An Intersectional Definition of Fairness

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, Shimei Pan
Department of Information Systems
University of Maryland, Baltimore County
ICDE 2020 Conference, Remote Presentation, April 2020

Work sponsored in part by NIST and NSF
Fairness in Machine Learning

- There is growing awareness that **biases inherent in data** can lead the behavior of machine learning algorithms to **discriminate against certain populations**.

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President
May 2016
Bias in Criminal Justice Risk Assessments

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), algorithm for risk assessment (Northpointe company)
  - Used for bail and sentencing decisions across the U.S.

ProPublica study (Angwin et al., 2016):
COMPAS almost twice as likely to incorrectly predict re-offending for African Americans than for white people.

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
Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants’ resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.
Typical Philosophical Assumption for AI Fairness

• **Infra-marginality:**
  
  attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

Typical Philosophical Assumption for AI Fairness

- **Infra-marginality:** attributes used by algorithm may have different distributions, depending on the protected attributes.

Typical Philosophical Assumption for AI Fairness

- **Infra-marginality:** attributes used by algorithm may have different distributions, depending on the protected attributes.

Typical Philosophical Assumption for AI Fairness

- **Infra-marginality:** attributes used by algorithm may have **different distributions**, depending on the **protected attributes**.

**Causal assumption**

- Protected attributes
- Features
- Outcomes

**Ideal World**

- Algorithm should behave differently for each group
- Individuals should get outcomes according to their “merit” or “risk”
- Algorithm is only biased if more inequitable than the data suggest

---

Fairness and Intersectionality

• **Intersectionality:**

  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**

  – gender, race, nationality, sexual orientation, disability status, socioeconomic class, ...

Fairness and Intersectionality

- **Intersectionality:** systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  - gender, race, nationality, sexual orientation, disability status, socioeconomic class, ...

  **Causal assumption**

Fairness and Intersectionality

• **Intersectionality:**
  systems of oppression built into society lead to **systematic disadvantages** along **intersecting dimensions**
  – gender, race, nationality, sexual orientation, disability status, socioeconomic class, ...

Fairness and Intersectionality

• **Intersectionality:**
systems of oppression built into society lead to **systematic disadvantages along intersecting dimensions**
  – gender, race, nationality, sexual orientation, disability status, socioeconomic class, ...

**Causal assumption**

**Ideal World**

• Rectify harmful effects of oppression

• Algorithm should **not** generally behave differently for each group

(unless justified, e.g. confounder variables)

Fairness and Intersectionality

We argue that an *intersectional definition of fairness* should satisfy:

- **Multiple protected attributes** should be considered

- **All** of the *intersecting values* of the protected attributes, e.g. *black women*, should be protected
  - We should still ensure that the individual protected attributes are protected overall, e.g. *women* are protected

- Systematic differences, due to structural oppression, are *rectified, rather than codified*.

- Protects *minority groups*

Our contributions

• We address fairness in machine learning from an **intersectional perspective**
  
  – Fairness definitions that **respect intersectionality**
    • Also provide a more politically conservative option
  
  – Theoretical results on our definitions’ properties
  
  – A learning algorithm to enforce our definitions
  
  – Experimental validation
Fairness and Intersectionality

- **Subgroup fairness** (Kearns et al., 2018)
  - Aims to prevent “fairness gerrymandering” a.k.a. *subset targeting*, by protecting specified subgroups
    
    **Definition 2.1** (Statistical Parity (SP) Subgroup Fairness). Fix any classifier $D$, distribution $\mathcal{P}$, collection of group indicators $\mathcal{G}$, and parameter $\gamma \in [0,1]$. For each $g \in \mathcal{G}$, define
    
    $$\alpha_{SP}(g,\mathcal{P}) = \mathbb{P}_D[g(x) = 1] \quad \text{and} \quad \beta_{SP}(g,D,\mathcal{P}) = |SP(D) - SP(D,g)|,$$

    where $SP(D) = \mathbb{P}_{D,X}[D(X) = 1]$ and $SP(D,g) = \mathbb{P}_{D,X}[D(X) = 1|g(x) = 1]$ denote the overall acceptance rate of $D$ and the acceptance rate of $D$ on group $g$ respectively. We say that $D$ satisfies $\gamma$-statistical parity (SP) Fairness with respect to $\mathcal{P}$ and $\mathcal{G}$ if for every $g \in \mathcal{G}$
    
    $$\alpha_{SP}(g,\mathcal{P})\beta_{SP}(g,D,\mathcal{P}) \leq \gamma.$$

  - punts on small groups (in order to prove generalization)

