Practical Individual Fairness

Mikhail Yurochkin
Machine Bias
There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016
2019 study of 13.2 million mortgage and refinancing applications

<table>
<thead>
<tr>
<th>BLACK &amp; LATINX</th>
<th>EVERYONE ELSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>61% rejection rates</td>
<td>48%</td>
</tr>
<tr>
<td>+5.3 basis points (bps) on mortgage interest rates for fintech lending</td>
<td></td>
</tr>
<tr>
<td>+7.9 bps on overall mortgage interest rates</td>
<td></td>
</tr>
</tbody>
</table>
Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that’s where the similarities end.
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric
What is a fair algorithm?

Group Fairness:

Algorithm is equally good on groups of individuals
What is a fair algorithm?

Group Fairness:

Algorithm is equally good on groups of individuals
What is a fair algorithm?

**Group Fairness:**

Algorithm is equally good on groups of individuals
What is a fair algorithm?

**Group Fairness:**

*Algorithm is equally good on groups of individuals*

- $Y$ – true label
- $A$ – protected attribute
- $\hat{Y}$ – prediction

**Statistical Parity:**

$\hat{Y} \perp A$

**Equalized Odds:**

$\hat{Y} | Y \perp A$

Most prior work is on group fairness
What is a fair algorithm?

**Individual Fairness:**

*Algorithm treats similar individuals similarly*

Ryan earns 50k and lives in a predominantly white zip-code

Tyree earns 50k and lives in a predominantly black zip-code
What is a fair algorithm?

**Individual Fairness:**

*Algorithm treats similar individuals similarly*

Ryan earns 50k and lives in a predominantly white zip-code

Tyree earns 50k and lives in a predominantly black zip-code
What is a fair algorithm?

**Individual Fairness:**

*Algorithm treats similar individuals similarly*

Ryan earns 50k and lives in a predominantly white zip-code

Tyree earns 50k and lives in a predominantly black zip-code
What is a fair algorithm?

**Individual Fairness (Dwork et al. 2012):**

\[ d_Y(h(x_1), h(x_2)) \leq d_X(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X} \]

- ML model is a map \( h : \mathcal{X} \rightarrow \mathcal{Y} \)

- **fair metric** \( d_X \) measures similarity between inputs

- \( d_Y \) measures similarity between outputs
Example: Sentiment analysis – classify words as positive or negative

Positive: admire, adorable, joy, lucky, talented, ...

Negative: aggressive, distrust, nasty, radical, ...

Why Individual Fairness?
Why Individual Fairness?

Example: Sentiment analysis – classify words as positive or negative

Positive: admire, adorable, joy, lucky, talented, ...

Negative: aggressive, distrust, nasty, radical, ...

Deep Learning + Word Embeddings -> 95% test accuracy.
Why Individual Fairness?

What is a sentiment of a name?

European-American names: Adam, Ryan, Paul, ... , Courtney, Meredith, Megan, ...

African-American names: Alonzo, Leroy, Tyree, ... , Shereen, Sharise, Tawanda, ...
Why Individual Fairness?

![Box plot showing sentiment scores for different protected attributes. The plot compares sentiment scores for 'Race: Caucasian' and 'Race: African-American' across baseline and different fairness methods. The diagram highlights differences in sentiment scores between the races, indicating potential unfairness issues.]
Why Individual Fairness?

Baseline

GrRuS Fairness

SenSR

−20
−10
0
10
20
sentiPent sFRre

PrRteFteG Attribute

Protected Attribute

Race: Caucasian
Race: African-American

sentiment score

Baseline

Group Fairness
Why Individual Fairness?

Baseline

GrRuS Fairness

SenSR

−20

−10

0

10

20

sentiPent sFRre

PrRteFteG Attribute

Protected Attribute

Race: Caucasian

Race: African-American

Ryan

Tyree

sentiment score

Baseline

Group Fairness
Why Individual Fairness?

Baseline
GrRuS Fairness
SenSR

−20
−10
0
10
20

sentiPent sFRre

Protected Attribute
Race: Caucasian
Race: African-American

Individual Fairness
Why Individual Fairness?

Baseline
GrRuS Fairness
SenSR

−20
−10
0
10
20
sentiPent
sFRre

PrRteFteG Attribute
RaFe: CauFasian
RaFe: AfriFan-APeriFan

95%
94%

Protected Attribute
Race: Caucasian
Race: African-American

sentiment score

Baseline
Group Fairness
Individual Fairness

95%
94%
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
  - Enforcing DIF
  - Subgroup Fairness
  - Learning the Fair Metric
Assessing Individual Fairness

$$\text{IF}(h, x_i) \triangleq \left\{ \begin{array}{l}
\max_{x'_i \in \mathcal{X}} \quad d_Y(h(x_i), h(x'_i)) \\
\text{subject to} \quad d_X(x_i, x'_i) \leq \epsilon
\end{array} \right\}$$

