overall, these patterns hold (and more)

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

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

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

This finding is *really* shocking

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

This finding is *really* shocking

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

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

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: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

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.

Takeaways: Reproducibility and empirical rigor

We ran **hundreds of millions** of models using our algorithm of bootstrapping, aggregating, and abstaining using ensembles

In nearly every single case: **Models are close-to-fair without doing anything to target (un)fairness**

Bootstrapping 101 models (rather than cross-validating 5) yields models that better estimate expected error — **and they also happen to be close-to-fair**

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.

Our results that much theory work in the field misses this point. Rigorous empirics cast doubt on the practical usefulness of prior theoretical formulation choices