See also *multicalibration*, a similar definition but for calibration of probabilities (Hebert-Johnson et al., 2018)
Differential Fairness (DF)

We propose a fairness definition with the following properties:

- **Measures** the fairness cost of algorithms and data
  - Can measure difference in fairness between algorithms and data: *bias amplification*

- **Privacy** and economic guarantees
  - Privacy perspective provides an *interpretation* of definition, based on *differential privacy*

- Implements *intersectionality*: e.g. fairness for (gender, race) provably ensures fairness for gender and for race separately

Essentially, differential fairness extends the 80% rule to multiple protected attributes and outcomes, and provides a privacy interpretation.
Fairness and the Law: Adverse Impact Analysis

- Title VII, other anti-discrimination laws prohibit employers from intentional discrimination against employees with respect to protected characteristics
  - gender, race, color, national origin, religion

- Uniform Guidelines for Employee Selection Procedures (Equal Employment Opportunity Commission)
Fairness and the Law: 
Adverse Impact Analysis

Uniform guidelines: the “four-fifths rule” (a.k.a. 80% rule)

“A selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact,

while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.”

If so, there is evidence of adverse impact
Differential Privacy vs the 80% Rule

Definition: \( M(X) \) is \( \epsilon \)-differentially private if

\[
e^{-\epsilon} \leq \frac{Pr(M(X) \in S)}{Pr(M(X') \in S)} \leq e^{\epsilon}
\]

for all outcomes \( S \), and pairs of databases \( X, X' \) differing in a single element.

Follows from taking the reciprocal. We want ratios close to 1

• 80% rule: Evidence of unfairness if:

\[
\frac{Pr(\text{hire}|\text{group } A)}{Pr(\text{hire}|\text{group } B)} < 0.8
\]

The ratio determines the degree of disparate impact between groups. Like differential privacy, we want to bound a ratio to be somewhere near 1
Scenario for Differential Fairness

Individuals’ data

\[ x_i \sim \theta , \]
\[ \theta \in \Theta \]

Fair algorithm

\[ M(x) \]

Outcomes \( y_i \)

Vendor
(may be untrusted)

Multiple protected attributes

\[ a_i \in A \]

Randomness in data and mechanism

\[ P_{M,\theta}(M(x) = y|\theta) \]
Our Proposed Fairness Definition: Differential Fairness (DF)

A mechanism $M(X)$ is $\epsilon$-differentially fair in a framework $(A, \Theta)$ if for all $\theta \in \Theta$ with $X \sim \theta$, and $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(x) = y|s_i, \theta)}{P_{M,\theta}(M(x) = y|s_j, \theta)} \leq e^\epsilon,$$

for all $(s_i, s_j) \in A \times A$ where $P(s_i|\theta) > 0$, $P(s_j|\theta) > 0$.

Key idea: ratios of probabilities of outcomes bounded for any pair of values of protected attributes.
Interpreting $\varepsilon$: Bayesian Privacy

• Untrusted vendor/adversary can learn very little about the protected attributes of the instance, relative to their prior beliefs, assuming their prior beliefs are in $\Theta$:

$$e^{-\varepsilon} \frac{P(s_i | \theta)}{P(s_j | \theta)} \leq \frac{P(s_i | M(x) = y, \theta)}{P(s_j | M(x) = y, \theta)} \leq e^\varepsilon \frac{P(s_i | \theta)}{P(s_j | \theta)}$$

