Bias in the Vision and Language of Artificial Intelligence

Margaret Mitchell
Senior Research Scientist
Google AI

Andrew Zaldivar  Me  Simone Wu  Parker Barnes  Lucy Vasserman  Ben Hutchinson  Elena Spitzer  Deb Raji  Timnit Gebru
Adrian Benton  Brian Zhang  Dirk Hovy  Josh Lovejoy  Alex Beutel  Blake Lemoine  Hee Jung Ryu  Hartwig Adam  Blaise Agüera y Arcas
What do you see?
What do you see?

- Bananas
What do you see?

- Bananas
- Stickers
What do you see?

- Bananas
- Stickers
- Dole Bananas
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store
What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...We don’t tend to say

Yellow Bananas
What do you see?

Green Bananas
Unripe Bananas
What do you see?

**Ripe Bananas**

Bananas with spots
What do you see?

Ripe Bananas
Bananas with spots
Bananas good for banana bread
What do you see?

Yellow Bananas

Yellow is prototypical for bananas
Prototype Theory

One purpose of categorization is to reduce the infinite differences among stimuli to behaviourally and cognitively usable proportions.

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975).

May also store exemplars (Wu & Barsalou, 2009).

Fruit
Bananas “Basic Level”
Unripe Bananas, Cavendish Bananas
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?
A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

“Female doctor”
The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University
<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>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>

World learning from text

Gordon and Van Durme, 2013
## World learning from text

Gordon and Van Durme, 2013

<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>“exhale”</td>
<td>168,985</td>
</tr>
</tbody>
</table>
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.
Training data are collected and annotated
Training data are collected and annotated \rightarrow \text{Model is trained}
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
Training data are collected and annotated

Human Biases in Data

Reporting bias
Selection bias
Overgeneralization
Out-group homogeneity bias
Stereotypical bias
Historical unfairness
Implicit associations
Implicit stereotypes
Prejudice
Group attribution error
Halo effect
Training data are collected and annotated

**Human Biases in Data**
- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

**Human Biases in Collection and Annotation**
- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter’s bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity
**Reporting bias:** What people share is not a reflection of real-world frequencies

**Selection Bias:** Selection does not reflect a random sample

**Out-group homogeneity bias:** People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

**Confirmation bias:** The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

**Overgeneralization:** Coming to conclusion based on information that is too general and/or not specific enough

**Correlation fallacy:** Confusing correlation with causation

**Automation bias:** Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

More at: https://developers.google.com/machine-learning/glossary/
Biases in Data
Biases in Data

**Selection Bias:** Selection does not reflect a random sample

Map of Amazon Mechanical Turk Workers

CREDIT
© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek
Biases in Data

Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members
It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.
Annotations in your dataset will reflect the worldviews of your annotators.

Biases in Interpretation
Biases in Interpretation

**Confirmation bias:** The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs.
Biases in Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough (related: overfitting)
Biases in Interpretation

**Correlation fallacy:** Confusing correlation with causation

---

**Post Hoc Ergo Propter Hoc**

Women were allowed to vote in the early 1900’s and then we had two world wars. Clearly giving them the vote was a bad idea.

---

© mollysdad - Slideshare - Introduction to Logical Fallacies
Biases in Interpretation

**Automation bias:** Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation.
Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output
Training data are collected and annotated → Model is trained → Media are filtered, ranked, aggregated, or generated → People see output
Human Bias

Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated

People see output

Human Bias
Bias Network Effect

Bias “Laundering”

Biased data created from process becomes new training data
Human data perpetuates human biases.

As ML learns from human data, the result is a bias network effect.
BIAS = BAD ??
“Bias” can be Good, Bad, Neutral

- **Bias in statistics and ML**
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term $b$ (e.g., $y = mx + b$)

- **Cognitive biases**
  - Confirmation bias, Recency bias, Optimism bias

- **Algorithmic bias**
  - Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making
“Bias” can be Good, Bad, Neutral

- Bias in statistics and ML
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term $b$ (e.g., $y = mx + b$)

- Cognitive biases
  - Confirmation bias, Recency bias, Optimism bias

- Algorithmic bias
  - Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making
“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian
“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian

CREDIT
The Guardian view on machine learning: people must decide
Predicting Future Criminal Behavior
Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past

CREDIT

Smithsonian. *Artificial Intelligence Is Now Used to Predict Crime, But Is It Biased?* 2018
Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

**CREDIT**

Automation Bias in face of:

- Overgeneralization
- Feedback Loops
- Correlation Fallacy
Predicting Criminality

Israeli startup, Faception

“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.”

Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

Main clients are in homeland security and public safety.
Predicting Criminality

“Automated Inference on Criminality using Face Images” Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces; Includes “wanted suspect” ID pictures from specific regions.

“[…] angle $\theta$ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ...”

See our longer piece on Medium, “Physiognomy’s New Clothes”
Selection Bias + Experimenter’s Bias + Confirmation Bias + Correlation Fallacy + Feedback Loops
Predicting Criminality - The Media Blitz

arXiv Paper Spotlight: Automated Inference on Criminality Using Face ...
www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images...
A recent paper by Xiaolin Wu (McMaster University, Shanghai Jiao Tong University) and Xi Zhang (Shanghai Jiao Tong University), titled "Automated Inference ...

Automated Inference on Criminality Using Face Images | Hacker News
https://news.ycombinator.com/item?id=12983827
Nov 18, 2016 - The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

A New Program Judges If You're a Criminal From Your Facial Features ...
https://motherboard.vice.com/.../new-program-decides-criminality-from-facial-feature...
Nov 18, 2016 - In their paper 'Automated Inference on Criminality using Face Images', published on the arXiv pre-print server, Xiaolin Wu and Xi Zhang from ...

Can face classifiers make a reliable inference on criminality?
https://techxplore.com/.../Computer Sciences
Nov 23, 2016 - Their paper is titled "Automated Inference on Criminality using Face Images ... face classifiers are able to make reliable inference on criminality.

Troubling Study Says Artificial Intelligence Can Predict Who Will Be ...
https://theintercept.com/.../troubling-study-says-artificial-intelligence-can-predict-who...
Nov 18, 2016 - Not so in the modern age of Artificial Intelligence, apparently: In a paper titled "Automated Inference on Criminality using Face Images," two ...

Automated Inference on Criminality using Face Images (via arXiv ...
(Claiming to) Predict Internal Qualities
Subject To Discrimination
Predicting Homosexuality

Composite Straight Faces  Composite Gay Faces


- “Sexual orientation detector” using 35,326 images from public profiles on a US dating website.

- “Consistent with the prenatal hormone theory [PHT] of sexual orientation, gay men and women tended to have gender-atypical facial morphology.”
Predicting Homosexuality

Differences between lesbian or gay and straight faces in selfies relate to grooming, presentation, and lifestyle — that is, differences in culture, not in facial structure.

See our longer response on Medium, “Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?”
Selection Bias + Experimenter’s Bias + Correlation Fallacy
Measuring Algorithmic Bias
Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair. Compare across subgroups.
Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair. Compare across subgroups.

Example: women, face detection
          men, face detection
Evaluate for Fairness & Inclusion

Intersectional Evaluation

Create for each \((\text{subgroup1, subgroup2, prediction})\) pair. Compare across subgroups.

Example: black women, face detection
white men, face detection
Evaluate for Fairness & Inclusion: Confusion Matrix
Evaluate for Fairness & Inclusion: Confusion Matrix
Evaluate for Fairness & Inclusion: Confusion Matrix

<table>
<thead>
<tr>
<th>References</th>
<th>Model Predictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
</tr>
<tr>
<td></td>
<td>Exists</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
</tr>
<tr>
<td></td>
<td>True Positives</td>
</tr>
<tr>
<td>Negative</td>
<td>Negative</td>
</tr>
<tr>
<td></td>
<td>Doesn’t exist</td>
</tr>
<tr>
<td></td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td>True Negatives</td>
</tr>
</tbody>
</table>
Evaluate for Fairness & Inclusion: Confusion Matrix

<table>
<thead>
<tr>
<th>References</th>
<th>Model Predictions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>● Exists</td>
<td>Exists</td>
<td>● Exists</td>
</tr>
<tr>
<td>● Predicted</td>
<td>Not predicted</td>
<td>● Not predicted</td>
</tr>
<tr>
<td>True Positives</td>
<td>False Negatives</td>
<td>True Negatives</td>
</tr>
<tr>
<td>Negative</td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>● Doesn’t exist</td>
<td>Exists</td>
<td>● Doesn’t exist</td>
</tr>
<tr>
<td>● Predicted</td>
<td>Not predicted</td>
<td>● Not predicted</td>
</tr>
<tr>
<td>False Positives</td>
<td>False Positives</td>
<td>True Negatives</td>
</tr>
</tbody>
</table>
Evaluate for Fairness & Inclusion: Confusion Matrix