- $d_X$: fair metric
- $d_Y$: metric on outputs
- $\epsilon$: small tolerance parameter
Distributional Individual Fairness (DIF)

\[ \text{DIF}(h) \triangleq \left\{ \sup_{T:x \rightarrow x} \mathbb{E}_{P_X} [d_Y(h(x), h(T(x)))] \right\} \]

subject to \( \mathbb{E}_{P_X} [d_X(x, T(x))] \leq \epsilon. \)

- \( P_X \) (marginal) distribution of inputs
- Optimal \( T \) maps \( x_i \) to \( x'_i \)
- Constraint enforces \( d_X(x_i, x'_i) \leq \epsilon \) on average
Comparing DIF and IF

Individual fairness map is feasible, but may not be optimal:

\[ T_{IF}(x_i) \triangleq \arg \max \ d_Y(h(x_i), h(x'_i)) \]
\[ \text{subject to } d_X(x_i, x'_i) \leq \epsilon \]

DIF may transport some points by more than \( \epsilon \).

IF does not imply DIF or vice a versa!
Comparing DIF and IF

Individual fairness map is feasible, but may not be optimal:

\[ T_{IF}(x_i) \triangleq \arg \max_{d\mathcal{X}(x_i, x'_i) \leq \epsilon} d\mathcal{Y}(h(x_i), h(x'_i)). \]

Let \( DIF(h) < \delta \), then

\[ P_X(d\mathcal{Y}(h(x), h(T_{IF}(x))) \geq \tau) \leq \frac{\delta}{\tau} \text{ for any } \tau > 0. \]

**DIF implies IF with high probability**
DIF in Social Science

\[
\text{DIF}(h) \triangleq \left\{ \sup_{T : \mathcal{X} \to \mathcal{X}} \mathbb{E}_{P_X} [d_Y(h(x), h(T(x)))] \right. \\
\text{subject to } \left. \mathbb{E}_{P_X} [d_{\mathcal{X}}(x, T(x))] \leq \epsilon. \right\}
\]

- \( P_X \) (marginal) distribution of inputs
- Optimal \( T \) maps \( x_i \) to \( x'_i \)
- Large \( \text{DIF}(h) \) implies unfairness
DIF in Social Science


- $P_X$ (marginal) distribution of inputs
- Optimal $T$ maps $x_i$ to $x'_i$
- Large DIF($h$) implies unfairness

- The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
- Optimal $T$ maps $x_i$ to $x'_i$.
- Large $\text{DIF}(h)$ implies unfairness.

- The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
- They randomly assigned African-American or white sounding names to the resumes.
- Large DIF(h) implies unfairness.
DIF in Social Science


- The investigators responded to job ads in Boston and Chicago newspapers with fictitious resumes.
- They randomly assigned African-American or white sounding names to the resumes.
- The investigators concluded there is discrimination against African-Americans: the resumes assigned white names received 50% more callbacks for interviews.
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric
Training DIF models

Usual AI: train algorithm accurate on the available data

- Observe data
- Update parameters to minimize prediction error
- Repeat
Individually Fair AI: train algorithm accurate on the available data and all possible similar data

- Observe data
- Generate similar data where algorithm performs differently to evaluate DIF
- Update parameters to minimize prediction error and DIF
- Repeat

Training DIF models
Training DIF models

\[ \min_{h \in \mathcal{H}} L(h) + \rho \text{DIF}(h) \]

\[ L(h) \triangleq \mathbb{E}[\ell(y, h(x))] \]

- \( \mathcal{H} \): model class (e.g. neural nets with a certain architecture)
- \( \ell \): \( \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \) is a loss function
- \( \rho \): regularization parameter
Training DIF models

\[
\text{DIF}(h) = \inf_{\lambda > 0} \{ \lambda \epsilon + \mathbb{E}_{P_X} [r_\lambda(h, x)] \}
\]

\[
r_\lambda(h, x) \triangleq \sup_{x' \in \mathcal{X}} \{ d_Y(h(x), h(x')) - \lambda d_X(x, x') \}
\]

DIF is infinite-dimensional, but its dual is univariate
Training DIF models

$$\min_{\theta \in \Theta, \lambda \in \mathbb{R}^+} \mathbb{E}[f(\theta, (x, y))]$$

$$f(\theta, (x, y)) \triangleq \ell(y, h_\theta(x)) + \rho(\lambda \epsilon + r_\lambda(h_\theta, x))$$

DIF is amendable to stochastic optimization
• Points on horizontal lines (identical $y$-values) are similar
• Training data is biased: $P_{Y|X}$ not constant on horizontal lines
• Use DIF regularizer to correct bias in training data
Achieving IF on toxicity classification

- Data: “Toxic Comment Classification Challenge”. 165k text comments
- Task: predict if a comment is toxic or not
- Measuring IF: does prediction change when changing identity tokens. Are predictions for “Some people are gay” and “Some people are straight” the same?
- 50 identity tokens: prediction on all variations should be the same to satisfy IF
Achieving IF on occupation prediction

- Data: Bias in Bios (De-Arteaga et al. 2019). 400k textual bio descriptions
- Task: predict one of the 28 occupations from a bio
- Measuring IF: does prediction change when changing names and gender pronouns in a bio?
Post-processing for Individual Fairness

