Responsible Artificial Intelligence

& the role of measurement

Manish Raghavan

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

MIT EECS

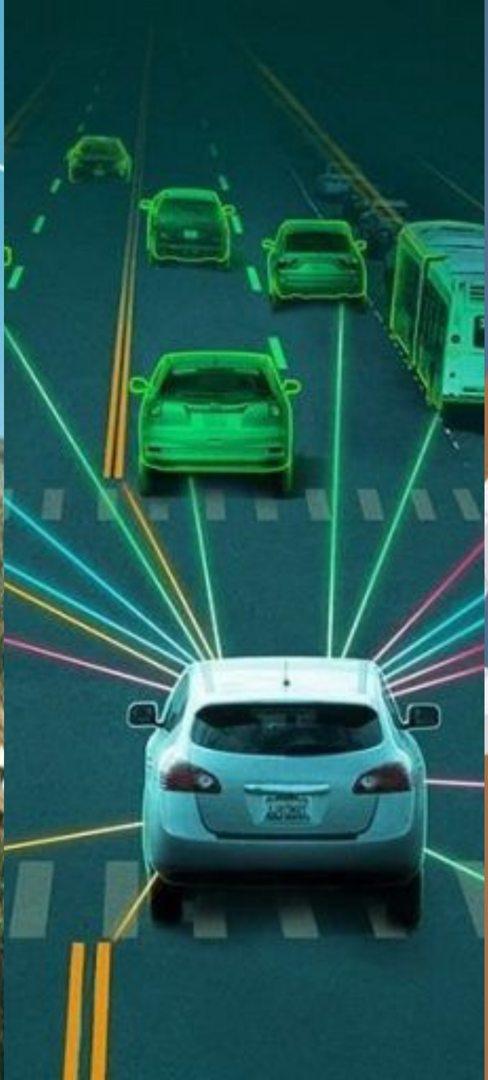
MIT MANAGEMENT SLOAN SCHOOL

AI is everywhere

Enterprise Artificial Intelligence (AI) Market is Expected To Reach USD 59.17 Billion By 2028

AI Adoption Skyrocketed Over the Last 18 Months


24°/18°
Percipitation 5%
1016 hPa
pH 5.0




RESUMÉ


SF Public and Private Schools (Shapes)
Proximity Filtering
Add Pod or Filter
Settings
No Filters Applied
Proximity Filter On
Remove
Include only
Apply as filter to:
Disable
LOCKED
Clay St
Golden Gate Ave
Grove St
Hayes St
Mission St
Natoma St
Folsom St
Harrison St
8th St
Francisco

Problems

Amazon scraps secret AI recruiting tool that showed bias against women

Increased use of AI in health care raises questions about fairness and equity

AI Bias Harms Black Families and Businesses. Howard University Is Working to Change That

Machine Bias

Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match

Responsible AI to the rescue?

Australia’s Artificial Intelligence Ethics Framework

PRINCIPLES OF ARTIFICIAL INTELLIGENCE ETHICS FOR THE INTELLIGENCE COMMUNITY

Business Roundtable Roadmap for Responsible Artificial Intelligence

What does this actually mean?

Whittlestone et al., 2019

What do we mean by AI?

- This talk: **supervised machine learning**
- Essentially just pattern matching
- Given input-output pairs, make output predictions for new inputs