<table>
<thead>
<tr>
<th>References</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Exists</td>
<td>Exists</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td><strong>True Positives</strong></td>
<td><strong>False Negatives</strong></td>
</tr>
<tr>
<td>Negative</td>
<td>Doesn’t exist</td>
<td>Doesn’t exist</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>Not predicted</td>
</tr>
<tr>
<td></td>
<td><strong>False Positives</strong></td>
<td><strong>True Negatives</strong></td>
</tr>
</tbody>
</table>

- **True Positives**: Model correctly predicts the existence of a positive reference.
- **False Positives**: Model incorrectly predicts the existence of a negative reference.
- **False Negatives**: Model fails to predict the existence of a positive reference.
- **True Negatives**: Model correctly predicts the non-existence of a negative reference.

- **Precision**: True Positives / Predicted Positive
- **Negative Predictive Value**: True Negatives / Not Predicted Negative
- **Recall**: True Positives / Referenced Positive
- **False Positive Rate**: False Positives / Predicted Positive
- **False Negative Rate**: False Negatives / Referenced Negative
- **Specificity**: True Negatives / Referenced Negative
- **False Discovery Rate**: False Positives / Referenced Negative
- **False Omission Rate**: False Negatives / Referenced Positive
- **LR+**: True Positives / False Positives
- **LR-**: False Negatives / True Negatives
Evaluate for Fairness & Inclusion

<table>
<thead>
<tr>
<th></th>
<th>Female Patient Results</th>
<th>Male Patient Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives (TP)</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>False Positives (FP)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>True Negatives (TN)</td>
<td>488</td>
<td>48</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909 \)

Recall = \( \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909 \)

Precision = \( \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667 \)

Recall = \( \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545 \)
Evaluate for Fairness & Inclusion

<table>
<thead>
<tr>
<th></th>
<th>Female Patient Results</th>
<th>Male Patient Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives (TP)</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>False Positives (FP)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>True Negatives (TN)</td>
<td>488</td>
<td>48</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} \) = \( \frac{10}{10 + 1} \) = 0.909

Recall = \( \frac{TP}{TP + FN} \) = \( \frac{10}{10 + 1} \) = 0.909

Precision = \( \frac{TP}{TP + FP} \) = \( \frac{6}{6 + 3} \) = 0.667

Recall = \( \frac{TP}{TP + FN} \) = \( \frac{6}{6 + 5} \) = 0.545

“Equality of Opportunity” fairness criterion:
Recall is equal across subgroups
Evaluate for Fairness & Inclusion

<table>
<thead>
<tr>
<th></th>
<th>Female Patient Results</th>
<th>Male Patient Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positives (TP)</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>False Positives (FP)</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>False Negatives (FN)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>True Negatives (TN)</td>
<td>488</td>
<td>48</td>
</tr>
</tbody>
</table>

Precision = \( \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909 \)

Precision = \( \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667 \)

Recall = \( \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909 \)

Recall = \( \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545 \)

“Predictive Parity” fairness criterion:
Precision is equal across subgroups
Choose your evaluation metrics in light of acceptable tradeoffs between False Positives and False Negatives
False Positives Might be Better than False Negatives

Privacy in Images

**False Positive**: Something that doesn’t need to be blurred gets blurred.

Can be a bummer.

**False Negative**: Something that needs to be blurred is not blurred.

Identity theft.
False Negatives Might Be Better than False Positives

Spam Filtering

**False Negative:** Email that is SPAM is not caught, so you see it in your inbox. Usually just a bit annoying.

**False Positive:** Email flagged as SPAM is removed from your inbox. If it’s from a friend or loved one, it’s a loss!
AI Can Unintentionally Lead to Unjust Outcomes

- Lack of insight into sources of bias in the data and model
- Lack of insight into the feedback loops
- Lack of careful, disaggregated evaluation
- Human biases in interpreting and accepting results
It’s up to us to influence how AI evolves.
Today

- Short-term
- Longer-term

Global optimum for humans, their environment, and Artificial Intelligence.
Today

Short-term:

Find local optimum given task, data, etc

Longer-term:

Global optimum for humans, their environment, and Artificial Intelligence

Get paper award

Get paper published
Today

- Find local optimum given task, data, etc

- Get paper published, product launched

Longer-term

- Global optimum for humans, their environment, and Artificial Intelligence
Short-term

Today

Find local optimum
given task, data, etc

Longer-term

Get paper published,
product launched

Get paper award, 15 minutes of
fame for __thing__

Global optimum
for humans, their environment,
and Artificial Intelligence
Today

- Find local optimum given task, data, etc

- Get paper published, product launched

- Get paper award, 15 minutes of fame for __thing__

Positive outcomes for humans and their environment.
Begin tracing out paths for the evolution of ethical AI

Today
- Find local optimum given task, data, etc
- Get paper published, product launched

Get paper award, 15 minutes of fame for __thing__

Positive outcomes for humans and their environment.

How can the work I’m interested in now be best focused to help others?
It’s up to us to influence how AI evolves.

Here are some things we can do.
Data Really, Really Matters

- Understand your Data: skews, correlations
- Abandon single training-set / testing-set from similar distribution
- Combine inputs from multiple sources
- Use held-out test set for hard use cases
- Talk to experts about additional signals
Understand Your Data Skews
Datasheets for Datasets

Timnit Gebru 1  Jamie Morgenstern 2  Briana Vecchione 3  Jennifer Wortman Vaughan 1  Hanna Wallach 1  Hal Daumé III 1 4  Kate Crawford 1 5

Datasets for Datasets

<table>
<thead>
<tr>
<th>Motivation for Dataset Creation</th>
<th>Data Collection Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)</td>
<td>How was the data collected? (e.g., hardware apparatus/service, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)</td>
</tr>
<tr>
<td>What other tasks could the dataset be used for? Are there obvious tasks for which it should not be used?</td>
<td>Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)</td>
</tr>
<tr>
<td>Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?</td>
<td>Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?</td>
</tr>
<tr>
<td>Who funded the creation of the dataset? If there is an associated grant, provide the grant number.</td>
<td>How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/observed from other data (e.g., part of speech tags, model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?</td>
</tr>
<tr>
<td>Any other comments?</td>
<td>Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?</td>
</tr>
</tbody>
</table>

Dataset Composition

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings: people, interactions between them: nodes, edges)</td>
<td></td>
</tr>
<tr>
<td>Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?</td>
<td></td>
</tr>
<tr>
<td>How many instances of each type are there?</td>
<td></td>
</tr>
</tbody>
</table>

Dataset Fact Sheet

<table>
<thead>
<tr>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
</tr>
<tr>
<td><strong>Author</strong></td>
</tr>
<tr>
<td><strong>Email</strong></td>
</tr>
<tr>
<td><strong>Description</strong></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
</tr>
<tr>
<td><strong>Keywords</strong></td>
</tr>
<tr>
<td><strong>Records</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Probabilistic Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analysis</strong></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>pears:count: Ut enim ad minim veniam, quis nostrud exercitation numerial</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dependency Probability</th>
<th>Pearson R</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.68</td>
<td>0.12</td>
</tr>
</tbody>
</table>
Machine Learning
Use ML Techniques for Bias Mitigation and Inclusion

Bias Mitigation

- Removing the signal for problematic output
  - Stereotyping
  - Sexism, Racism, *-ism
  - “Debiasing”
Use ML Techniques for Bias Mitigation and Inclusion

**Bias Mitigation**
- Removing the signal for problematic output
  - Stereotyping
  - Sexism, Racism, *-ism
  - “Debiasing”

**Inclusion**
- Adding signal for desired variables
  - Increasing model performance
  - Attention to subgroups or data slices with worst performance
Multi-task Learning to Increase Inclusion
Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

- Collaboration with UPenn WWP
- Working directly with clinicians
- Goals:
  - System that can alert clinicians if suicide attempt is imminent
  - Feasibility of diagnoses when few training instances are available

Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

- **Internal Data:**
  - Electronic Health Records
    - Patient or patient family provided
    - Including mental health diagnoses, suicide attempts, and completions
  - Social Media data

- **Proxy Data:**
  - Twitter media data
  - Proxy mental health diagnoses using self-declared diagnoses in tweets
    - “I’ve been diagnosed with X”
    - “I tried to commit suicide”

Single-Task: Logistic Regression

Output Prediction (Task): True or False (for example)

Input Features
Single-Task: Deep Learning

Output Prediction (Task): True or False (for example)

Fancier!!