- Works with any trained model without re-training
- Fast and easy to implement
- Does not require knowledge of the model parameters
- Only needs access to the outputs
- **Key idea**: represent individuals as a graph and quantify fairness with the graph Laplacian quadratic form
Post-processing for Individual Fairness
Post-processing for Individual Fairness

\[ \arg \min_f \left\{ \left\| f - \hat{y} \right\|_2^2 + \lambda f^\top \mathbf{L} f \right\} \]

Original predictions

Post-processed fair predictions

Stay close to original predictions

Penalize individual fairness violations

Dave
Charlie

Eve
Bob
Alice

JS

Stay close to original predictions
Penalize individual fairness violations
Post-processing for Individual Fairness

\[ \text{arg min}_f \left\| f - \hat{y} \right\|^2_2 + \lambda f^\top L f \]

Original predictions
Graph Laplacian

Stay close to original predictions
Penalize individual fairness violations

Closed-form solution!

\[ f = (I + \lambda L)^{-1} \hat{y} \]
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric
The @AppleCard is such a sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple’s black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

1:34 PM · Nov 7, 2019 · Twitter for iPhone

9K Retweets   3.5K Quote Tweets   28K Likes
Fairness problems Today
What is in-between?
What is in-between?

Subgroup fairness violation
What is in-between?

Enforcing Subgroup fairness with Group fairness requires expert knowledge to define all subgroups.
Subgroup fairness can be enforced with Individual fairness generalizing to unforeseen subgroups.
Example: Income prediction

Predict individuals' income based on Census data (ADULT dataset).

Evaluate SenSR against baseline and Adversarial Debiasing (group fairness).

Spousal consistency measures counterfactual sensitivity to “relationship” status, i.e. how often the classification remains unchanged when relationship status is perturbed.

SenSR + EXPLORE is best for S-Con; the group fairness method is worse than baseline.

*SenSR is our earlier algorithm for enforcing individual fairness.
Announcement 🚨

inFairness Python package coming out next year

```python
from inFairness.default_cfg import get_cfg_defaults
from inFairness.trainer import DefaultTrainer

cfg = get_cfg_defaults()
cfg.merge_from_list(["MODEL.SENSEI_RHO", 5.0])
cfg.freeze()
print(cfg)
trainer50 = DefaultTrainer(cfg)
trainer50.train()
```
Announcement

Demonstrations and interactive visualization tools

- Age 32 black male working a full time job

![Graphs showing fair and unfair predictions based on race, sex, and age.](image)
Roadmap

- AI is prone to biases
- What is a fair algorithm
- Distributional Individual Fairness (DIF)
- Enforcing DIF
- Subgroup Fairness
- Learning the Fair Metric
What are similar individuals?

Ryan earns 50k and lives in a predominantly white zip-code area

Tyree earns 50k and lives in a predominantly black zip-code area

Problem:
- Requires significant subject expertise to cover all possibilities
- Time consuming
- Very hard on some data types (images, customer records, ...)
Fair Metric Learning

\[
d_X(x_1, x_2) = (x_1 - x_2) \top \Sigma (x_1 - x_2)
\]

Option 1 (EXPLORE)

**Input:** pairs of comparable \((y = 1)\) and incomparable \((y = 0)\) samples \(\{x_{i1}, x_{i2}, y_i\}_{i=1}^n\)

Fit a model to find \(\Sigma\) such that: \(P(y = 1|x_1, x_2) \propto \frac{1}{1 + e^{d_X(x_1, x_2)}}\)

Example: Loan applicants with similar income such as Ryan and Tyree
Fair Metric Learning

\[ d_{\mathcal{X}}(x_1, x_2) = (x_1 - x_2)^\top \Sigma (x_1 - x_2) \]

Option 2 (Sensitive subspace)

**Input:** group (or groups) of comparable samples

Find directions of major variation with PCA, i.e. \( V = \{v_1, \ldots, v_K\} \).

Ignore them in the fair metric: \( \Sigma = I - P_{\text{span}(V)} \).

Example: Word embeddings of popular baby names in the sentiment classification experiment
References


Blog-posts

- SenSR: the first practical algorithm for individual fairness.  
  https://mitibmwatsonailab.mit.edu/research/blog/training-individually-fair-ml-models-with-sensitive-subspace-robustness

- Black Loans Matter: Fighting Bias for AI Fairness in Lending.  
  https://mitibmwatsonailab.mit.edu/research/blog/black-loans-matter-fighting-bias-for-ai-fairness-in-lending

- New research helps make AI fairer in decision-making.  
  https://www.research.ibm.com/blog/make-ai-fairer

Paper links, videos, news, and code are on my website moonfolk.github.io
Collaborators

University of Michigan: Yuekai Sun, Amanda Bower, Songkai Xue, Debarghya Mukherjee, Moulinath Banerjee, Alexander Vargo, Fan Zhang, Subha Maity, Hamid Eftekhar

University of Konstanz: Felix Petersen

IBM: Mark Weber, Ben Hoover, Hendrik Strobelt, Mayank Agarwal, Aldo Pareja, Onkar Bhardwaj, Ioana Baldini

Wells Fargo: Sherif Botros, Vanio Markov
Thank You!