Data, Power, and AI Ethics

Emily Denton
Research Scientist, Google Brain
The man at bat readies to swing at the pitch while the umpire looks on.
“The potential of AI”

“Imagine for a moment that you’re in an office, hard at work.

But it’s no ordinary office. By observing cues like your posture, tone of voice, and breathing patterns, it can sense your mood and tailor the lighting and sound accordingly. Through gradual ambient shifts, the space around you can take the edge off when you’re stressed, or boost your creativity when you hit a lull. Imagine further that you’re a designer, using tools with equally perceptive abilities: at each step in the process, they riff on your ideas based on their knowledge of your own creative persona, contrasted with features from the best work of others.”

[Landay (2019). “Smart Interfaces for Human-Centered AI”]
“The potential of AI”

“Imagine for a moment that you’re in an office, hard at work.

But it’s no ordinary office. By observing cues like your posture, tone of voice, and breathing patterns, it can sense your mood and tailor the lighting and sound accordingly. Through gradual ambient shifts, the space around you can take the edge off when you’re stressed, or boost your creativity when you hit a lull. Imagine further that you’re a designer, using tools with equally perceptive abilities: at each step in the process, they riff on your ideas based on their knowledge of your own creative persona, contrasted with features from the best work of others.”

Potential for who?

[Landay (2019). “Smart Interfaces for Human-Centered AI”]
Another future

“Someday you may have to work in an office where the lights are carefully programmed and tested by your employer to hack your body’s natural production of melatonin through the use of blue light, eking out every drop of energy you have while you’re on the clock, leaving you physically and emotionally drained when you leave work. Your eye movements may someday come under the scrutiny of algorithms unknown to you that classifies you on dimensions such as “narcissism” and “psychopathy”, determining your career and indeed your life prospects.”

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research
Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research
Patterns of exclusion: Object recognition

Object classification accuracy dependent on geographical location and household income

DeVries et al. (2019). Does Object Recognition Work for Everyone?

Ground truth: Soap
Nepal, 288 $ / month
Common machine classifications: food, cheese, food product, dish, cooking

Ground truth: Soap
UK, 1890 $ / month
Common classification: soap dispenser, toiletry, faucet, lotion
Patterns of exclusion: Image classification

“Wearing a white mask worked better than using my actual face” -- Joy Buolamwini

Patterns of exclusion: Facial analysis

The Coded Gaze: Unmasking Algorithmic Bias

When the Robot Doesn’t See Dark Skin

By Joy Buolamwini
Ms. Buolamwini is the founder of the Algorithmic Justice League.

June 21, 2018
We’ve seen this before...

Technology has a long history of encoding whiteness as a default

“Shirley cards” calibrated color film for lighter skin tones

Roth (2009). *Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity*
Representational harms: Gender stereotypes in language models

Garg et al. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes
Representational harms: Racial stereotypes in search engines

Ads suggestive of arrest record served for queries of Black-associated names

Sweeney (2013). Discrimination in Online Ad Delivery.
Representational harms: Racial stereotypes in search engines
Discrimination in automated decision making tools: Carceral system

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

<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>

Discrimination in automated decision making tools: Healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer¹,², Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan⁵,⁶

+ See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

NEWS • 24 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms
Discrimination in automated decision making tools: Employment

Why Amazon’s Automated Hiring Tool Discriminated Against Women

By Rachel Goodman, Staff Attorney, ACLU Racial Justice Program

October 12, 2018 | 1:00 PM

Amazon Created a Hiring Tool Using A.I. It Immediately Started Discriminating Against Women.

By Jordan Weissmann

October 10, 2018 • 4:52 PM

Amazon scraps secret AI recruiting tool that showed bias against women

Business News • October 9, 2018 / 11:12 PM / A YEAR AGO
Discrimination in automated decision making tools
AI systems are tools that operate within existing systems of inequality.

One in two American adults is in a law enforcement face recognition network used in unregulated searches employing algorithms with unaudited accuracy.

The Perpetual Line Up
(Garvie, Bedoya, Frankle 2016)

Facial Recognition is the Plutonium of AI

It's dangerous, racializing, and has few legitimate uses; facial recognition needs regulation and control on par with nuclear waste.

By Luke Stark
AI systems are tools that operate within existing systems of inequality.

Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research
“Every data set involving people implies subjects and objects, those who collect and those who make up the collected. It is imperative to remember that on both sides we have human beings.”

- Mimi Onuoha (2016)
Sampling bias

The selected data is not representative of the relevant population.
Buolamwini & Gebru (2018). \textit{Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification}

DeVries et al. (2019). \textit{Does Object Recognition Work for Everyone?}
Sampling bias

Approx 50% of verbs in imSitu visual semantic role labeling (vSRL) dataset are extremely biased in the male or female direction.

[Zhao et al. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints]
Human reporting bias

The frequency with which people write about actions, outcomes, or properties is not a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals.
# Reporting bias

## World learning from text

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“was late”</td>
<td>368,922</td>
</tr>
<tr>
<td>“exhaled”</td>
<td>168,985</td>
</tr>
<tr>
<td>“was punctual”</td>
<td>5,045</td>
</tr>
</tbody>
</table>

Gordon and Van Durme (2013). [Reporting Bias and Knowledge Acquisition](#)
### Reporting bias

#### World learning from text

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency in corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>“spoke”</td>
<td>11,577,917</td>
</tr>
<tr>
<td>“laughed”</td>
<td>3,904,519</td>
</tr>
<tr>
<td>“murdered”</td>
<td>2,834,529</td>
</tr>
<tr>
<td>“inhaled”</td>
<td>984,613</td>
</tr>
<tr>
<td>“breathed”</td>
<td>725,034</td>
</tr>
<tr>
<td>“hugged”</td>
<td>610,040</td>
</tr>
<tr>
<td>“blinked”</td>
<td>390,692</td>
</tr>
<tr>
<td>“was late”</td>
<td>368,922</td>
</tr>
<tr>
<td>“exhaled”</td>
<td>168,985</td>
</tr>
<tr>
<td>“was punctual”</td>
<td>5,045</td>
</tr>
</tbody>
</table>

Gordon and Van Durme (2013). *Reporting Bias and Knowledge Acquisition*
Reporting bias

What do you see?

“Bananas”

“Green bananas”

“Unripe bananas”

[Misra et al. (2016). Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels]
Reporting bias

Social stereotypes can affect implicit prototypicality judgements
Implicit stereotypes

Unconscious attribution of characteristics, traits and behaviours to members of certain social groups.

Data annotation tasks can activate implicit social stereotypes.
Implicit gender stereotypes

Implicit biases can also affect how people classify images

Filter into a computer vision system through annotations

“Doctor”

“Nurse”
Historical bias

Biases that arise from the world as it was when the data was sampled.
If historical hiring practices favor men, gendered cues in the data will be predictive of a ‘successful candidate’
Historical bias

Historical (and ongoing) injustices encoded in datasets
Systemic racism and sexism is foundational to all our major institutions.

Data is generated through social processes and reflects the social world.

‘Unbiased’ data is a myth that obscures the entanglement between tech development and structural inequality.
Policing and surveillance applications

Predictive policing tools predict “crime hotspots” based on policing data that reflects corrupt and racially discriminatory practices of policing and documentation.

Lum & Isaac (2016). To predict and serve?

Richardson et al. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice

Estimated number of drug users, based National Survey on Drug Use and Health

Drug arrests made by Oakland police department
“When bias is routed through technoscience and coded ‘scientific’ and ‘objective’ ... it becomes even more difficult to challenge it and hold individuals and institutions accountable.”

- Ruha Benjamin, Race After Technology
Policing and surveillance applications: Who defines ‘high risk’?

Clifton et al. (2017). *White Collar Crime Risk Zones*
Healthcare applications

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer\textsuperscript{1,2,*}, Brian Powers\textsuperscript{3}, Christine Vogeli\textsuperscript{4}, Sendhil Mullainathan\textsuperscript{5,*,†}

* See all authors and affiliations

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

NEWS  24 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms
“New Jim Code”: ‘race neural’ algorithms that reproduce racial inequality
Datasets construct a particular view of the world -- a view that is often laden with subjective values, judgements, & imperatives

Data is always always socially and culturally situated (Gitelman, 2013; Elish and boyd, 2017)
Datasets construct a particular view of the world -- a view that is often laden with subjective values, judgements, & imperatives

This is inescapable

There is no “view from nowhere” (Haraway, 1991)
The view of the world through ImageNet

“To produce a dataset at ‘the scale of the web’ implies to impose a particular way of seeing images, of pointing and naming.” -- Malevé (2019)
The view of the world through ImageNet

The women of ImageNet ➔ Bikinis and mini-skirts

The men of ImageNet ➔ Music, sports, and fishing

Prabhu & Birhane (2020). Large image datasets: A pyrrhic win for computer vision?
Classifications within machine learning datasets reflect sociotechnical decisions and embed politics, values, and power imbalances.