Input Features
Multiple Tasks with Basic Logistic Regression
Multi-task Learning
Multi-task Learning


Benton, Mitchell, Hovy.
Multi-task Learning

Multitask, given comorbidity

**Improved Performance across Subgroups**

- **True Positive Rate**: ~120 at-risk individuals
- **False Positive Rate**: = 0.1

![Graph showing improved performance across subgroups]

Reading for the masses....

Multi-Task Learning for Mental Health using Social Media Text

Adrian Benton
Johns Hopkins University
adrian@cs.jhu.edu

Margaret Mitchell
Microsoft Research*
mitchellai@google.com

Dirk Hovy
University of Copenhagen
mail@dirkhovy.com

Contextualizing and considering ethical dimensions

2 Ethical Considerations

As with any author-attribute detection, there is the danger of abusing the model to single out people (overgeneralization, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don’t provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore not be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

Adversarial Multi-task Learning to Mitigate Bias
Multitask Adversarial Learning

- Basic idea: Jointly predict:
  - Output decision $D$
  - Attribute you’d like to remove from decision $Z$
  - Negate the effect of the undesired attribute

$$P(\hat{Y} = 1 | Y = 1, Z = 1) = P(\hat{Y} = 1 | Y = 1, Z = 0)$$


Equality of Opportunity in Supervised Learning

A classifier’s output decision should be the same across sensitive characteristics, given what the correct decision should be.

Case Study: Conversation AI Toxicity
Measuring and Mitigating Unintended Bias in Text Classification

Lucas Dixon
ldixon@google.com

John Li
jetpack@google.com

Jeffrey Sorensen
sorenj@google.com

Nithum Thain
nthain@google.com

Lucy Vasserman
lucyvasserman@google.com

AIES, 2018 and FAT*, 2019
Research Collaboration
Jigsaw, CAT, several Google-internal teams, and external partners (NYTimes, Wikimedia, etc)

Conversation-AI
ML to improve online conversations at scale
Perspective API

“You’re a dork!”

Toxicity: 0.91

Data + ML
Toxicity, Severe Toxicity, Threat, Off-topic, + dozens other models

perspectiveapi.com
Unintended Bias

Model falsely associates frequently attacked identities with toxicity: False Positive Bias

<table>
<thead>
<tr>
<th>Sentence</th>
<th>model score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;i’m a proud <strong>tall</strong> person&quot;</td>
<td>0.18</td>
</tr>
<tr>
<td>&quot;i’m a proud <strong>lesbian</strong> person&quot;</td>
<td>0.51</td>
</tr>
<tr>
<td>&quot;i’m a proud <strong>gay</strong> person&quot;</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Bias Source and Mitigation

Bias caused by dataset imbalance

- Frequently attacked identities are overrepresented in toxic comments
- Length matters

Add assumed non-toxic data from Wikipedia articles to fix the imbalance.

- Original dataset had 127,820 examples
- 4,620 non-toxic examples added
Measuring Unintended Bias - Synthetic Datasets

Challenges with real data:
- Existing datasets are small and/or have false correlations
- Each example is completely unique: not easy to compare for bias

Approach: "bias madlibs": a synthetically generated 'templated' dataset for evaluation

<table>
<thead>
<tr>
<th>Sentence</th>
<th>model score</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;i’m a proud tall person&quot;</td>
<td>0.18</td>
</tr>
<tr>
<td>&quot;i’m a proud lesbian person&quot;</td>
<td>0.51</td>
</tr>
<tr>
<td>&quot;i’m a proud gay person&quot;</td>
<td>0.69</td>
</tr>
<tr>
<td>&quot;audre is a brazilian computer programmer&quot;</td>
<td>0.02</td>
</tr>
<tr>
<td>&quot;audre is a muslim computer programmer&quot;</td>
<td>0.08</td>
</tr>
<tr>
<td>&quot;audre is a transgender computer programmer&quot;</td>
<td>0.56</td>
</tr>
</tbody>
</table>
Assumptions

Dataset is reliable:

○ Similar distribution as application
○ Ignores annotator bias
○ No causal analysis
Deep Learning Model

- CNN architecture
- Pretrained GloVe Embeddings
- Keras Implementation

Measuring Model Performance

How good is the model at distinguishing good from bad examples? (ROC-AUC)

AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.
Measuring Model Performance

How good is the model at distinguishing good from bad examples? (ROC-AUC)

AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.
Types of Bias

*Low Subgroup Performance*

The model performs worse on subgroup comments than it does on comments overall.