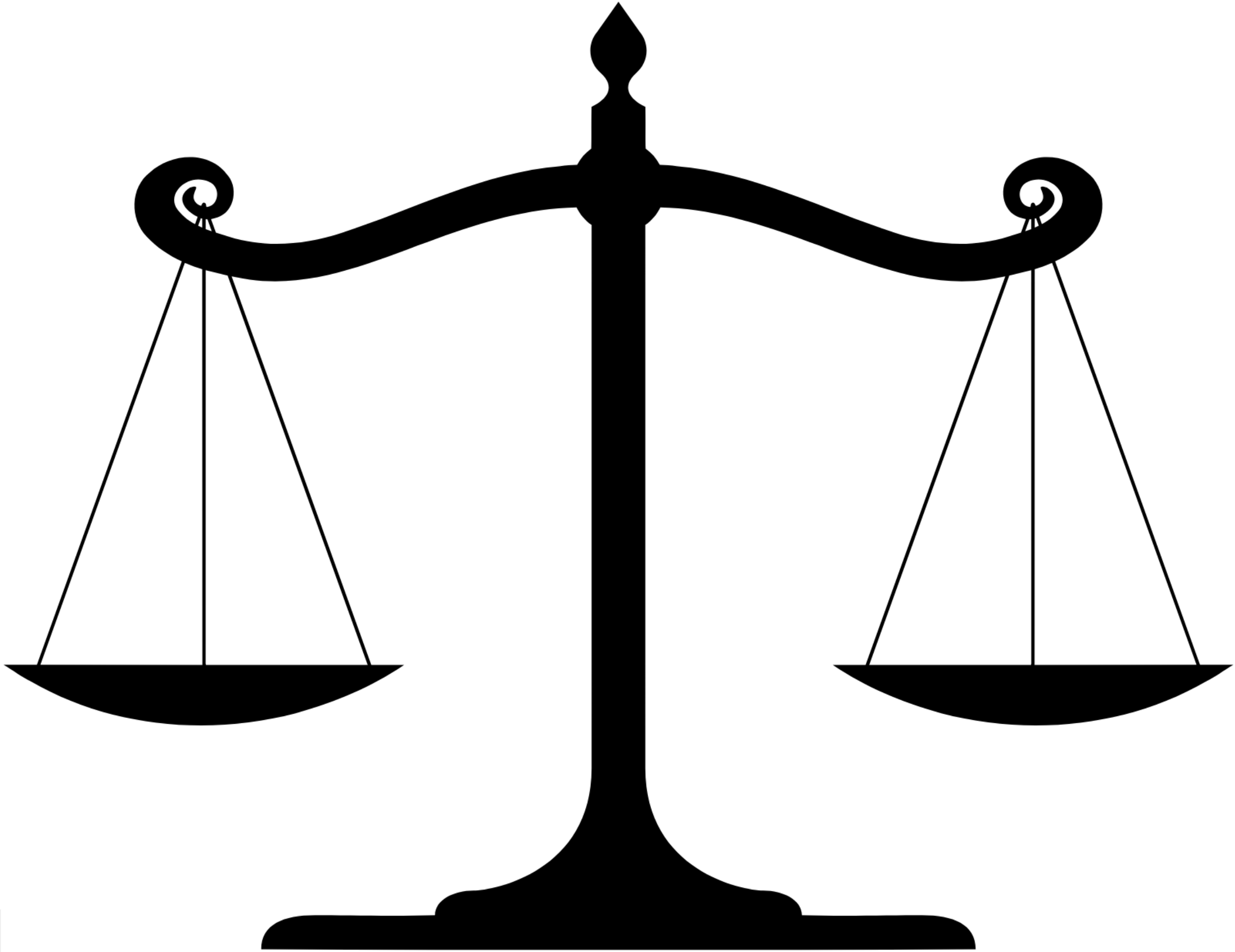
Why is AI any different?

- We don’t necessarily understand it
- New levers for observation & control
- Regulatory uncertainty

This talk: Measurement

- **Not** about governance, organizational principles, etc.
- **Not** about the technical methods
- What can quantitative measures tell us?
- Where should we be cautious?

Measurement

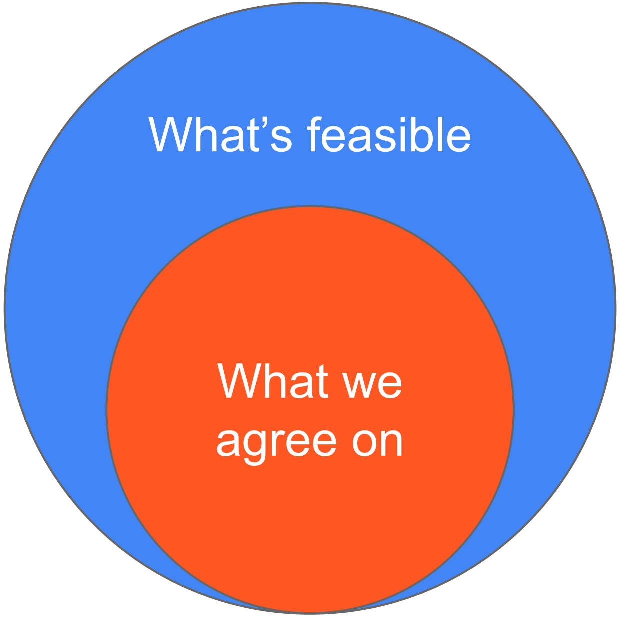


Metrics: a solution for responsible AI?