• E.g., if a loan was given to an individual, the vendor or adversary's Bayesian posterior beliefs about their race and gender will not be substantially changed

• This can prevent subsequent discrimination, e.g. in retaliation for a correction against bias.
Intersectionality Property of DF: Fairness with Multiple Protected Attributes

- **Intersectionality theory**: gender is not the only dimension upon which power structures in society impose systems of oppression and marginalization.
  - The intersection of a number of aspects must be considered, including race, sexual orientation, class, and disability status

**Theorem**: Let $M$ be an $\epsilon$-differentially fair mechanism in $(A, \Theta)$, $A = S_1 \times S_2 \times \ldots \times S_p$, and let $D = S_a \times \ldots \times S_k$ be the Cartesian product of a nonempty proper subset of the variables included in $A$. Then $M$ is $\epsilon$-differentially fair in $(D, \Theta)$.

E.g., if $M$ is differentially fair in (race, gender, nationality), it is differentially fair to a similar degree in gender alone

Other Theoretical Properties

• Generalization Guarantee

**Theorem** Fix a class of functions $\mathcal{H}$, which without loss of generality aim to discriminate the outcome $y = 1$ from any other value, denoted here as $y = 0$. For any conditional distribution $P(y, x|s)$ given a group $s$, let $S \sim P^m$ be a dataset consisting of $m$ examples $(x_i, y_i)$ sampled i.i.d. from $P(y, x|s)$. Then for any $0 < \delta < 1$, with probability $1 - \delta$, for every $h \in \mathcal{H}$, we have:

$$|P(y = 1|s, h) - P_S(y = 1|s, h)| \leq \tilde{O}\left(\sqrt{\text{VCDIM} (\mathcal{H}) \log m + \log (1/\delta)} \right).$$

• Economic guarantee

An $\epsilon$-differentially fair mechanism admits a disparity in expected utility of as much as a factor of $\exp(\epsilon) \approx 1 + \epsilon$ (for small values of $\epsilon$) between pairs of protected groups with $s_i \in A$, $s_j \in A$, for any utility function that could be chosen.

Protected groups have similar economic outcomes
Measuring Bias Amplification

• We can measure the extent to which an algorithm increases the bias over the original data

• Calculate differential fairness of data, $\epsilon_1$

• Calculate differential fairness of algorithm, $\epsilon_2$

• Bias amplification: $\epsilon_2 \rightarrow \epsilon_1$

This is a more politically conservative fairness definition: implements infra-marginality
Learning with DF Penalty

• Objective: \( \min_{W} [L_X(W) + \lambda R_X(\epsilon)] \)

\[ R_X(\epsilon) = \max(0, \epsilon_{M_W(x)} - \epsilon_1) \]

Fairness penalty term

Determines whether to penalize DF or DF-bias amplification

• Optimize via \textbf{gradient descent}: backprop + auto-diff (DF-Classifier)
• We use a similar algorithm to enforce subgroup fairness (SF-Classifier)
Learning Results

- Both algorithms improve both metrics, both per-group and overall
- DF-classifier improves fairness for minority groups, even under SF metric
Thank you!

• Contact information:
  – James Foulds
    Assistant professor
    Department of Information Systems
    UMBC
    Email: jfoulds@umbc.edu
    Webpage: http://jfoulds.informationsystems.umbc.edu

• An extended version of our work is online at arxiv.org:

• An accepted SDM 2020 paper on modeling uncertainty in estimating DF:
Bonus Slides
Subgroup Fairness and Intersectionality

Size of group, as a proportion of the population

Our metric does not down-weight small intersectional groups

Subgroup fairness down-weights small intersectional groups
Fairness and Privacy: the Untrusted Vendor

Individuals’ data $x_i$, including protected attribute(s) → Fair algorithm $M(x)$ → Outcomes $y$

The user of the algorithm’s outputs (the vendor) may discriminate, e.g. in retaliation for a fairness correction (Dwork et al., 2012)

Interlude: Differential Privacy (Dwork et al., 2006)

• DP is a promise:
  – “If you add your data to the database, you will not be affected much”