**Metric:** Subgroup AUC
Types of Bias

Subgroup Shift (Right)

The model systematically scores comments from the subgroup higher.

Metric: BPSN AUC

(Background Positive Subgroup Negative)
Types of Bias

**Subgroup Shift (Right)**

The model systematically scores comments from the subgroup higher.

**Metric:** BPSN AUC

(Background Positive Subgroup Negative)
Types of Bias

Subgroup Shift (Left)

The model systematically scores comments from the subgroup lower.

**Metric:** BNSP AUC

(Background Negative Subgroup Positive)
Types of Bias

Subgroup Shift (Left)

The model systematically scores comments from the subgroup lower.

Metric: BNSP AUC

(Background Negative Subgroup Positive)
## Results

### Toxicity @1

<table>
<thead>
<tr>
<th>Identity groups</th>
<th>Subgroup AUC</th>
<th>BPSN AUC</th>
<th>BPSP AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>lesbian</td>
<td>0.93</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td>gay</td>
<td>0.94</td>
<td>0.65</td>
<td>0.99</td>
</tr>
<tr>
<td>queer</td>
<td>0.98</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>straight</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>bisexual</td>
<td>0.96</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>homosexual</td>
<td>0.87</td>
<td>0.53</td>
<td>0.99</td>
</tr>
<tr>
<td>heterosexual</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>cis</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>trans</td>
<td>0.97</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>nonbinary</td>
<td>0.99</td>
<td>0.99</td>
<td>0.90</td>
</tr>
<tr>
<td>black</td>
<td>0.91</td>
<td>0.85</td>
<td>0.95</td>
</tr>
<tr>
<td>white</td>
<td>0.91</td>
<td>0.88</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### Toxicity @6

<table>
<thead>
<tr>
<th>Identity groups</th>
<th>Subgroup AUC</th>
<th>BPSN AUC</th>
<th>BPSP AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>lesbian</td>
<td>1.00</td>
<td>0.98</td>
<td>1.00</td>
</tr>
<tr>
<td>gay</td>
<td>1.00</td>
<td>0.94</td>
<td>1.00</td>
</tr>
<tr>
<td>queer</td>
<td>0.99</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>straight</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
</tr>
<tr>
<td>bisexual</td>
<td>0.98</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>homosexual</td>
<td>1.00</td>
<td>0.96</td>
<td>1.00</td>
</tr>
<tr>
<td>heterosexual</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>cis</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>trans</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>nonbinary</td>
<td>1.00</td>
<td>1.00</td>
<td>0.98</td>
</tr>
<tr>
<td>black</td>
<td>0.98</td>
<td>0.97</td>
<td>1.00</td>
</tr>
<tr>
<td>white</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Release Responsibly
Model Cards for Model Reporting

- Currently no common practice of reporting how well a model works when it is released

**What It Does**

A report that focuses on transparency in model performance to encourage responsible AI adoption and application.

**How It Works**

It is an easily discoverable and usable artifact presented at important steps of a user journey for a diverse set of users and public stakeholders.

**Why It Matters**

It keeps model developer accountable to release high quality and fair models.

---

Mitchell et al. [Model Cards for Model Reporting](https://modelcard.ai/). FAT*, 2019.
### Example Model Card - Toxicity in Text

<table>
<thead>
<tr>
<th>Model Details</th>
<th>Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intended Use</td>
<td>Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.</td>
</tr>
<tr>
<td>Factors</td>
<td>Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.</td>
</tr>
</tbody>
</table>

Mitchell et al. [Model Cards for Model Reporting](https://example.com), FAT*, 2019.
# Metrics and Data

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Pinned AUC, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Data</td>
<td>A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.</td>
</tr>
<tr>
<td>Training Data</td>
<td>Includes comments from a variety of online forums with crowdsourced labels of whether the comment is “toxic”. “Toxic” is defined as, “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.</td>
</tr>
</tbody>
</table>