Plan

1. Formally define “responsible”
2. Build systems that respect this definition

Problem #1: Metrics need normative values


What's feasible
What we
agree on

Metrics in criminal justice

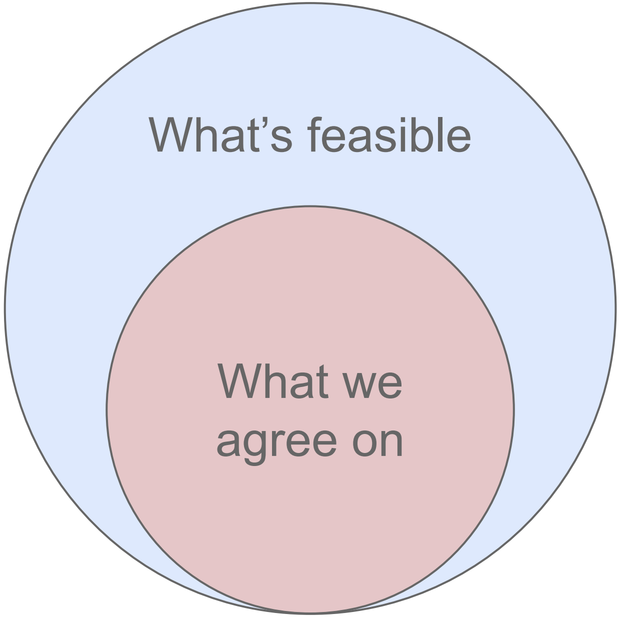
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend (FPR)	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend (FNR)	47.7%	28.0%

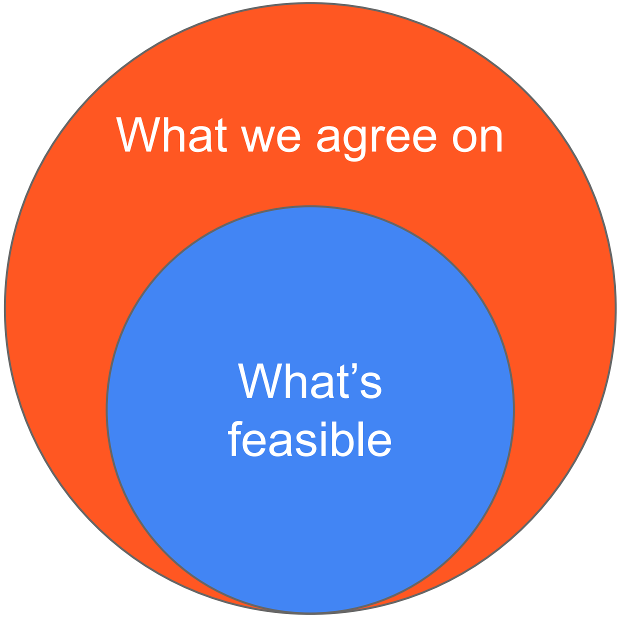
Angwin, Larson, Mattu, and Kirchner, 2016

Theorem: Unless predictions are systematically dishonest, they will **necessarily** generate error rate disparities.

Kleinberg, Mullainathan, and Raghavan, 2017

Problem #2: Metrics can't make disparities disappear


What's feasible
What we
agree on


What we agree on
What's
feasible

Metrics as diagnosis?

Abebe, Barocas, Kleinberg, Levy, Raghavan, and Robinson, 2020

Diagnosis in facial analysis

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini & Gebru, 2018

Diagnosis → improvement?

IBM Research Blog Topics Labs

AI

Mitigating Bias in AI Models

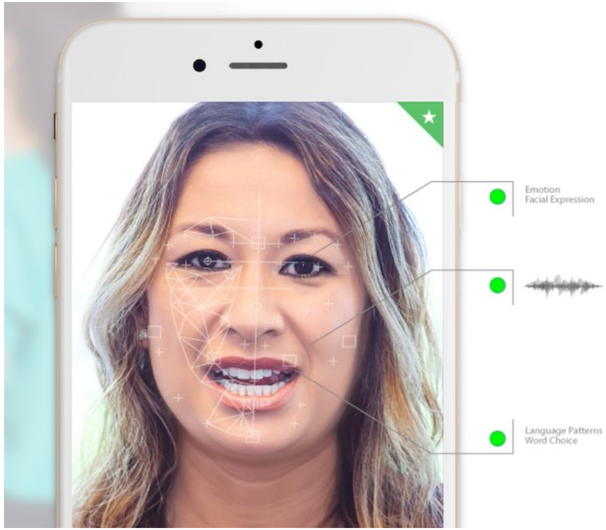
Microsoft | The AI Blog Our Company News and Stories Press Tools

Microsoft improves facial recognition technology to perform well across all skin tones, genders

Beyond diagnosis

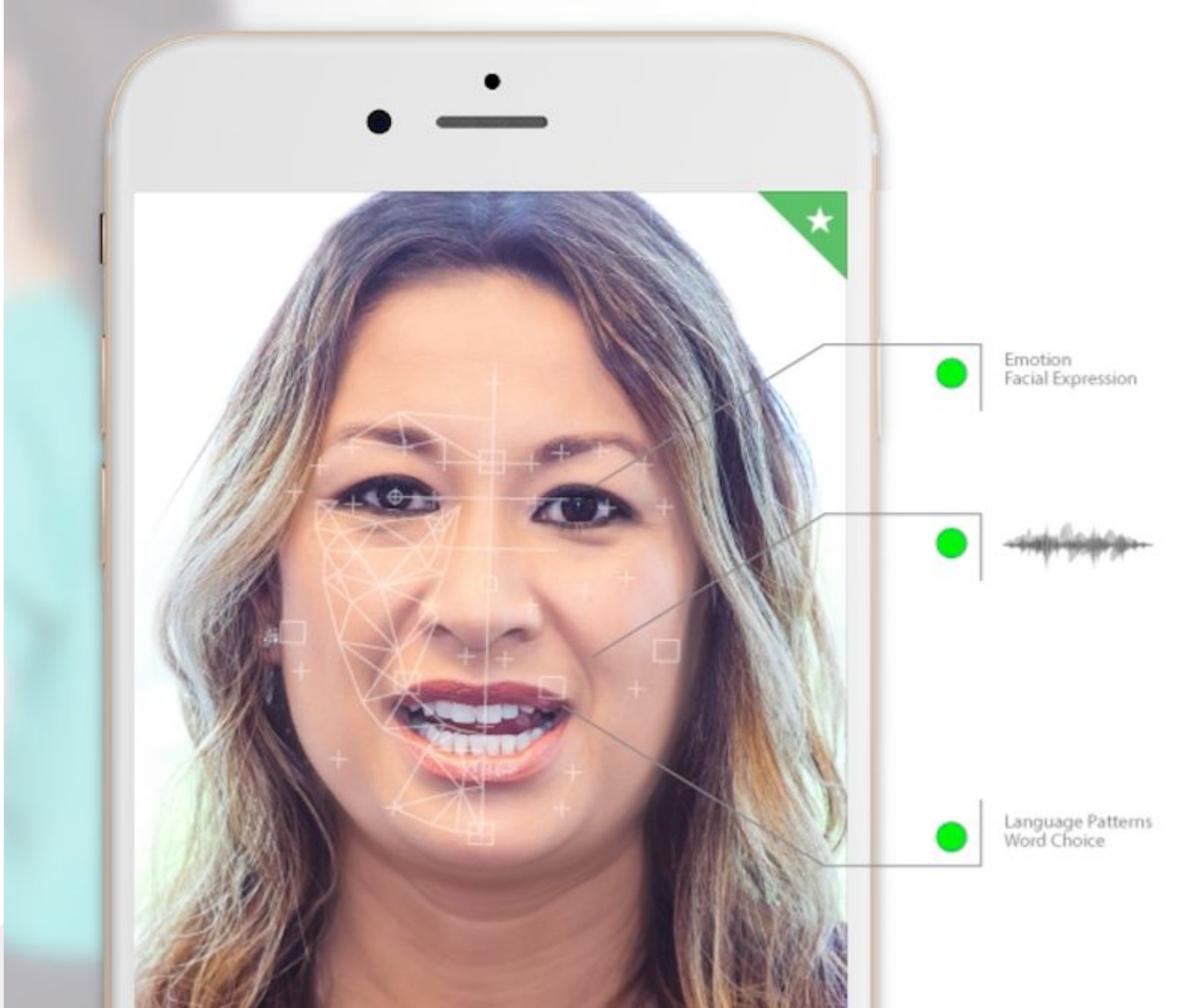
Today: two settings

Algorithmic hiring

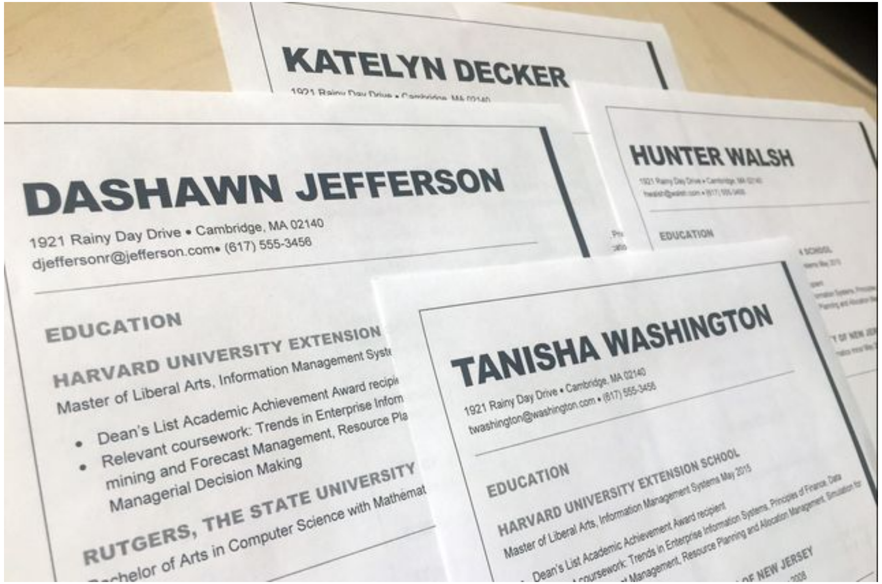


Online platforms

Algorithmic hiring

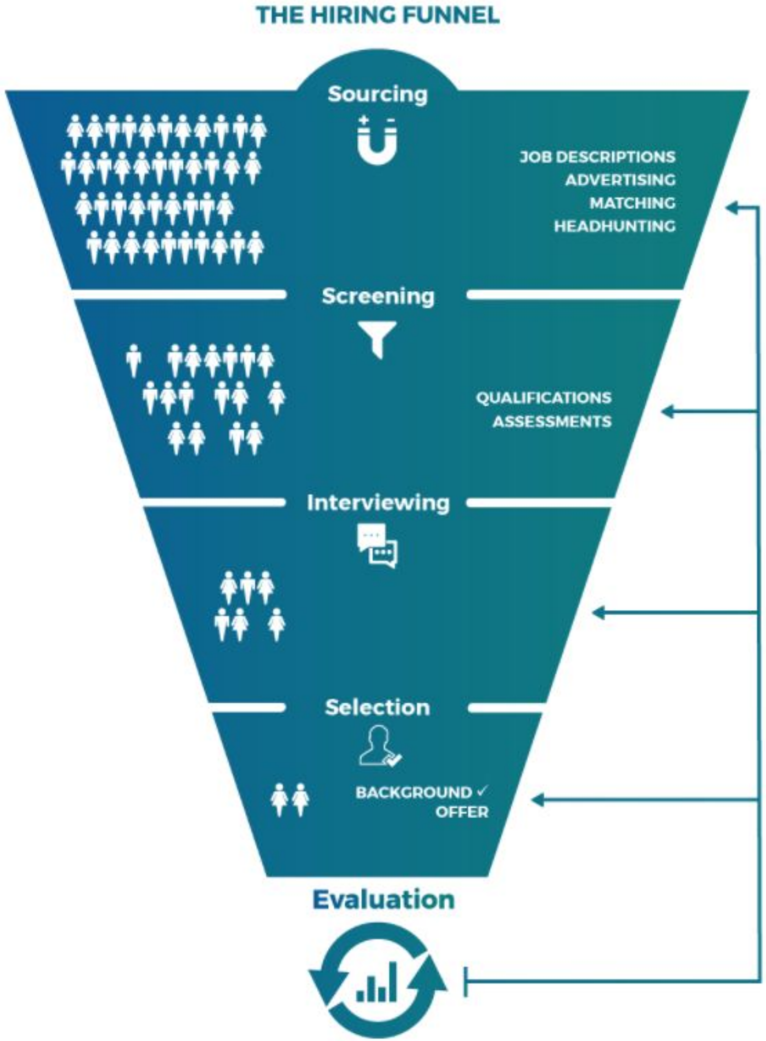


Basic problem: discrimination



Jessica Wolf/UCLA

Same resumes, different names
[Bertrand and Mullainathan, 2004]



Bogen & Rieke, 2018

Advertising

Targeting

Optimization

Alex Slobodkin | Getty Images

Case study: HUD v. Facebook (2018)

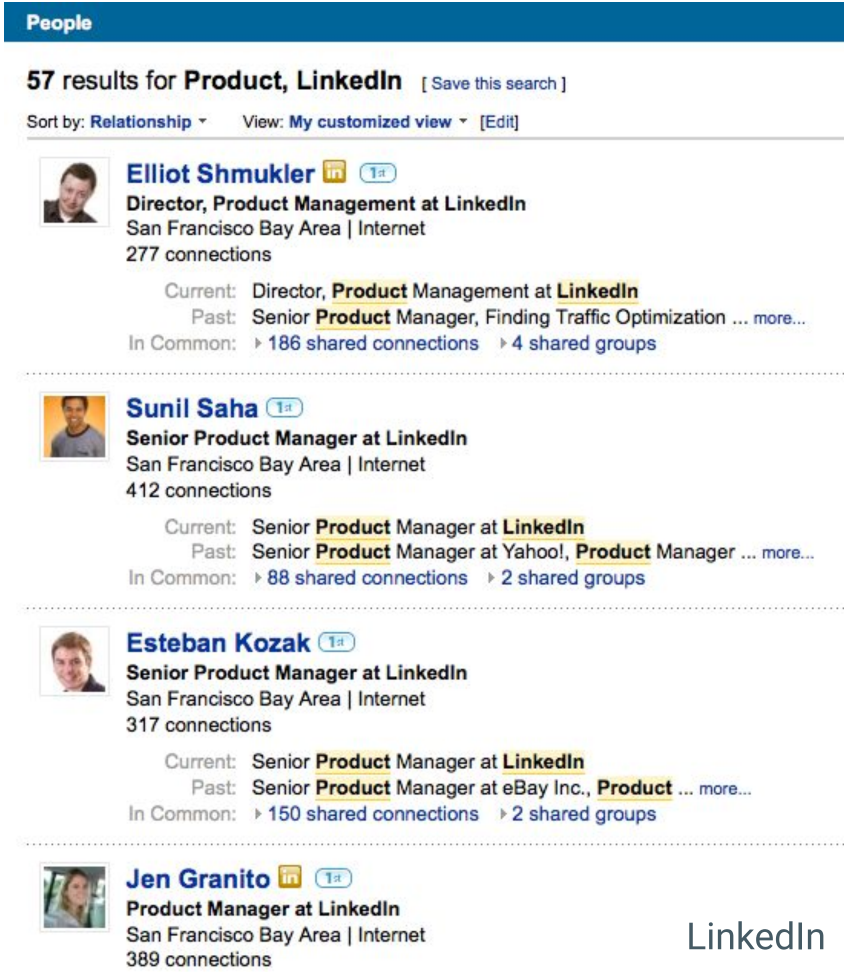
UNITED STATES OF AMERICA
DEPARTMENT OF HOUSING AND URBAN DEVELOPMENT
OFFICE OF ADMINISTRATIVE LAW JUDGES

The Secretary, United States Department of Housing and Urban Development, on behalf of Complainant Assistant Secretary for Fair Housing and Equal Opportunity,	
Charging Party,	HUD ALJ No. FHEO No. 01-18-0323-8
v.	
Facebook, Inc.,	
Respondent	

CHARGE OF DISCRIMINATION

Search

Who do recruiters see?



LinkedIn

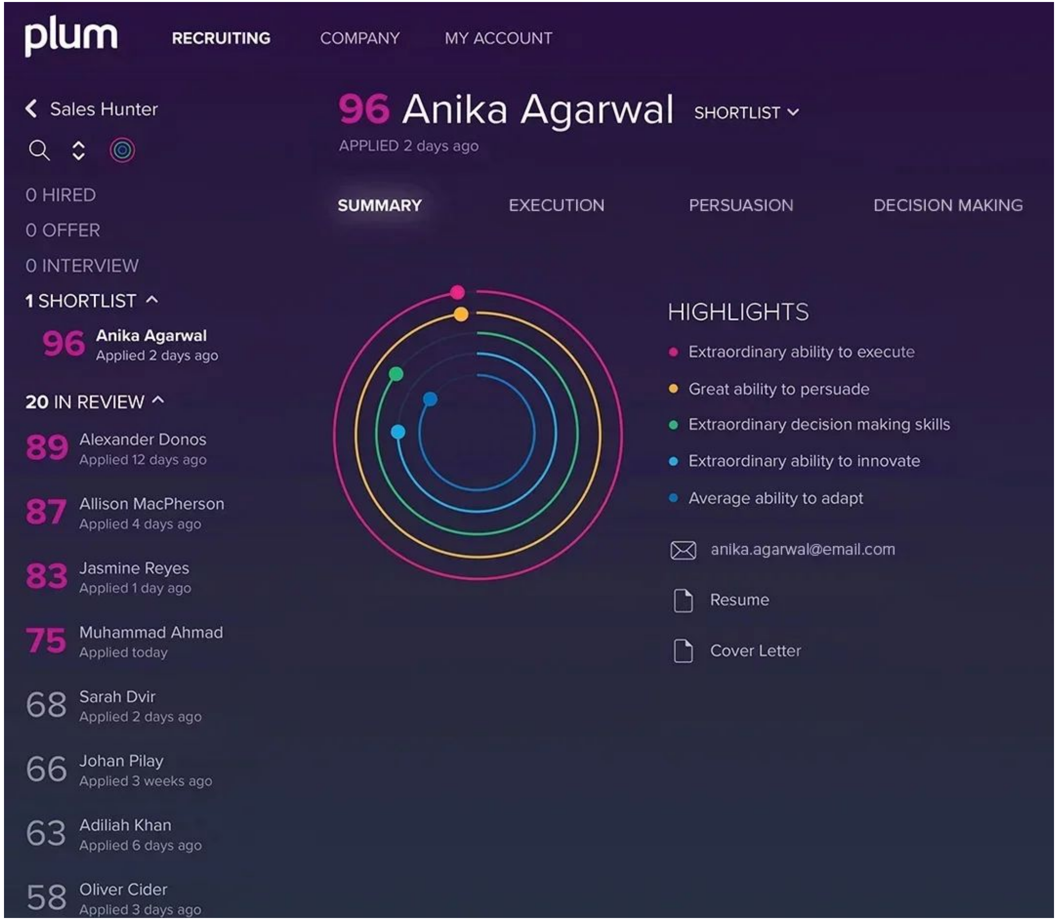
Case study: LinkedIn Search (2018)

As of last week, LinkedIn's search results for recruiters are designed to reflect the gender distribution for particular types of job in each industry. For example, if 44 percent of the talent pool for account executives in the U.S. are women, each page of candidate results for that position will reflect that mix, said John Jersin, head of product for talent solutions at LinkedIn.

