Accountability and Compliance in the Age of AI

Anupam Datta
Professor, ECE & CSD
Carnegie Mellon University
AI Systems are Ubiquitous

How Big Data Could Replace Your Credit Score

Credit scores are useful in determining who gets loans, but they’re far from perfect. AvantCredit determines loan-worthiness based on all sorts of factors, including your use of social media and prepaid cell phones.

Big Data in Education

Learn how and when to use key methods for educational data mining and learning analytics on large-scale educational data.

TEACHERS COLLEGE
COLUMBIA UNIVERSITY
Themes

1. How AI black boxes threaten societal values, including privacy and fairness

2. Research progress on opening up AI black boxes to discover and mitigate their problems

3. Engineering tools to support accountability and compliance activities in the age of AI
Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

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

Public Safety

Police are using software to predict crime. Is it a ‘holy grail’ or biased against minorities?

By Justin Jouvenal
November 17, 2016

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Forbes Staff
Welcome to The Not-So Private Parts where technology & privacy collide
External Audit Tools for AI Systems
WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.
Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Amit Datta, Michael Carl Tschantz, and Anupam Datta

Volume/Issue: Volume 2015: Issue 1
First Online: 18 Apr 2015
Page Count: 92–112
DOI: https://doi.org/10.1515/popets-2015-0007
"We'll be back": Hong Kong protesters chant as camp site dismantled

Reuters | Dec 19, 2014, 08:39 AM IST

HONG KONG: Hong Kong police arrested pro-democracy activists and cleared most of the main protest site on Thursday, marking an end to more than two months of street demonstrations in the Chinese-controlled city, but many chanted: "We will be back".

Most activists chose to leave the Admiralty site, next to the Central business area, peacefully, despite their demands for a free vote not being met. But the overall mood remained defiant.

Hong Kong Federation of Students leader Alex Chow said: "You might have the clearance today but people will come back on to the streets."

Police officers stand guard before they move on to remove protesters from a road written 'We Will Be Back' with tarps at an occupied area outside government headquarters in Hong Kong.
Antidepressant Medication - Info On An Rx Antidepressant Drug
knowmydepression.com/antidepressant ▼
Visit For Treatment Info & Facts.
### Settings for Google ads

Ads enable free web services and content. These settings help control the types of Google ads you see.

<table>
<thead>
<tr>
<th>Ads on Google</th>
<th>Google ads across the web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Female [Edit]</td>
</tr>
<tr>
<td>Age</td>
<td>25-34 [Edit]</td>
</tr>
<tr>
<td>Languages</td>
<td>English [Edit]</td>
</tr>
<tr>
<td>Interests</td>
<td>Air Travel, and 30 more [Edit]</td>
</tr>
</tbody>
</table>

**Opt-out settings**

- You've opted out of **interest-based** ads on Google.
  - **Opt in** to **interest-based** ads on Google
  - **Opt out** of **interest-based** Google ads across the web
AdFisher

Contribution: The rigor of experimental science

- Causal effects
- Statistical significance
- Automation

Is there a difference? P-value
Discrimination

Web browsing

Browse websites related finding a new job

Google Ad ecosystem

Ad settings

Advertisements

Set the gender bit to female or male

Significant difference ads on news website 
\( p < 0.000006 \)
Discrimination Finding

Fails the 80% rule for disparate impact

$200k+ Jobs - Execs Only

Find Next $200k+ Job

Female

Male

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>0% - 20%</td>
<td>311</td>
<td>1816</td>
</tr>
<tr>
<td>20% - 40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40% - 60%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60% - 80%</td>
<td>7</td>
<td>36</td>
</tr>
<tr>
<td>80% - 100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$200k+ Jobs - Execs Only

Find Next $200k+ Job
Discrimination in Online Advertising: A Multidisciplinary Inquiry

Section 704(b), Title VII of Civil Rights Act

- Unlawful “to print or publish or cause to be printed or published any … advertisement relating to employment … indicating any preference … based on … sex …”
Classified ads in newspapers

350
HELP WANTED, MEN

ACCOUNTANT — CPA or CPA
candidate for small S.F. finan. dist.
CPA firm perm. Resume to this
paper AD No. 54081.¹

Source: San Francisco Chronicle,
Jan. 21, 1972
The ad ecosystem has many parties

- The Barrett Group
- Google
- Male users
- Female users
- Other advertisers
- Other entities

Ad Ecosystem
Possible cause: direct advertiser targeting

Input → Output

The Barrett Group → Show to males → Other advertisers
Google → OK, will do.
Male users → Other entities
Female users
Google allows targeting on gender

Choose how to target your ads

- Demographics
- Interests & remarketing (affinity audiences) – show ads to people based on their interests.
- Use a different targeting method

### Demographics

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Parental Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18-24</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>25-34</td>
<td>Parent</td>
</tr>
<tr>
<td></td>
<td>35-44</td>
<td>Not a parent</td>
</tr>
<tr>
<td></td>
<td>45-54</td>
<td></td>
</tr>
<tr>
<td></td>
<td>55-64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>65 or more</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reach a significantly wider audience by showing ads to people whose Age, Gender, and Parental status we do not know.
Applicability of 704(b)

Applies only to an

1. employer,
2. labor organization,
3. employment agency, or
4. joint labor-management committee
Analogous Statutes

Title VII (employment):

It shall be ... unlawful ... for an employer, labor organization, employment agency, or joint labor-management committee ... to ... publish ... any ... advertisement relating to employment ... indicating any preference, ... based on ... sex, ...