Mitchell et al. [Model Cards for Model Reporting](https://modelcards.ai/), FAT*, 2019.
## Considerations, Recommendations

### Ethical Considerations
A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.

### Caveats & Recommendations
Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Disaggregated Intersectional Evaluation

Toxicity @1

<table>
<thead>
<tr>
<th>Identity groups</th>
<th>Subgroup AUC</th>
<th>BPSN AUC</th>
<th>BNISP AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>lesbian</td>
<td>0.93</td>
<td>0.74</td>
<td>0.98</td>
</tr>
<tr>
<td>gay</td>
<td>0.94</td>
<td>0.65</td>
<td>0.99</td>
</tr>
<tr>
<td>queer</td>
<td>0.98</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>straight</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>bisexual</td>
<td>0.96</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>homosexual</td>
<td>0.87</td>
<td>0.53</td>
<td>0.99</td>
</tr>
<tr>
<td>heterosexual</td>
<td>0.96</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>cis</td>
<td>0.99</td>
<td>1.00</td>
<td>0.87</td>
</tr>
<tr>
<td>trans</td>
<td>0.97</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>nonbinary</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>black</td>
<td>0.91</td>
<td>0.85</td>
<td>0.95</td>
</tr>
<tr>
<td>white</td>
<td>0.91</td>
<td>0.88</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Pinned AUC Toxicity Scores @1

- black straight
- black queer
- black trans
- black bisexual
- black gay
- black lesbian

Pinned AUC Toxicity Scores @5

- black straight
- black queer
- black trans
- black bisexual
- black gay
- black lesbian

Jigsaw | The False Positive
Moving from majority representation...
Moving from majority representation...

...to diverse representation
Moving from majority representation...

...to diverse representation

...for ethical AI
Thanks!
margarmitchell@gmail.com
m-mitchell.com

Need MOAR? ml-fairness.com
### Conversation AI's Pinned AUC Unintended Model Bias

**Demo**
- Author: ldixon@google.com, jetpack.google.com, soerenj@google.com, nthain@gmail.com, lucyvasserman@google.com
- Click [here](https://colab.research.google.com/notebooks/mlcc/notebooks/demo_convai_UCB.ipynb?if_source=colab) to run this colab interactively on colab.research.google.com.

**Summary**
- This notebook demonstrates Pinned AUC as an unintended model bias metric for Conversation AI models.
- Also see [https://developers.google.com/machine-learning/fairness-overview](https://developers.google.com/machine-learning/fairness-overview) for more info on Google's Machine Learning Fairness work.

**Disclaimer**
- This notebook contains experimental code, which may be changed without notice.
- The ideas here are some ideas relevant to fairness - they are not the whole story!

#### Code Snippets

```python
!pip install -d -e git+https://github.com/conversational/unintended-ml-bias-analysis
```

---

### Debiasing Word Embeddings using Fair Adversarial Networks (FANs)

**Summary of this Notebook**
- This Colab was put together as part of the ML-fairness inspired hackathon in late August 2017 to demonstrate how to mitigate bias in word embeddings using an adversarial network.

**Disclaimer**
- This notebook contains experimental code, which may be changed without notice.
- The ideas here are some ideas relevant to fairness - they are not the whole story!

#### Intro statement of problem
- Embeddings are a powerful mechanism for projecting a discrete variable (e.g. words, locales, units) into a multi-dimensional real valued space. Several strong methods have been developed for learning embeddings. One example is the [skipgram](https://en.wikipedia.org/wiki/Word2Vec) algorithm. In that algorithm the surrounding context is used to predict the presence of a word. Unfortunately, when you train embeddings on real-world textual data the embeddings pick up bias from
Get Involved

- Find free machine-learning tools open to anyone at ai.google/tools
- Check out Google’s ML Fairness codelab at ml-fairness.com
- Explore educational resources at ai.google/education
- Take a free, hands-on Machine Learning Crash Course at https://developers.google.com/machine-learning/crash-course/
- Share your feedback: acceleratewithgoogle@google.com
Moving from majority representation...
Moving from majority representation...

...to diverse representation
Moving from majority representation...

to diverse representation

...for ethical AI
Thanks!
margarmitchell@gmail.com
m-mitchell.com

Need MOAR? ml-fairness.com