Privacy-preserving interface: randomized algorithms

Differential Fairness Example

Scenario: Given an applicant’s score on a standardized test, an applicant is hired if their test score is greater than a threshold $t$. Here, $t = 10.5$. Each group of applicant has a different distribution over scores:

Group 1: $N(X; \mu_1 = 10, \sigma = 1)$
Group 2: $N(X; \mu_2 = 12, \sigma = 1)$

M(X): Hire applicant if test score $> 10.5$
Differential Fairness Example

M(X): Hire applicant if test score > 10.5

<table>
<thead>
<tr>
<th>Probability of Hiring Outcome Given Group</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>yes</td>
<td>0.3085</td>
</tr>
<tr>
<td>no</td>
<td>0.6915</td>
</tr>
</tbody>
</table>

Log Ratios of Probabilities

| y   | $s_i$ | $s_j$ | $\log \frac{P_{M,\theta}(M(X)=y|s_i,\theta)}{P_{M,\theta}(M(X)=y|s_j,\theta)}$ |
|-----|-------|-------|------------------------------------------------------------------|
| no  | 1     | 2     | 2.337                                                            |
|     | 2     | 1     | -2.337                                                           |
| yes | 1     | 2     | -1.107                                                           |
|     | 2     | 1     | 1.107                                                            |

Find the worst case: $\epsilon = 2.337$
Measuring Bias in Data

• Can measure bias in a dataset

Special case of differential fairness, in which the algorithm is the data distribution

Empirical differential fairness (EDF) of a labeled dataset:

Corresponds to verifying that for any $y$, $s_i$, $s_j$, we have

$$e^{-\epsilon} \leq \frac{N_{y,s_i}}{N_{s_i}} \cdot \frac{N_{s_j}}{N_{y,s_j}} \leq e^{\epsilon}$$

• Also applies to a probabilistic model of the data
Learning Results

<table>
<thead>
<tr>
<th>Models</th>
<th>DF-Classifier</th>
<th>SF-Classifier</th>
<th>Typical Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.686</td>
<td>0.684</td>
<td>0.690</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.633</td>
<td>0.642</td>
<td>0.622</td>
</tr>
<tr>
<td>ROC AUC</td>
<td>0.730</td>
<td>0.723</td>
<td>0.719</td>
</tr>
<tr>
<td>(\epsilon = 0.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\epsilon = 0.2231)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = 0.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = \gamma_{data})</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>SF-Classifier</th>
<th>Typical Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\epsilon = \epsilon_{data})</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(\gamma = \gamma_{data})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fairness Measures (using soft counts)</th>
<th>DF-Classifier</th>
<th>SF-Classifier</th>
<th>Typical Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\epsilon = 0.0)</td>
<td>0.180</td>
<td>0.410</td>
<td>0.468</td>
</tr>
<tr>
<td>(\epsilon = 0.2231)</td>
<td>0.281</td>
<td>0.404</td>
<td>0.468</td>
</tr>
<tr>
<td>(\gamma = 0.0)</td>
<td>0.006</td>
<td>0.033</td>
<td>0.028</td>
</tr>
<tr>
<td>(\gamma = \gamma_{data})</td>
<td>0.007</td>
<td>0.007</td>
<td>0.014</td>
</tr>
<tr>
<td>Bias Amp-DF</td>
<td>-0.360</td>
<td>-0.130</td>
<td>-0.072</td>
</tr>
<tr>
<td>Bias Amp-SF</td>
<td>-0.015</td>
<td>-0.012</td>
<td>0.014</td>
</tr>
<tr>
<td>Bias Amp-DF</td>
<td>-0.025</td>
<td>0.020</td>
<td>0.008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fairness Measures (using hard counts)</th>
<th>DF-Classifier</th>
<th>SF-Classifier</th>
<th>Typical Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\epsilon = 0.0)</td>
<td>0.207</td>
<td>0.884</td>
<td>0.860</td>
</tr>
<tr>
<td>(\epsilon = 0.2231)</td>
<td>0.671</td>
<td>0.825</td>
<td>0.860</td>
</tr>
<tr>
<td>(\gamma = 0.0)</td>
<td>0.015</td>
<td>0.017</td>
<td>0.048</td>
</tr>
<tr>
<td>(\gamma = \gamma_{data})</td>
<td>0.045</td>
<td>0.017</td>
<td>0.048</td>
</tr>
<tr>
<td>Bias Amp-DF</td>
<td>-0.339</td>
<td>0.338</td>
<td>0.314</td>
</tr>
<tr>
<td>Bias Amp-SF</td>
<td>-0.025</td>
<td>0.020</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 3: Comparison of intersectionally fair classifiers with the typical classifier on the COMPAS dataset (\(\epsilon_1 = 0.2231\) is the 80% rule).

