Bias and Fairness in AI/ML models

Swati Gupta
Assistant Professor
School of Industrial and Systems Engineering,
Georgia Institute of Technology

October 25, 2018
Digital Data Flows Master Class: Emerging Technologies
Machine Learning Pipeline

Data → Machine Learning/AI → Data driven decisions

What is the effect of these decisions on human well-being?
Bias and Fairness in AI/ML models | Swati Gupta | Georgia Institute of Technology

What is Bias/Fairness?

“Bias. When scientific or technological decisions are based on a narrow set of systemic, structural or social concepts and norms, the resulting technology can privilege certain groups and harm others.” – Nature comment

Four ethical priorities for neurotechnologies and AI

Rafael Yuste, Sara Goering, Blaise Agüera y Arcas, Guoqiang Bi, Jose M. Carmena, Adrian Carter, Joseph J. Fins, Phoebe Friesen, Jack Gallant, Jane E. Huggins, Judy Illes, Philipp Kellmeyer, Eran Klein, Adam Marblestone, Christine Mitchell, Erik Parens, Michelle Pham, Alan Rubel, Norihiro Sadato, Laura Specker Sullivan, Mina Teicher, David Wasserman, Anna Wexler, Meredith Whittaker & Jonathan Wolpaw

08 November 2017
What is Bias/Fairness?

“Bias. When scientific or **technological decisions** are based on a narrow set of systemic, structural or **social concepts and norms**, the resulting technology can **privilege certain groups** and harm others.” – Nature comment
Amazon to Bring Same-Day Delivery to Roxbury After Outcry

by Spencer Soper
April 26, 2016, 5:19 PM EDT  Updated on April 26, 2016, 8:22 PM EDT

Rating systems may discriminate against Uber drivers

By Leslie Morris | December 15, 2016

Bias and Fairness in AI/ML models | Swati Gupta | Georgia Institute of Technology
Amazon to Bring Same-Day Delivery to Roxbury After Outcry

by Spencer Soper
April 26, 2016, 5:19 PM EDT Updated on April 26, 2016, 8:22 PM EDT

CORNELL CHRONICLE

Rating systems may do more harm than good

PROPUBLICA

MACHINE BIAS

Facebook Lets Advertisers Exclude Users by Race

Facebook’s system allows advertisers to exclude black, Hispanic, and other “ethnic affinities” from seeing ads.

by Julia Angwin and Terry Parris Jr., Oct. 28, 2016, 1 p.m. EDT

Bias and Fairness in AI/ML models | Swati Gupta | Georgia Institute of Technology
Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people’s concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google’s ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google’s DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user’s behaviors.

Key background reading before our discussion on predictive policing on Wednesday, February 24th.
Outline of the talk

- Bias in the data, models and variables
- Fairness Metrics
  - Statistical measures
  - Equity measures
- Trolley Problem of Choice
**Predictive Policing**

“application of analytical techniques to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions”

![Heat map of drug arrests made](image)

**FIGURE 1** (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health
ML finds patterns in data
ML finds patterns in data
PredPol: crime type, time, loc

Heat map of drug arrests made

[Kristian Lum, William Isaac, 2016]
We live in a biased society, so it’s inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..
We live in a biased society, so it’s inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..
We live in a biased society, so it’s inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..

“We also found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male.”
Bias and Fairness in AI/ML models | Swati Gupta | Georgia Institute of Technology

Not just about collection

We live in a biased society, so it’s inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies.

We do not want such biases to propagate into systems that make life-changing decisions.

Proxies:
- predicting crime using data on arrests,
- Not on incidence of crime.
Outline of the talk

- Bias in the data, models and variables
- Fairness Metrics
  - Statistical measures
  - Equity measures
- Trolley Problem of Choice
Machine Learning Pipeline

What is the effect of these decisions on human well-being?
Classification

Hired for job or not, will re-offend or not (prison), given a loan or not.
Statistical Definitions of Fairness

Hired for job or not, will re-offend or not (prison), given a loan or not.

Is it fair to achieve highest accuracy in classification?
Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison)**, given a loan or not.

**YES**

**NO**

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

COMPAS Risk Score: ProPublica
Statistical Definitions of Fairness

Hired for job or not, will re-offend or not (prison), given a loan or not.

YES

NO

Prediction Fails Differently for Black Defendants

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Overall, Northpointe’s assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)
Hired for job or not, will re-offend or not (prison), given a loan or not.

Is it fair to achieve highest accuracy in classification?

Or is it fair to balance false positives across the groups?
Statistical Definitions of Fairness

Hired for job or not, will re-offend or not (prison), given a loan or not.

Yes

Is it fair to achieve highest accuracy in classification?

Or is it fair to balance false positives across the groups?

No
Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison)**, given a loan or not.

Is it **fair** to achieve highest accuracy in classification?

Or is it **fair** to balance **false positives** across the groups?