Data-driven doesn't inherently imply empirically grounded and scientific.
Technologies of human classification

Francis Galton (1877). Composite portraits of human ‘types’

Wu and Zhang (2016). Automated Inference on Criminality using Face Images
Technologies of human classification

Physiognomy’s New Clothes

Aguera y Arcas (2017). Physiognomy’s New Clothes

Jo & Gebru (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning
“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for **profiling people** and revealing their personality based only on their facial image.”

- **Faception** startup

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```

```
```
Datasets represent specific formulations of a problem

Fairness concerns often stem from decisions about how to operationalize social constructs within a datasets (Jacobs and Wallach, 2018)

\[
\begin{align*}
\text{Crime patterns} & \leftrightarrow \text{Policing patterns} \\
\text{Illness} & \leftrightarrow \text{Health care costs} \\
\text{Successful job candidate} & \leftrightarrow \text{Hiring and retention patterns}
\end{align*}
\]
Outline

Part I: Algorithmic (un)fairness

Part II: Data, power, and inequity

Part III: Equitable and accountable AI research
Ethics-informed model testing

Consider **multiple evaluation metrics** - they each provide different information

<table>
<thead>
<tr>
<th>Target</th>
<th>Model Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive (Y = 1)</td>
<td>True positives</td>
</tr>
<tr>
<td>Negative (Y = 0)</td>
<td>False negatives</td>
</tr>
</tbody>
</table>
Ethics-informed model testing

Consider **multiple evaluation metrics** - they each provide different information.

Compute metrics over subgroups defined along cultural, demographic, phenotypical lines.

- How you define groups will be context specific.

Evaluate for each (metric, subgroup) pair.
## Ethics-informed model testing

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Metric</th>
<th>All</th>
<th>F</th>
<th>M</th>
<th>Darker</th>
<th>Lighter</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSFT</strong></td>
<td>PPV(%)</td>
<td>93.7</td>
<td>89.3</td>
<td>97.4</td>
<td>87.1</td>
<td>99.3</td>
<td>79.2</td>
<td>94.0</td>
<td>98.3</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>6.3</td>
<td>10.7</td>
<td>2.6</td>
<td>12.9</td>
<td>0.7</td>
<td><strong>20.8</strong></td>
<td>6.0</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>93.7</td>
<td>96.5</td>
<td>91.7</td>
<td>87.1</td>
<td>99.3</td>
<td>92.1</td>
<td>83.7</td>
<td><strong>100</strong></td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>6.3</td>
<td>8.3</td>
<td>3.5</td>
<td>12.9</td>
<td>0.7</td>
<td><strong>16.3</strong></td>
<td>7.9</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Face++</strong></td>
<td>PPV(%)</td>
<td>90.0</td>
<td>78.7</td>
<td>99.3</td>
<td>83.5</td>
<td>95.3</td>
<td>65.5</td>
<td>99.3</td>
<td>94.0</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>10.0</td>
<td>21.3</td>
<td>0.7</td>
<td>16.5</td>
<td>4.7</td>
<td><strong>34.5</strong></td>
<td>0.7</td>
<td>6.0</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>90.0</td>
<td>98.9</td>
<td>85.1</td>
<td>83.5</td>
<td>95.3</td>
<td>98.8</td>
<td>76.6</td>
<td><strong>98.9</strong></td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>10.0</td>
<td>14.9</td>
<td>1.1</td>
<td>16.5</td>
<td>4.7</td>
<td><strong>23.4</strong></td>
<td>1.2</td>
<td>7.1</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>IBM</strong></td>
<td>PPV(%)</td>
<td>87.9</td>
<td>79.7</td>
<td>94.4</td>
<td>77.6</td>
<td>96.8</td>
<td>65.3</td>
<td>88.0</td>
<td>92.9</td>
<td><strong>99.7</strong></td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>12.1</td>
<td>20.3</td>
<td>5.6</td>
<td>22.4</td>
<td>3.2</td>
<td><strong>34.7</strong></td>
<td>12.0</td>
<td>7.1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>87.9</td>
<td>92.1</td>
<td>85.2</td>
<td>77.6</td>
<td>96.8</td>
<td>82.3</td>
<td>74.8</td>
<td><strong>99.6</strong></td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>12.1</td>
<td>14.8</td>
<td>7.9</td>
<td>22.4</td>
<td>3.2</td>
<td><strong>25.2</strong></td>
<td>17.7</td>
<td>5.20</td>
<td>0.4</td>
</tr>
</tbody>
</table>