- **Little to no loss in accuracy metrics when trained to prevent bias amplification**
- **Differential fairness is protected or improved vs training data (“bias de-amplification”)**
Dealing With Confounders

• UC Berkeley admissions: Simpson’s paradox
  – “Department applied to” is a confounder
  – Demographic parity no longer ideal

• Solution:
  protect DF per department

  **DFC:** Differential fairness w/ confounders

• **Theorem:** overall admissions
  DF no worse than DFC
  (i.e. DF of the “worst” dept)
Proof of Intersectionality Theorem

Proof. Define $E = S_1 \times \ldots \times S_{a-1} \times S_{a+1} \ldots \times S_{k-1} \times S_{k+1} \times \ldots \times S_p$, the Cartesian product of the protected attributes included in $A$ but not in $D$. Then for any $\theta \in \Theta$, $y \in \text{Range}(M)$,

$$\log \max_{s \in D: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|D = s, \theta)$$

$$= \log \max_{s \in D: P(s|\theta) > 0} \sum_{e \in E} P_{M, \theta}(M(x) = y|E = e, s, \theta)P_{\theta}(E = e|s, \theta)$$

$$\leq \log \max_{s \in D: P(s|\theta) > 0} \sum_{e \in E} \max_{e' \in E: P_{\theta}(E = e'|s, \theta) > 0} (P_{M, \theta}(M(x) = y|E = e', s, \theta)) \times P_{\theta}(E = e|s, \theta)$$

$$= \log \max_{s \in D: P(s|\theta) > 0} \max_{e' \in E: P_{\theta}(E = e'|s, \theta) > 0} P_{M, \theta}(M(x) = y|E = e', s, \theta)$$

$$= \log \max_{s' \in A: P(s'|\theta) > 0} P_{M, \theta}(M(x) = y|s', \theta)$$

By a similar argument, $\log \min_{s \in D: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|D = s, \theta) \geq \log \min_{s' \in A: P(s'|\theta) > 0} P_{M, \theta}(M(x) = y|s', \theta)$. Applying Lemma 7.1, we hence bound $\epsilon$ in $(D, \Theta)$ as

$$\log \max_{s \in D: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|D = s, \theta)$$

$$- \log \min_{s \in D: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|D = s, \theta)$$

$$\leq \log \max_{s' \in A: P(s'|\theta) > 0} P_{M, \theta}(M(x) = y|s', \theta)$$

$$- \log \min_{s' \in A: P(s'|\theta) > 0} P_{M, \theta}(M(x) = y|s', \theta) \leq \epsilon .$$

\textbf{Lemma 7.1}
$e$-DF criterion can be rewritten as: for any $\theta \in \Theta$, $y \in \text{Range}(M)$,

$$\log \max_{s \in A: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|s, \theta)$$

$$- \log \min_{s \in A: P(s|\theta) > 0} P_{M, \theta}(M(x) = y|s, \theta) \leq \epsilon .$$
References


(9/16 female first authors, indicated in bold)
Conclusion

“There is a rise of big-data optimism here, and if ever there were a time when politicians, industry leaders, and academics were enamored with artificial intelligence as a superior approach to sense-making, it is now.

This should be a wake-up call for people living in the margins, and people aligned with them, to engage in thinking through the interventions we need.”