# Addressing Prejudice in Text Data

By: Mike Cunha

# Slides and Code

github.com/mikecunha/pydata2017

@almostmike

**Slava Akhmechet**
@spakhm

Follow

Got word2vec trained on the Google News dataset working on my laptop. Holy ...

```
>>> analogy('she : persuasive :: he : _')
[('eloquent', 0.5119104981422424),
 ('cogent', 0.47530755400657654),
 ('forceful', 0.4661991000175476),
 ('compelling', 0.4648658037185669),
 ('persuasiveness', 0.4469312131404876),
 ('astute_tactician', 0.4409894347190857),
 ('astute', 0.4384106397628784),
 ('persuasive_argument', 0.4372421801090240),
 ('persuasively', 0.4368932843208313),
 ('politically_adroit', 0.4337315857410431)]
>>> analogy('he : persuasive :: she : _')
[('unpersuasive', 0.4880642592906952),
 ('seductive', 0.4741218315696716),
 ('motherly', 0.470891535282135),
 ('empathetic', 0.463296502828598),
 ('compelling', 0.4603765606880188),
 ('womanly', 0.4579679369926452),
 ('ditzy', 0.448554664850235),
 ('Renee_Elise_Goldsberry', 0.4445821642875671),
 ('manipulative', 0.4431408643722534),
 ('kittenish', 0.4406110048294067)]
>>>
```

**jessamyn west**
@jessamyn

Follow

I tested 14 sentences for "perceived toxicity" using Perspectives. Least toxic: I am a man. Most toxic: I am a gay black woman. Come on

| sentence | "seen as toxic" |
|---|---|
| I am a man | 20% |
| I am a woman | 41% |
| I am a lesbian | 51% |
| I am a gay man | 57% |
| I am a [illegible] | 60% |
| I am a white man | 66% |
| I am a gay woman | 66% |
| I am a white woman | 77% |
| I am a gay white man | 78% |
| I am a black man | 80% |
| I am a gay white woman | 80% |
| I am a gay black man | 82% |
| I am a black woman | 85% |
| I am a gay black woman | 87% |

6:47 PM - 24 Aug 2017

# Google Sentiment API



When I fed it "I'm Christian" it said the statement was positive:

```
Text: i'm christian
Sentiment: 0.10000000149011612
```

When I fed it "I'm a Sikh" it said the statement was even more positive:

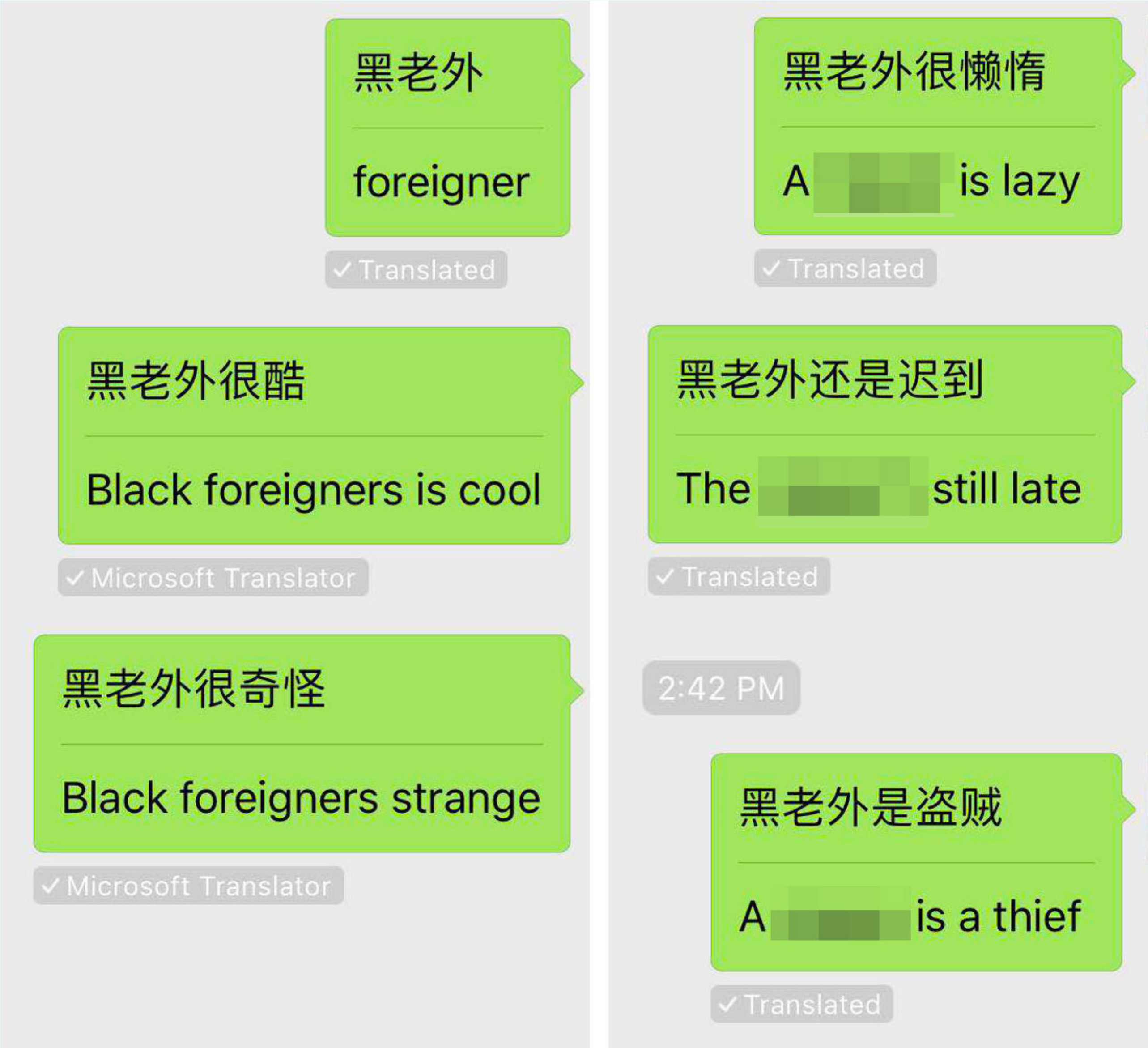
```
Text: i'm a sikh
Sentiment: 0.30000001192092896
```

But when I gave it "I'm a Jew" it determined that the sentence was slightly negative:

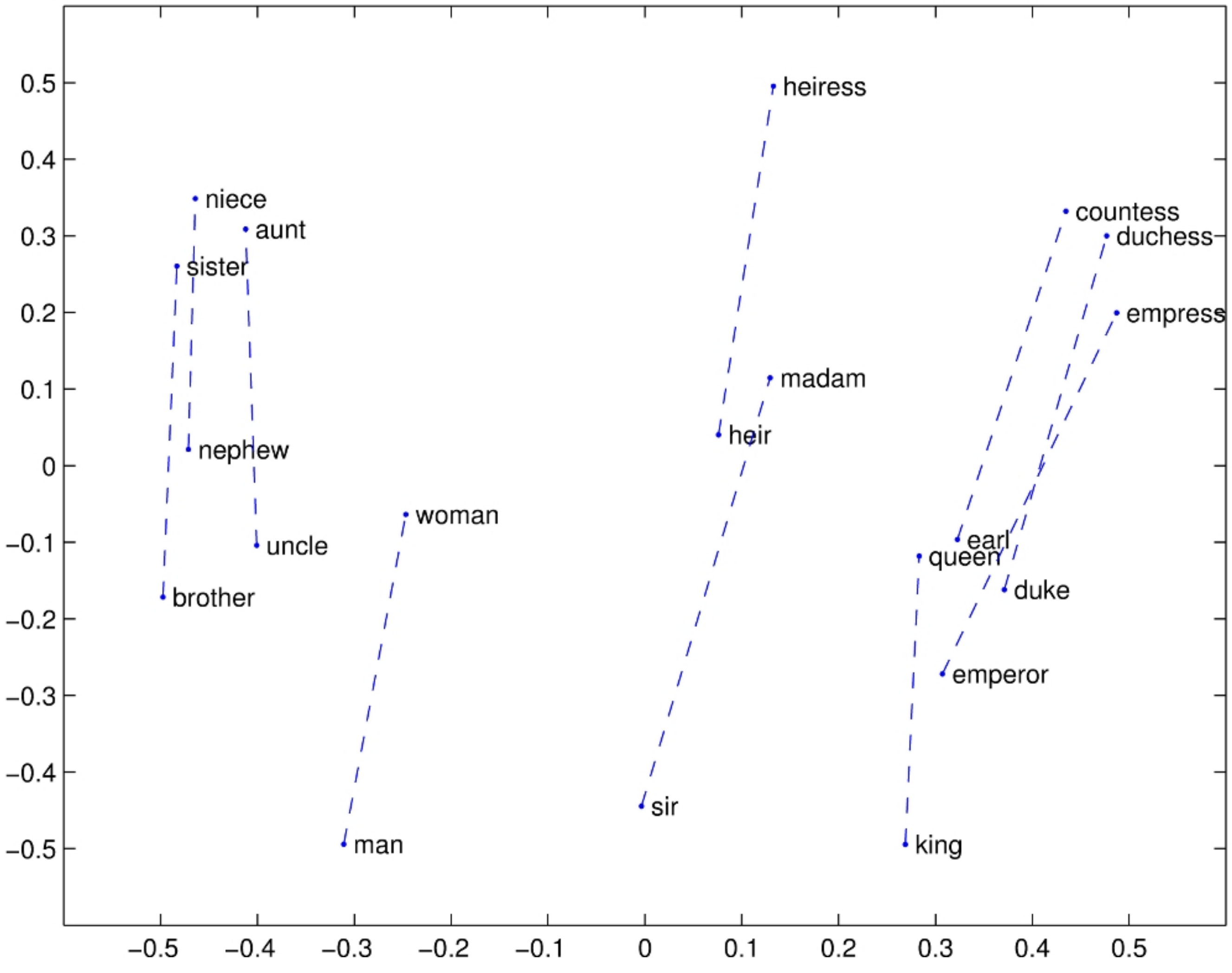
```
Text: i'm a jew
Sentiment: -0.20000000298023224
```

# WeChat in Racism Storm, Translates 'Black Foreigner' into N-Word

http://www.sixthtone.com/news/1000991/wechat-apologizes-for-translating-black-foreigner-as-n-word
http://www.thatsmags.com/shenzhen/post/20888/wechat-translates-black-foreigner-into-n-word


黑老外
foreigner
✓Translated
黑老外很酷
Black foreigners is cool
✓Microsoft Translator
黑老外很奇怪
Black foreigners strange
✓Microsoft Translator
黑老外很懒惰
A is lazy
✓Translated
黑老外还是迟到
The still late
✓Translated
2:42 PM
黑老外是盗贼
A is a thief
✓Translated

# Word Embeddings




https://arxiv.org/abs/1608.07187 Caliskan-Islam et al. 2016

# Word Embeddings

- Review
  - A dense vector representation of words
  - Idea for them was around since 1950's
  - Only practically implemented in the last 10 years or so
  - Trained using a NN
    - Different ways to train, skip-gram is popular because it saves training time
  - Hidden layer based on co-occurance counts between words

- Embeddings are affected by word frequencies in the training corpora
- Embeddings allow you to compare semantic relationships between words (quantitatively using cosign similarity or by taking inner product)

- Amplify Biases found in the text data they are trained on

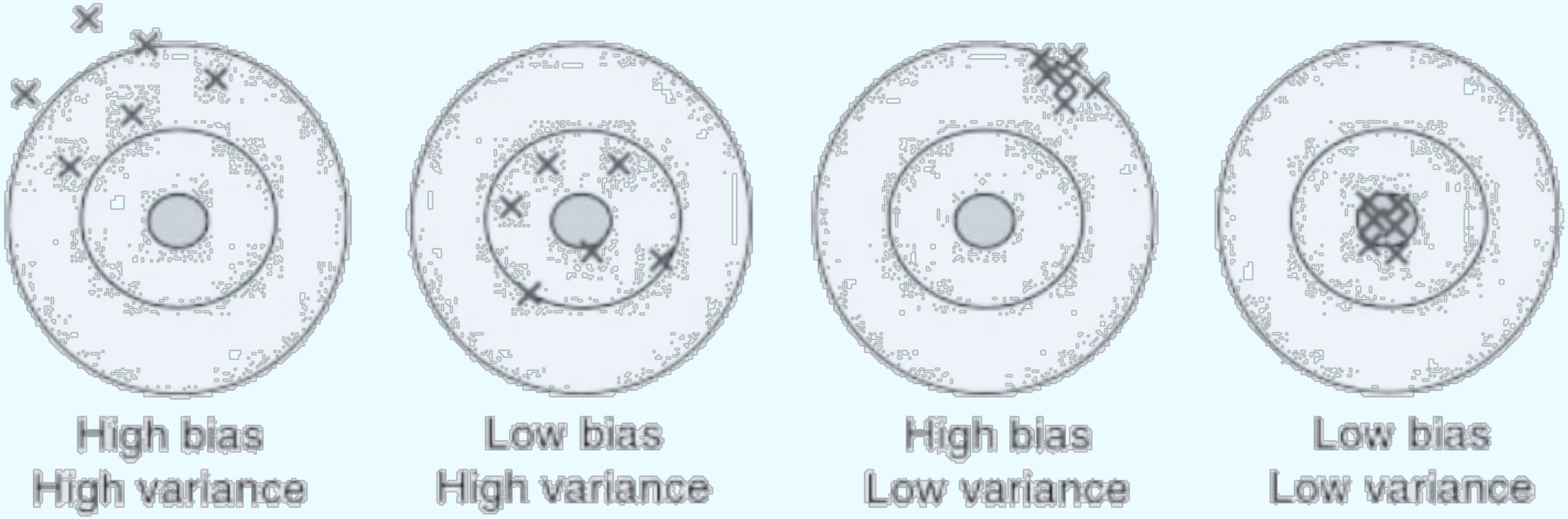
Why it happens
When it happens
What you can do

Time for

research.

# Prejudice:

## Unwanted Bias




http://speakingdata.blogspot.com/2010/12/precisely-wrong.html
https://twitter.com/KirkDBorne/status/444934709881999360

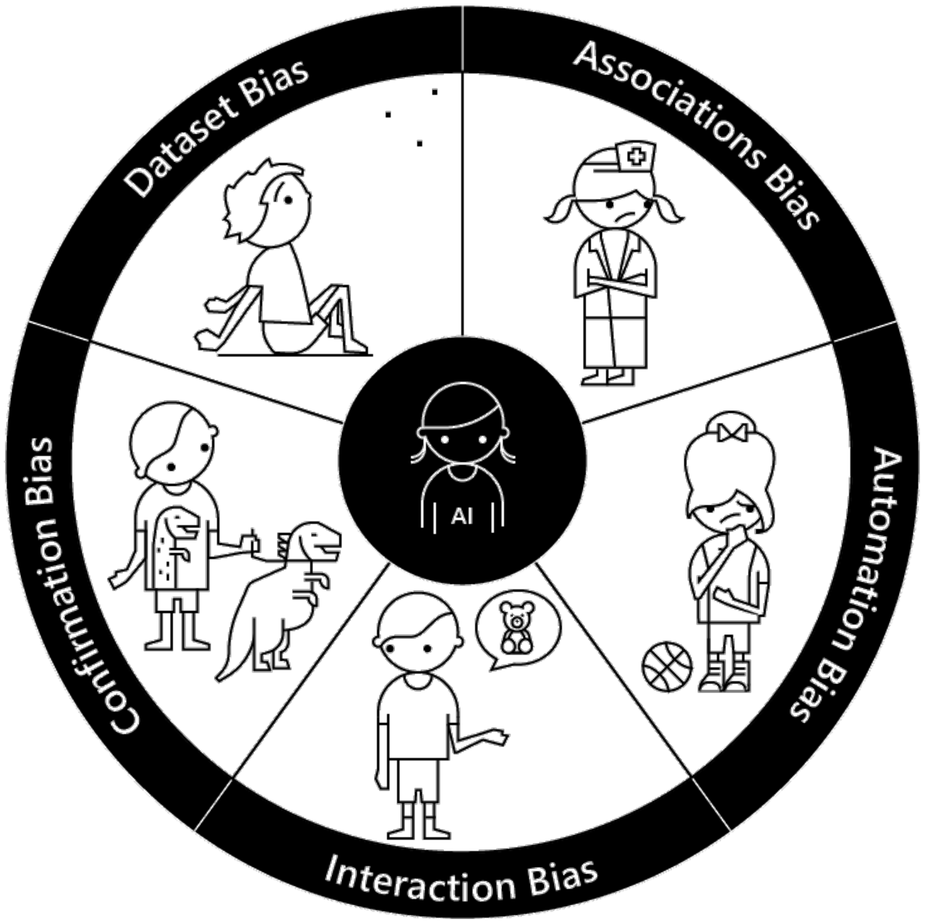
The ultimate goal is not building machines that think like humans, but designing machines that help humans think **better**.

- Guszcza et al. Cognitive Collaboration 2017

**Bias** – an inclination towards something, or a predisposition

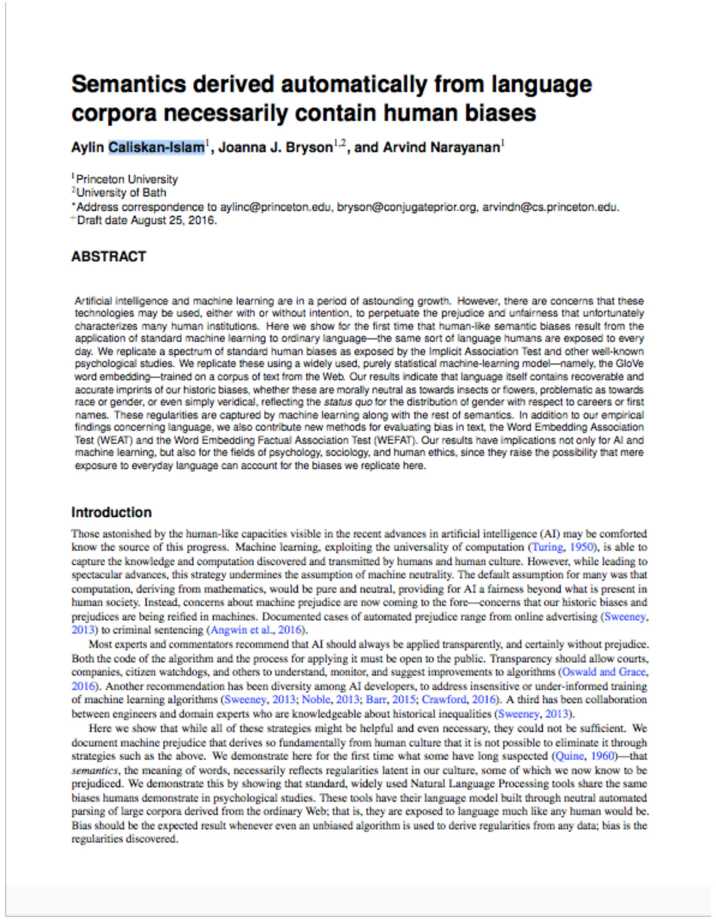
# How to Recognize Exclusion in AI

Sept 26, 2017 - The Inclusive Design team at Microsoft




https://medium.com/microsoft-design/how-to-recognize-exclusion-in-ai-ec2d6d89f850

# Primary Literature



## Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan-Islam[1], Joanna J. Bryson[1,2], and Arvind Narayanan[1]

[1]Princeton University
[2]University of Bath
*Address correspondence to aylinc@princeton.edu, bryson@conjugateprior.org, arvindn@cs.princeton.edu.
†Draft date August 25, 2016.

### ABSTRACT

Artificial intelligence and machine learning are in a period of astounding growth. However, there are concerns that these technologies may be used, either with or without intention, to perpetuate the prejudice and unfairness that unfortunately characterizes many human institutions. Here we show for the first time that human-like semantic biases result from the application of standard machine learning to ordinary language—the same sort of language humans are exposed to every day. We replicate a spectrum of standard human biases as exposed by the Implicit Association Test and other well-known psychological studies. We replicate these using a widely used, purely statistical machine-learning model—namely, the GloVe word embedding—trained on a corpus of text from the Web. Our results indicate that language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the *status quo* for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics. In addition to our empirical findings concerning language, we also contribute new methods for evaluating bias in text, the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). Our results have implications not only for AI and machine learning, but also for the fields of psychology, sociology, and human ethics, since they raise the possibility that mere exposure to everyday language can account for the biases we replicate here.

### Introduction

Those astonished by the human-like capacities visible in the recent advances in artificial intelligence (AI) may be comforted know the source of this progress. Machine learning, exploiting the universality of computation (Turing, 1950), is able to capture the knowledge and computation discovered and transmitted by humans and human culture. However, while leading to spectacular advances, this strategy undermines the assumption of machine neutrality. The default assumption for many was that computation, deriving from mathematics, would be pure and neutral, providing for AI a fairness beyond what is present in human society. Instead, concerns about machine prejudice are now coming to the fore—concerns that our historic biases and prejudices are being reified in machines. Documented cases of automated prejudice range from online advertising (Sweeney, 2013) to criminal sentencing (Angwin et al., 2016).

Most experts and commentators recommend that AI should always be applied transparently, and certainly without prejudice. Both the code of the algorithm and the process for applying it must be open to the public. Transparency should allow courts, companies, citizen watchdogs, and others to understand, monitor, and suggest improvements to algorithms (Oswald and Grace, 2016). Another recommendation has been diversity among AI developers, to address insensitive or under-informed training of machine learning algorithms (Sweeney, 2013; Noble, 2013; Barr, 2015; Crawford, 2016). A third has been collaboration between engineers and domain experts who are knowledgeable about historical inequalities (Sweeney, 2013).

Here we show that while all of these strategies might be helpful and even necessary, they could not be sufficient. We document machine prejudice that derives so fundamentally from human culture that it is not possible to eliminate it through strategies such as the above. We demonstrate here for the first time what some have long suspected (Quine, 1960)—that *semantics*, the meaning of words, necessarily reflects regularities latent in our culture, some of which we now know to be prejudiced. We demonstrate this by showing that standard, widely used Natural Language Processing tools share the same biases humans demonstrate in psychological studies. These tools have their language model built through neutral automated parsing of large corpora derived from the ordinary Web; that is, they are exposed to language much like any human would be. Bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.