**False negatives?**
### Statistical Definitions of Fairness

<table>
<thead>
<tr>
<th>True condition</th>
<th>Prevalence</th>
<th>Accuracy (ACC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition positive</td>
<td>$\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$</td>
<td>$\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$</td>
</tr>
<tr>
<td>Condition negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Predicted condition positive</th>
<th>False positive, Type I error</th>
<th>Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$</th>
<th>False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted condition negative</td>
<td>False negative, Type II error</td>
<td>False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$</td>
<td>Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$</td>
</tr>
</tbody>
</table>

- **True positive** (TPR), Recall, Sensitivity, probability of detection = $\frac{\sum \text{True positive}}{\sum \text{Condition positive}}$
- **False negative rate (FNR), Miss rate** = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$
- **False positive rate (FPR), Fall-out**, probability of false alarm = $\frac{\sum \text{False positive}}{\sum \text{Condition positive}}$
- **Specificity (SPC), Selectivity, True negative rate (TNR)** = $\frac{\sum \text{True negative}}{\sum \text{Condition negative}}$
- **Positive likelihood ratio (LR+)** = $\frac{\text{TPR}}{\text{FPR}}$
- **Negative likelihood ratio (LR-)** = $\frac{\text{FNR}}{\text{TNR}}$
- **Diagnostic odds ratio (DOR)** = $\frac{\text{LR+}}{\text{LR-}}$
- **$F_1$ score** = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$

**In fact,** different stakeholders might have different points of view.
Equity Metrics of Fairness

What about general decisions: *how much loan to give? where to place an emergency room?* Where to schedule deliveries?

Is it **fair** to minimize **total distance** travelled by any group?
Equity Metrics of Fairness

What about general decisions: *how much loan to give?* **where to place an emergency room?** Where to schedule deliveries?

Is it *fair* to minimize *total distance* travelled by any group?
Equity Metrics of Fairness

What about general decisions: *how much loan to give? where to place an emergency room?* Where to schedule deliveries?

Is it *fair* to minimize *total distance* travelled by any group?
Equity Metrics of Fairness

What about general decisions: how much loan to give? where to place an emergency room? Where to schedule deliveries?

Is it fair to minimize average distance travelled by any group (per person)?
Equity Metrics of Fairness

What about general decisions: *how much loan to give? where to place an emergency room?* Where to schedule deliveries?

Is it *fair* to minimize *average distance* travelled by any group (per person)?
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Scaling</th>
<th>Reference</th>
<th>Framework for equity measure</th>
<th>Mean</th>
<th>Value</th>
<th>Normalized</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>$A$</td>
<td>$B$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\bar{E} - \bar{E}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
</tr>
<tr>
<td>$E$</td>
<td>$A$</td>
<td>$B$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\bar{E} - \bar{E}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
</tr>
<tr>
<td>$E$</td>
<td>$A$</td>
<td>$B$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\bar{E} - \bar{E}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
<td>$\frac{1}{N} \sum_{i=1}^{N} \frac{E_i - E}{A_i - B_i}$</td>
</tr>
</tbody>
</table>

**Equity Metrics of Fairness**

Marsh and Schilling 1993
Outline of the talk

- Bias in the data, models and variables
- Fairness Metrics
  - Statistical measures
  - Equity measures
- Trolley Problem of Choice
Which **fairness** do we want?

At least 50 ways to be fair

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the application.
Which **fairness** do we want?

Social Scientist:
Arrest data is not a good proxy for crime data

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.
Which fairness do we want?

Lawyer/Policy maker: Cannot use protected classes for making decisions.


Disparate Treatment v/s Impact

This has to be a collective decision we need to consciously reach at, after a deeper dive into the application.
Which **fairness** do we want?

**Lawyer/Policy maker**: Cannot use protected classes for making decisions.

**Disparate Treatment v/s Impact**

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.
Which **fairness** do we want?

Algorithm designer: awareness of protected classes can fix bias

Machine Bias

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

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.
Which **fairness** do we want?

**Statistician**: cannot have equal false positive, negative rates & calibration simultaneously

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

COMPAS Debate: Northpointe v/s ProPublica

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.
Which **fairness** do we want?

Optimizer: *can at times have approximately fair solutions for multiple metrics together*

Group by **race**? **income**? **insurance**?

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Bias and Fairness in AI/ML models | Swati Gupta | Ongoing work with Jalan, Ranade, Yang, Zhuang, 2018
Which **fairness** do we want?

This has to be a **collective decision** we need to *consciously* reach at, after a **deeper dive** into the **application**.

Economists, Behavioral scientists, Humans-in the loop, ..
Which **fairness** do we want?

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.
Summary

- **Bias in the data, models and variables**
  - Collection, Feedback, Proxies, Test Data, Representation..

- **Fairness Metrics**
  - Statistical measures: accuracy, false positive rate, true positive rate, calibration, ...
  - Equity measures: general decisions, average metric, total metric, group choice, ...

- **Trolley Problem of Choice:** it’s an inclusive story

Questions? swatig@gatech.edu, www.swatigupta.tech