[Buolamwini and Gebru, 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*]
### Ethics-informed model testing

Intersectional groups

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Metric</th>
<th>All</th>
<th>F</th>
<th>M</th>
<th>Darker</th>
<th>Lighter</th>
<th>DF</th>
<th>DM</th>
<th>LF</th>
<th>LM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MSFT</strong></td>
<td>PPV(%)</td>
<td>93.7</td>
<td>89.3</td>
<td>97.4</td>
<td>87.1</td>
<td>99.3</td>
<td>79.2</td>
<td>94.0</td>
<td>98.3</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>6.3</td>
<td>10.7</td>
<td>2.6</td>
<td>12.9</td>
<td>0.7</td>
<td>20.8</td>
<td>6.0</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>93.7</td>
<td>96.5</td>
<td>91.7</td>
<td>87.1</td>
<td>99.3</td>
<td>92.1</td>
<td>83.7</td>
<td>100</td>
<td>98.7</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>6.3</td>
<td>8.3</td>
<td>3.5</td>
<td>12.9</td>
<td>0.7</td>
<td>16.3</td>
<td>7.9</td>
<td>1.3</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Face++</strong></td>
<td>PPV(%)</td>
<td>90.0</td>
<td>78.7</td>
<td>99.3</td>
<td>83.5</td>
<td>95.3</td>
<td>65.5</td>
<td>99.3</td>
<td>94.0</td>
<td>99.2</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>10.0</td>
<td>21.3</td>
<td>0.7</td>
<td>16.5</td>
<td>4.7</td>
<td>34.5</td>
<td>0.7</td>
<td>6.0</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>90.0</td>
<td>98.9</td>
<td>85.1</td>
<td>83.5</td>
<td>95.3</td>
<td>98.8</td>
<td>76.6</td>
<td>98.9</td>
<td>92.9</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>10.0</td>
<td>14.9</td>
<td>1.1</td>
<td>16.5</td>
<td>4.7</td>
<td>23.4</td>
<td>1.2</td>
<td>7.1</td>
<td>1.1</td>
</tr>
<tr>
<td><strong>IBM</strong></td>
<td>PPV(%)</td>
<td>87.9</td>
<td>79.7</td>
<td>94.4</td>
<td>77.6</td>
<td>96.8</td>
<td>65.3</td>
<td>80.0</td>
<td>92.9</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>Error Rate(%)</td>
<td>12.1</td>
<td>20.3</td>
<td>5.6</td>
<td>22.4</td>
<td>3.2</td>
<td>34.7</td>
<td>12.0</td>
<td>7.1</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>TPR (%)</td>
<td>87.9</td>
<td>92.1</td>
<td>85.2</td>
<td>77.6</td>
<td>96.8</td>
<td>82.3</td>
<td>74.8</td>
<td>99.6</td>
<td>94.8</td>
</tr>
<tr>
<td></td>
<td>FPR (%)</td>
<td>12.1</td>
<td>14.8</td>
<td>7.9</td>
<td>22.4</td>
<td>3.2</td>
<td>25.2</td>
<td>17.7</td>
<td>5.20</td>
<td>0.4</td>
</tr>
</tbody>
</table>

[Buolamwini and Gebru, 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification]
# Model and data transparency

## Model cards: Standardized framework for transparent model reporting

### Model creators:
Encourage thorough and critical evaluations
Outline potential risks or harms, and implications of use