Title VIII (housing):

[l]t shall be unlawful ... [t]o ... publish, ... any ... advertisement, with respect to the sale or rental of a dwelling that indicates any preference, ... based on ... sex, ...

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac
Communications Decency Act § 230

• Law designed to protect companies that provide spaces for speech online

• Shields “interactive computer service” from liability for content created by others

• Protects these computer services when they provide “neutral tools” that are used by third parties to create content
Possible cause: targeting a correlate

- The Barrett Group
- Other advertisers
- Google
- Websites
- Male users
- Female users
- Show to high earners

High earners are male

Input ➔

Output ➞
Exceptions to CDA Section 230?

• Protection is not absolute

• “Information content providers”: Entities responsible, in whole or in part, for the creation or development of information
Exceptions to CDA Section 230? (cont’d)

• Fair Housing Council v. Roommates (9th Cir. 2008)

• Targeting of ads is itself discriminatory (even if the content of the ad on its face is not)

• Advertising platform is not a “neutral tool”
What’s next

• Mismatch between responsibility and capability

• Policy changes
  – Revise Section 704(b) to make it applicable to all actors in the context of employment advertising

• Technological changes
  – Revise targeting algorithms
Tools for Explaining AI Systems’ Decisions
Discrimination

Algorithms and bias: What lenders need to know

Amazon scraps secret AI recruiting tool that showed bias against women

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

Police are using software to predict crime. Is it a 'holy grail' or biased against minorities?

By Justin Jouvenal
November 17, 2016
Discrimination

20 JAN 2017 | Insight

Kevin Petrasic

**Algorithms and bias: What lenders need to know**

The algorithms that power fintech may discriminate in ways that can be difficult to anticipate—and financial institutions can be held accountable even when alleged discrimination is clearly unintentional.
AI Systems are Opaque Black Boxes

User data

Credit Classifier

Decisions

DENIED
Adverse action notices

When a credit application is denied, the consumer or business has to be provided with the principal reasons behind the denial.

- Equal Credit Opportunity Act
  - To guard against discrimination and provide transparency into underwriting
- Federal Credit Reporting Act
  - To allow consumers to correct errors in their credit report
Algorithmic Transparency via Quantitative Input Influence:
Theory and Experiments with Learning Systems

Anupam Datta    Shayak Sen    Yair Zick
Carnegie Mellon University, Pittsburgh, USA
{danupam, shayaks, yairzick}@cmu.edu
Lending Club Loans Data

- All loans issued by Lending Club from 2007-2015
  - 900k data points
  - 75 variables

- Build AI models to predict charge-offs
Q Explanations

51-variable GBM, individualized explanation

Most variables aren't important for any particular individual

Average number of negative influencers: 4.6
Replace feature with random values from the population, and examine distribution over outcomes.
Key Idea | Aggregating Marginal Influence

Think of features as states in an election

What is the effect of PA results after results from IN, GA, MD are in?
Disparate impact and business necessity

If a protected group gets significantly worse outcomes then the onus is on the employer to provide a business necessity defense

• Title VII of Civil Rights Act
Proxy use

[Datta, Fredrikson, Ko, Mardziel, Sen 2017; Yeom, Datta, Fredrikson 2018]

Protected information types: Race, sex

- Gang affiliation
- Age during latest arrest
- ...

Likelihood of involvement in shooting incident
Use Privacy Violations
[Datta, Fredrikson, Ko, Mardziel, Sen 2017]

Protected information type:
Pregnancy status

- Scent-free lotion
- Prenatal vitamins
- ...

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill, Forbes Staff
Welcome to The Not-So Private Parts where technology & privacy collide

Use of sensitive health information for targeting of Google ads raises privacy concerns
GDPR Explanations (Article 15)

“...the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic* involved, as well as the significance and the envisaged consequences of such processing for the data subject.”

*emphasis added
Cluster Explanations
[Datta, Sen with Bracke, Jung of Bank of England 2018]

Cluster 0
- \( \text{year} \leq 2011 \)
- type = not remortgagors
- rate type = fixed

Low PD

Cluster 3
- \( \text{year} \leq 2011 \)
- type = not remortgagors
- rate type = not fixed

Cluster 2
- \( \text{year} \geq 2011 \)
- type = remortgagors
- rate type = trackers

Higher PD

Cluster 4
- \( \text{year} \leq 2011 \)
- type = remortgagors
- rate type = trackers

Higher PD
Themes

1. How AI black boxes threaten societal values, including privacy and fairness

2. Research progress on opening up AI black boxes to discover and mitigate their problems

3. Engineering tools to support accountability and compliance activities in the age of AI
Additional slides
Sexist ad targeting

Targeted to females

Targeted to males
Sexist ad delivery

56,497 impressions, all to females

73,607 impressions, all to males
Possible cause: direct advertiser targeting

The Barrett Group

Other advertisers

Google

Male users

Female users

Show to males

OK, will do.
Some relevant regulations

• United States
  • Credit, employment, housing
    • ECOA, FCRA, Title VII, Title VIII, FHA
  • Protection from discrimination
  • Explanations for unfavorable outcomes

• Europe
  • GDPR
  • Protection from discrimination
  • Right to explanation for automated decisions

• Legal interpretation and compliance for AI systems being explored
Inferential Privacy Violations

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill  Forbes Staff
Welcome to The Not-So Private Parts where technology & privacy collide

New AI can guess whether you're gay or straight from a photograph

An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions