COMPSCI 389
Introduction to Machine Learning

Days: Tu/Th.  Time: 2:30 – 3:45  Building: Morrill 2  Room: 222

Topic 15.0: Fairness
Prof. Philip S. Thomas (pthomas@cs.umass.edu)
Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- Fairness research
- Everything we talked about is wrong (not incorrect)
Claim: AI systems have produced what some might call “unfair” behavior.
<table>
<thead>
<tr>
<th>Gender</th>
<th>by Google Translate (via Turkish Pronouns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>he is a soldier</td>
<td>he is a president</td>
</tr>
<tr>
<td>she’s a teacher</td>
<td>he is an entrepreneur</td>
</tr>
<tr>
<td>he is a doctor</td>
<td>she is a singer</td>
</tr>
<tr>
<td>she is a nurse</td>
<td>he is a student</td>
</tr>
<tr>
<td>he is a writer</td>
<td>he is a translator</td>
</tr>
<tr>
<td>he is a dog</td>
<td>he is hard working</td>
</tr>
<tr>
<td>she is a nanny</td>
<td>she is lazy</td>
</tr>
<tr>
<td>it is a cat</td>
<td></td>
</tr>
</tbody>
</table>
Amazon Doesn’t Consider the Race of Its Customers. Should It?

By David Ingold and Spencer Soper
April 21, 2016
77.6m Americans live in ZIP codes where Amazon offers Prime Free Same-Day Delivery

Source: Bloomberg analysis of data from Amazon.com and the American Community Survey
The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.
<table>
<thead>
<tr>
<th>Gender Classifier</th>
<th>Darker Male</th>
<th>Darker Female</th>
<th>Lighter Male</th>
<th>Lighter Female</th>
<th>Largest Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>94.0%</td>
<td>79.2%</td>
<td>100%</td>
<td>98.3%</td>
<td>20.8%</td>
</tr>
<tr>
<td>FACE++</td>
<td>99.3%</td>
<td>65.5%</td>
<td>99.2%</td>
<td>94.0%</td>
<td>33.8%</td>
</tr>
<tr>
<td>IBM</td>
<td>88.0%</td>
<td>65.3%</td>
<td>99.7%</td>
<td>92.9%</td>
<td>34.4%</td>
</tr>
</tbody>
</table>
The official account of Tay, Microsoft’s A.I. farm from the internet that’s got zero chill! The more you talk the smarter Tay gets.

the internets
tay.ai/#about

hellooooooo w sıdırlld!!!

c u soon humans need sleep now so many conversations today thx❤️
@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32
@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45
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
<table>
<thead>
<tr>
<th>Did not reoffend</th>
<th>Black</th>
<th>White</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did reoffend</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>White</td>
</tr>
<tr>
<td>---------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Did not reoffend</td>
<td>Green</td>
<td>Green</td>
</tr>
<tr>
<td>Did reoffend</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>White</td>
</tr>
<tr>
<td>----------------</td>
<td>-------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Did not reoffend</td>
<td>44.9% labeled as high risk</td>
<td>23.5% labeled as high risk</td>
</tr>
<tr>
<td>Did reoffend</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Google Mistakenly Tags Black People as ‘Gorillas,’ Showing Limits of Algorithms

By Alistair Barr
Updated July 1, 2019 3:03 pm ET

Google is a leader in artificial intelligence and machine learning. But the company’s computers still have a lot to learn, judging by a major blunder by its Photos app this week.

The app tagged two black people as “Gorillas,” according to Jacky Alcindé, a Web developer who spotted the error and tweeted a photo of it.

"Google Photos, y'all f**ked up. My friend’s not a gorilla," he wrote on Twitter.

Google apologized and said it’s tweaking its algorithms to fix the problem.

“We’re appalled and genuinely sorry that this happened,” a company spokeswoman said.

“There is still clearly a lot of work to do with automatic image labeling, and we’re looking..."
Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- Fairness research
- Everything we talked about is wrong (not incorrect)
• 9 Entrance Exams
  • Physics
  • Biology
  • History
  • Second language
  • Geography
  • Literature
  • Portuguese and Essay
  • Math
  • Chemistry
• GPA from first 3 semesters
• Gender

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/O35FW8
```python
import pandas as pd

df = pd.read_csv('data/GPA_full.csv')
display(df)
```

<table>
<thead>
<tr>
<th>gender</th>
<th>physics</th>
<th>biology</th>
<th>history</th>
<th>English</th>
<th>geography</th>
<th>literature</th>
<th>Portuguese</th>
<th>math</th>
<th>chemistry</th>
<th>gpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>622.60</td>
<td>491.56</td>
<td>439.93</td>
<td>707.64</td>
<td>663.65</td>
<td>557.09</td>
<td>711.37</td>
<td>731.31</td>
<td>509.80</td>
<td>1.33333</td>
</tr>
<tr>
<td>1</td>
<td>538.00</td>
<td>490.58</td>
<td>406.59</td>
<td>529.05</td>
<td>532.28</td>
<td>447.23</td>
<td>527.58</td>
<td>379.14</td>
<td>488.64</td>
<td>2.98333</td>
</tr>
<tr>
<td>2</td>
<td>455.18</td>
<td>440.00</td>
<td>570.86</td>
<td>417.54</td>
<td>453.53</td>
<td>425.87</td>
<td>475.63</td>
<td>476.11</td>
<td>407.15</td>
<td>1.97333</td>
</tr>
<tr>
<td>3</td>
<td>756.91</td>
<td>679.62</td>
<td>531.28</td>
<td>583.63</td>
<td>534.42</td>
<td>521.40</td>
<td>592.41</td>
<td>783.76</td>
<td>588.26</td>
<td>2.53333</td>
</tr>
<tr>
<td>4</td>
<td>584.54</td>
<td>649.84</td>
<td>637.43</td>
<td>609.06</td>
<td>670.46</td>
<td>515.38</td>
<td>572.52</td>
<td>581.25</td>
<td>529.04</td>
<td>1.58667</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>43300</td>
<td>798.75</td>
<td>817.58</td>
<td>731.98</td>
<td>648.42</td>
<td>751.30</td>
<td>648.67</td>
<td>662.05</td>
<td>773.15</td>
<td>835.25</td>
<td>3.75000</td>
</tr>
<tr>
<td>43301</td>
<td>527.66</td>
<td>443.82</td>
<td>545.88</td>
<td>624.18</td>
<td>420.25</td>
<td>676.80</td>
<td>583.41</td>
<td>395.46</td>
<td>509.80</td>
<td>2.50000</td>
</tr>
<tr>
<td>43302</td>
<td>512.56</td>
<td>415.41</td>
<td>517.36</td>
<td>532.37</td>
<td>592.30</td>
<td>382.20</td>
<td>538.35</td>
<td>448.02</td>
<td>496.39</td>
<td>3.16667</td>
</tr>
</tbody>
</table>

43303 rows × 11 columns
Can we predict GPAs from entrance exams?

• Let’s focus on one exam, “biology”
Can we predict GPAs from entrance exams?

- **Linear fit:**
  - Slope: 0.0019
  - Y-intercept: 1.7

- **Question:** Would it be fair and/or responsible to use this system to predict student GPAs? Why or why not?
Desirable fairness properties

• The model should not over-predict for one gender and under-predict for another.
  • $\text{abs}(E[Y - \hat{Y}|\text{Male}] - E[Y - \hat{Y}|\text{Female}])$ should be small

• The model should not predict higher values on average for one gender.
  • $\text{abs}(E[\hat{Y}|\text{Male}] - E[\hat{Y}|\text{Female}])$ should be small
What if we consider gender?

- Male $\rightarrow$ shift prediction down by 0.15 GPA points.
- Female $\rightarrow$ shift prediction up by 0.15 GPA points.
- Average over-prediction for men: $0.15 - 0.15 = 0!$
- Average over-prediction for women: $-0.15 - (-0.15) = 0!$

Note: Actually -.137... for men and +0.146... for women.
Is the model now fair?

• Average prediction error for men: $\approx 0$
• Average prediction error for women: $\approx 0$

• Average predicted GPA for men: $\approx 2.6$
• Average predicted GPA for women: $\approx 3.0$

Desirable fairness properties

• The model should not over-predict for one gender and under-predict for another.
  • $\text{abs}(E[Y \mid \text{Male}] - E[Y - \hat{Y} \mid \text{Female}])$ should be small

• The model should not predict higher values on average for one gender.
  • $\text{abs}(E[\hat{Y} \mid \text{Male}] - E[\hat{Y} \mid \text{Female}])$ should be small
Overview

• AI systems have produced unfair behavior
• An illustrative example: Predicting student GPAs
• **Impossibility results**
• Sources of “bias”
• Fairness research
• Everything we talked about is wrong (not incorrect)
Do not (on average):
- Predict higher values for one gender
- Over-predict more for one gender
Fairness definitions often conflict!

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jou Kleinberg *  Seadhil Mullainathan 1  Munish Raghavan 1

Abstract
Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of fairness. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly contrived special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying any two conditions approximately requires that the data be in an approximate version of one of the constrained special cases identified by our framework. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human experts to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative risk tools to assess a defendant’s probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [13,23]. One of their main contentions was that the tool’s errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool’s errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [24,25,26].

Proposition 2. Assume that $A$ and $Y$ are not independent. Then sufficiency and independence cannot both hold.

Proposition 5. Assume $Y$ is not independent of $A$ and assume $Y$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.
In any effort to regulate the use of machine learning to ensure fairness, a critical first step is to define precisely what fairness means. This may require recognizing that certain behaviors that appear to be unfair may necessarily be permissible, in order to enable enforcement of a conflicting and more appropriate notion of fairness.
EDUCATIONAL REDLINING

Student Borrower Protection Center

February 2020

PROTECTBORROWERS.ORG

Our findings from our broader analysis and the highlighted case studies are consistent: holding all else constant, borrowers who attend community colleges, Historically Black Colleges and Universities (HBCUs), and Hispanic-Serving Institutions (HSIs) will pay significantly more for credit, because of people’s prejudices regarding those who sit next to them in the classroom.
Slippery Slope!
• Every decision making system will be unfair from some perspective.
• When accusing a system of being unfair, make sure that there is an established notion of what fair means in the given context.
• [Defense] When you hear about a system being unfair, check if the accusation discusses conflicting definitions.
• [Prosecution] When the accused claims innocence due to a conflicting fairness definition, 1) ensure that they actually enforce that definition and 2) determine which fairness definition should take precedence.
• It is critical that we agree on the “right” definition of fairness for key applications like automated loan approval.
The right definition of fairness
Overview

• AI systems have produced unfair behavior
• An illustrative example: Predicting student GPAs
• Impossibility results
• Sources of “bias”
• Fairness research
• Everything we talked about is wrong (not incorrect)
Source of Bias (1/3): Malicious intent

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45
Source of Bias (2/3): “Biased” data

The thing about data is...

Garbage in, Garbage out.
Source of Bias (3/3): “Biased” algorithms

Over/under-predicted relative to the data.

Additional bias added by the machine learning algorithm, on top of any bias in the data!
Source of Bias (3/3): Conflicting Objectives

• Drive to Boston as fast as possible, but stop at red lights.
• Eat lunch as fast as possible between meetings, but don’t choke.
• Order the tastiest food, but don’t make future you unhappy.
• Jail as many murderers as possible, but don’t jail innocent people.
• Make predictions as accurate as possible, but make sure they are fair.

• In order to make fair predictions, you (usually) cannot make predictions as accurately as possible.
Overview

• AI systems have produced unfair behavior
• An illustrative example: Predicting student GPAs
• Impossibility results
• Sources of “bias”
• **Fairness research**
• Everything we talked about is wrong (not incorrect)
• Creating fair algorithms
ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.
Fair Seldonian algorithms

- Allow the user to define fairness
- Allow the user to pick a probability, $p$
- Guarantee with probability $p$ that they will not produce unfair decision-making rules

Check out Seldonian.cs.umass.edu!
Past and Current Research Projects

• Can we make fairness guarantees robust to *demographic shift*?
• Can we make fairness guarantees robust to general *distributional shift*?
• Can we make fairness guarantees robust to adversarial data corruptions?
• Can we achieve the same fairness guarantees with less data?
• Can we enforce fairness guarantees in other machine learning settings, like contextual bandits and reinforcement learning?
• Can we broaden the class of fairness definitions that our algorithms can handle?
Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- Everything we talked about is wrong (not incorrect)
The Minneapolis Domestic Violence Experiment (MDVE) evaluated the effectiveness of various police responses to domestic violence calls in Minneapolis, Minnesota. This experiment was implemented during 1981-82 by Lawrence W. Sherman, Director of Research at the Police Foundation, and by the Minneapolis Police Department with funding support from the National Institute of Justice.[1] Among a pool of domestic violence offenders for whom there was probable cause to make an arrest, the study design called for officers to randomly select one third of the offenders for arrest, one third would be counseled and one third would be separated from their domestic partner.

The results of the study, showing a deterrent effect for arrest, had a "virtually unprecedented impact in changing then-current police practices."[2] Subsequently, numerous states and law enforcement agencies enacted policies for mandatory arrest, without warrant, for domestic violence cases in which the responding police officer had probable cause that a crime had occurred.
Metaphorical use

The expression *fig leaf* has a pejorative metaphorical sense meaning a flimsy or minimal cover for anything or behaviour that might be considered shameful, with the implication that the cover is only a token gesture and the truth is obvious to all who choose to see it.[7]
• Fairness 1: I was equally likely to give loans to black and white people.
• Fairness 2: Of the people who would repay their loan, I was equally likely to give them a loan.
• Fairness 3: I did not consider race when deciding whether to give a loan.

Short-Term Notions of Fairness

Fig-Leaf Fairness

Delayed Impact: The automated loan approval system makes choices that reduce a variety of measures of racial inequality over 10–50 years.
DELAYED IMPACT OF FAIR MACHINE LEARNING

Lydia T. Liu (UC Berkeley)

Joint work with Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt
Past and Current Research Projects

• Can we make fairness guarantees robust to demographic shift?
• Can we make fairness guarantees robust to general distributional shift?
• Can we make fairness guarantees robust to adversarial data corruptions?
• Can we achieve the same fairness guarantees with less data?
• Can we enforce fairness guarantees in other machine learning settings, like contextual bandits and reinforcement learning?
• Can we broaden the class of fairness definitions that our algorithms can handle?
• Can we enforce delayed impact fairness definitions?