### Model consumers:
Provide information to facilitate informed decision making

---

**Model Card - Smiling Detection in Images**

<table>
<thead>
<tr>
<th>Model Details</th>
<th>Quantitative Analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Developed by researchers at Google and the University of Toronto, 2018, v1.</td>
<td>- False Positive Rate @ 0.5</td>
</tr>
<tr>
<td>- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.</td>
<td>- 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
</tr>
<tr>
<td><strong>Intended Use</strong></td>
<td>- False Negative Rate @ 0.5</td>
</tr>
<tr>
<td>- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing data points for people who are blind, or assisting applications such as automatically finding smiling photos.</td>
<td>- 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
</tr>
<tr>
<td>- Particularly intended for younger audiences.</td>
<td>- False Discovery Rate @ 0.5</td>
</tr>
<tr>
<td>- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.</td>
<td>- 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
</tr>
<tr>
<td><strong>Factors</strong></td>
<td>- False Omission Rate @ 0.5</td>
</tr>
<tr>
<td>- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.</td>
<td>- 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000</td>
</tr>
<tr>
<td>- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].</td>
<td></td>
</tr>
<tr>
<td><strong>Metrics</strong></td>
<td></td>
</tr>
<tr>
<td>- Evaluation metrics include False Positive Rate and False Negative Rate to measure disproportionate model performance errors across subgroups. False Discovery Rate and False Omission Rate, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]</td>
<td></td>
</tr>
<tr>
<td>- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.</td>
<td></td>
</tr>
<tr>
<td>- These also correspond to metrics in recent definitions of “fairness” in machine learning [cf. 6, 26], where parity across subgroups for different metrics correspond to different fairness criteria.</td>
<td></td>
</tr>
<tr>
<td>- Using confidence intervals calculated with bootstrap resampling.</td>
<td></td>
</tr>
<tr>
<td>- All metrics reported at the 5 decision threshold, where all error types (FPR, FNR, FDR, FPR) are within the same range (8.94 - 14.14).</td>
<td></td>
</tr>
</tbody>
</table>

### Training Data
- CelebA [36], training data split.

### Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

### Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

### Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

---

Model and data transparency

Standardized framework for transparent dataset documentation

Dataset creators:
Reflect on on process of creation, distribution, and maintenance
Making explicit any underlying assumptions
Outline potential risks or harms, and implications of use

Dataset consumers:
Provide information to facilitate informed decision making

Timnit, et al. (2018). Datasheets for datasets
Holland et al. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards
Bender and Friedman (2018). Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science
Measurement and construct validity

Fairness concerns often stem from decisions about how to operationalize social constructs within a dataset (Jacobs and Wallach, 2018)

Crime patterns ↔ Policing patterns
Illness ↔ Health care costs
Successful job candidate ↔ Hiring and retention patterns
As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)
As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)
As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented ([Geiger et al. 2020; Scheuerman et al. 2020](#))

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural ([Scheuerman et al. 2020](#))

- Annotation and labelling is rarely viewed as interpretive work ([Miceli et al. 2020](#))
  - Annotation demographics often underspecified -- annotators presumed interchangeable
As a field, we need to rethink how we develop and use datasets

Currently:

- Data decisions go heavily undocumented (Geiger et al. 2020; Scheuerman et al. 2020)

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)

- Annotation and labelling is rarely viewed as interpretive work (Miceli et al. 2020)
  - Annotation demographics often underspecified -- annotators presumed interchangeable

- Ground truth often presumed to be fact (Aroyo & Welty, 2015; Muller et al. 2019)
As a field, we need to rethink how we develop and use datasets

Currently:
- Data work is heavily undervalued, relative to model work
  - NLP dataset publications devalued within peer-review processes (Heinzerling, 2019); ongoing work indicates similar pattern in computer vision
As a field, we need to rethink how we develop and use datasets

Currently:

● Data work is heavily undervalued, relative to model work
  ○ NLP dataset publications devalued within peer-review processes (Heinzerling, 2019); ongoing work indicates similar pattern in computer vision

● ML curriculums and textbooks don’t treat dataset development as a specialty
  ○ Jo & Gebru, 2020 characterize resulting practices by a laissez faire attitude
As a field, we need to rethink how we develop and use datasets

Contingent → Datasets are contingent on the social conditions of creation

Constructed → Data is not objective; ‘Ground truth’ isn’t truth

Value-laden → Datasets are shaped by patterns of inclusion and exclusion

Our data collection and data use practices should reflect this
Data is contingent, constructed, value-laden

Who is reflected in the data?
What taxonomies are imposed?
How are images categorized?
Who is doing the categorization?

CelebA dataset
AI research is not a value-neutral endeavor

“I’m just an engineer”

“I’m just doing basic research”

Data Science as Political Action
Grounding Data Science in a Politics of Justice

Ben Green
bgreen@g.harvard.edu
Berkman Klein Center for Internet & Society at Harvard University
Harvard John A. Paulson School of Engineering and Applied Sciences
Accountability for the intended and unintended impacts of our work

Status quo is the default, but the status quo is political

“Detachment in the face of history ensures its ongoing codification” -- Ruha Benjamin

Shift focus from intent → impact
Research is contingent and situated -- be attentive to your own positionality

Our social positions in the world and set of experiences shapes and bounds our view of the world; this in turn affects the research questions we pursue and how we pursue them.

Suggested readings:
Kaeser-Chen et al. (2020). Positionality-Aware Machine Learning
Research is contingent and situated -- be attentive to your own positionality

Limits in your knowledge don’t absolve you of responsibility

Voice-to-face synthesis:

Fun application of conditional generative models?

Assistive technology?

Surveillance technology?

Trans-exclusionary technology?

Oh, et al. (2019). [Speech2Face: Learning the Face Behind a Voice](#)
Wen et al. (2019). [Reconstructing faces from voices](#)
Value knowledge and experience of individuals holding marginalized identities

AI development cannot be divorced from the larger social and political landscape

Who gets a say in the development of AI? Who is most likely to experience positive benefit of AI technologies?

Who is marginalized from AI development? Who is most likely to be harmed by AI technologies?
Diversity and inclusion efforts are part and parcels of responsible AI development.

Suggested reading:
West et al. (2019). Discriminating Systems: Gender, Race and Power in AI
80% of AI professors are male

On average, 80% of professors from UC Berkeley, Stanford, UIUC, CMU, UC London, Oxford, and ETH Zurich are male.
Facebook (as of 2018)
- 22% of technical roles filled by women
- 15% of AI researchers were women

Google (as of 2018)
- 21% of technical roles filled by women
- 10% of AI researchers were women

No reported data on trans and non-binary employees, or other gender minorities

Facebook (as of 2018)
❖ 4% Black workers
❖ 5% Hispanic workers

Microsoft (as of 2018)
❖ 4% Black workers
❖ 6% Latinx workers

Google (as of 2018)
❖ 2.5% Black workers
❖ 3.6% Latinx workers

West et al. (2019). Discriminating Systems: Gender, Race and Power in AI
Minority tax

- Fixing D&I problems
- Calling out unethical practices
Interrogate how structural racism, sexism, etc. shape academic and industry hiring practices, cultures, and incentive structures.
Value interdisciplinarity and ‘non-technical’ work

Building AI is simultaneously a technical and social endeavour

Racial literacy is important for every AI developer (see Data and Society’s Advancing Racial Literacy in Tech)

Knowledge hierarchies embedded within STEM structure the types of knowledge that is seen as valuable

Lived experiences of individuals experiencing the harms of AI technologies is a form of valuable knowledge
Value knowledge and experience of individuals holding marginalized identities

Those belonging to marginalized groups experience the world in ways that give them access to knowledge that those with the dominant perspective do not

Suggested reading:
Donna Haraway (1988). *Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective*
Patricia Hill Collins (1990). *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*
Value knowledge and experience of individuals holding marginalized identities

Actively follow the perspectives of people in marginalized groups

Listen to your colleagues who have personal experiences with the harms of AI systems

Use your voice and position of power to amplify the voices of marginalized individuals

Learn about design frameworks and organizations that are privilege the perspectives of marginalized stakeholders and are leveraging data to empower marginalized communities (e.g. Design Justice Network, Our Data Bodies, Data for Black Lives)
Thanks!

Emily Denton
dentone@google.com
@cephaloponderer