Jordan Novet, CNBC

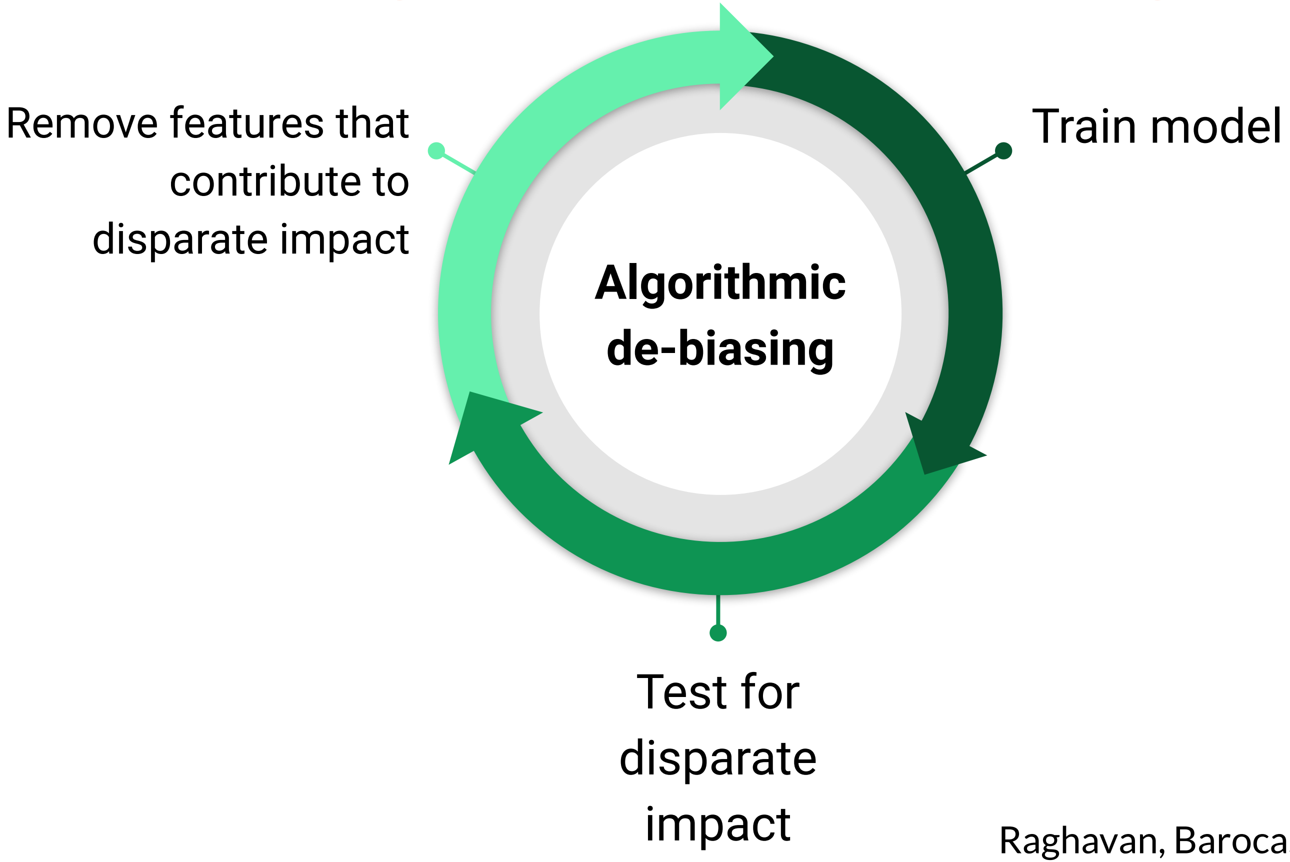
Assessments

Who should employers interview?



plum.io

Case study: Algorithmic De-biasing



Raghavan, Barocas, Kleinberg, Levy 2020

Takeaways

- Wide range of possible metrics
- Contextually dependent (advertising, search, assessments...)
- Rapidly changing legal environment

Online platforms

What determines what people see?

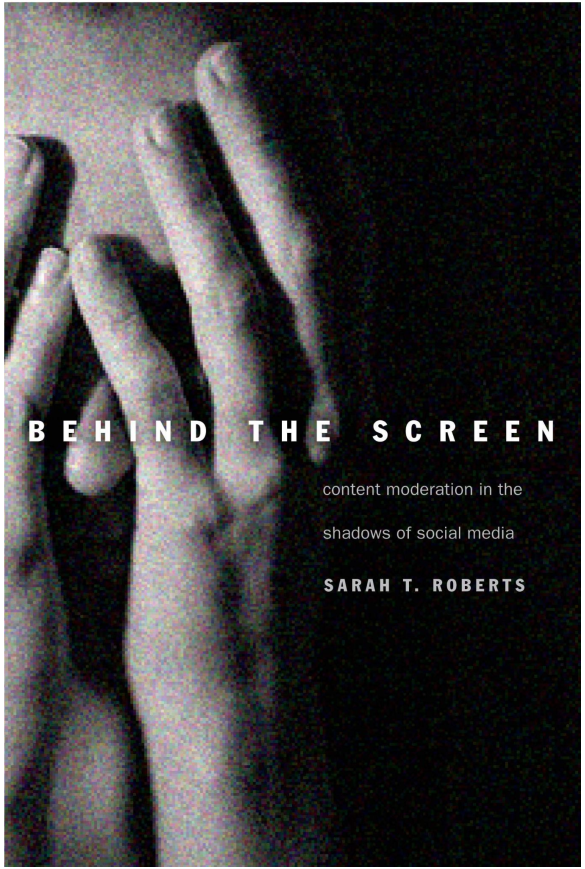
Content moderation

Content curation

Content moderation

Basics of content moderation

- Volume
- Hard, traumatizing work
- AI seems well-suited?



An AI problem

Given content, predict whether it violates standards…

…but how would you actually build this?

And how would you measure if it’s working?

Not just an AI problem

01/21/2022

Sen. Ted Cruz: Big Tech Censorship is the Greatest Threat to Free Speech in America

Where is the difference between moderation and censorship on tech platforms?

We ask Eric Berkowitz, author of "Dangerous Ideas" and human-rights lawyer

A Twitter censorship conundrum for Elon Musk, champion of free speech

- Musk's biggest challenge will be how to deal with governments like Beijing, Moscow, Tehran and Kabul that censor Twitter while using it to push anti-US propaganda
- To truly protect freedom of speech, he may have to block the government-linked accounts of any country that doesn't respect it

AI and content policy

- Given a policy, how do we use AI to implement it?
  - Where do we get labeled examples from?
- Is a given policy feasible to implement through AI?
  - More complex and nuanced rules are harder to implement
- How do you measure whether AI is doing what you want it to?
  - Are different groups held to different standards?
  - Can the public trust your measurement?

Bakalar et al., 2021

Example: Misinformation and censorship

Personality Type, as well as Politics, Predicts Who Shares Fake News

Highly impulsive people who lean conservative are more likely to share false news stories. They have a desire to create chaos and won't be deterred by fact-checkers

By many measures, conservatives share more fake news

How do we know if these are "unbiased" measures of misinformation?

Does such a measure exist?

And it it does, how would you convince a skeptic?

Content curation

Basics of content curation

- Lots of content out there
- You don't want to see most of it
- Platforms have to decide for you
- This is "The Algorithm"

Soft censorship, bias, etc.

- Content curation presents **similar problems to moderation**
  - Which voices get elevated & suppressed?
  - And who decides?
- Measurement is similarly difficult

Is this the full story?

Engagement optimization

Engagement optimization

- The basis for how most platforms operate
- Fundamental assumption: people do what they want to do
- Therefore, we should measure and learn from behavior
- Behavioral data is ubiquitous

Is this a good assumption?

Measuring behavior isn't enough

Want vs. should

- Online groceries vs. in person [Milkman, Rogers, Bazerman 2010]
  - "Want": ice cream
  - "Should": vegetables
- Mail-order DVDs vs. streaming [Milkman, Rogers, Bazerman 2009]
  - "Want": action movies
  - "Should": documentaries

How would you measure "should" for online platforms?

Kleinberg, Mullainathan, and Raghavan 2022

Ongoing efforts

How do we learn whether users truly think content is good for them?

- Do survey responses match behavior?
- Can we learn from interventions (app crashes, break reminders, ...)?
- What else should we be measuring?

Kleinberg, Mullainathan, and Raghavan 2022

Caveats

Goodhart's Law

Scope


Link to bodycam_out.mp4
03/31/2017
20:14:22
Camera 1
ID: 22
ID: 19
ID: 18
ID: 16
ID: 15
ID: 14
ID: 13
ID: 11
ID: 9
ID: 6
ID: 5
ID: 1
Males
Females
SONY
NVIDIA

Quantifiability

Measurement and metrics are necessary, but require caution

Thanks!

mraghavan@g.harvard.edu