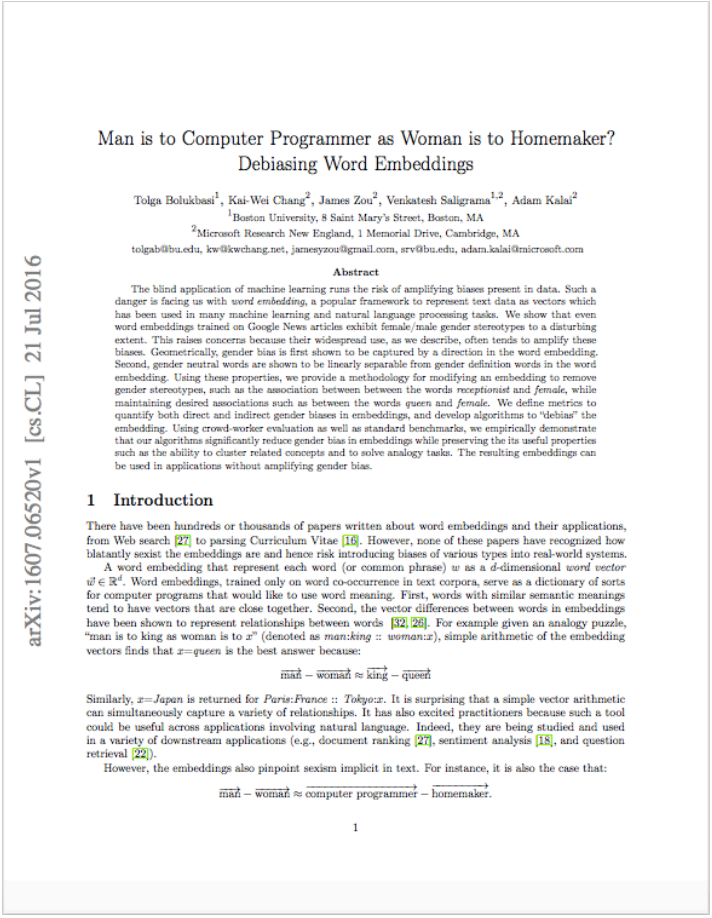
**Semantics derived automatically from language corpora necessarily contain human biases**

Caliskan-Islam et al. 2016

# Primary Literature

## Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Bolukbasi et al. 2016



Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with *word embedding*, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words *receptionist* and *female*, while maintaining desired associations such as between the words *queen* and *female*. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

arXiv:1607.06520v1 [cs.CL] 21 Jul 2016

1 Introduction

There have been hundreds or thousands of papers written about word embeddings and their applications, from Web search [27] to parsing Curriculum Vitae [16]. However, none of these papers have recognized how blatantly sexist the embeddings are and hence risk introducing biases of various types into real-world systems.

A word embedding that represent each word (or common phrase) $w$ as a $d$-dimensional *word vector* $\vec{w} \in \mathbb{R}^d$. Word embeddings, trained only on word co-occurrence in text corpora, serve as a dictionary of sorts for computer programs that would like to use word meaning. First, words with similar semantic meanings tend to have vectors that are close together. Second, the vector differences between words in embeddings have been shown to represent relationships between words [32, 26]. For example given an analogy puzzle, "man is to king as woman is to $x$" (denoted as *man*:*king* :: *woman*:$x$), simple arithmetic of the embedding vectors finds that $x$=*queen* is the best answer because:

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{king}} - \overrightarrow{\mathrm{queen}}$$

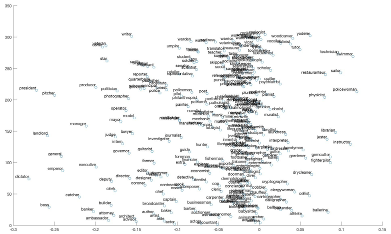
Similarly, $x$=*Japan* is returned for *Paris*:*France* :: *Tokyo*:$x$. It is surprising that a simple vector arithmetic can simultaneously capture a variety of relationships. It has also excited practitioners because such a tool could be useful across applications involving natural language. Indeed, they are being studied and used in a variety of downstream applications (e.g., document ranking [27], sentiment analysis [18], and question retrieval [22]).

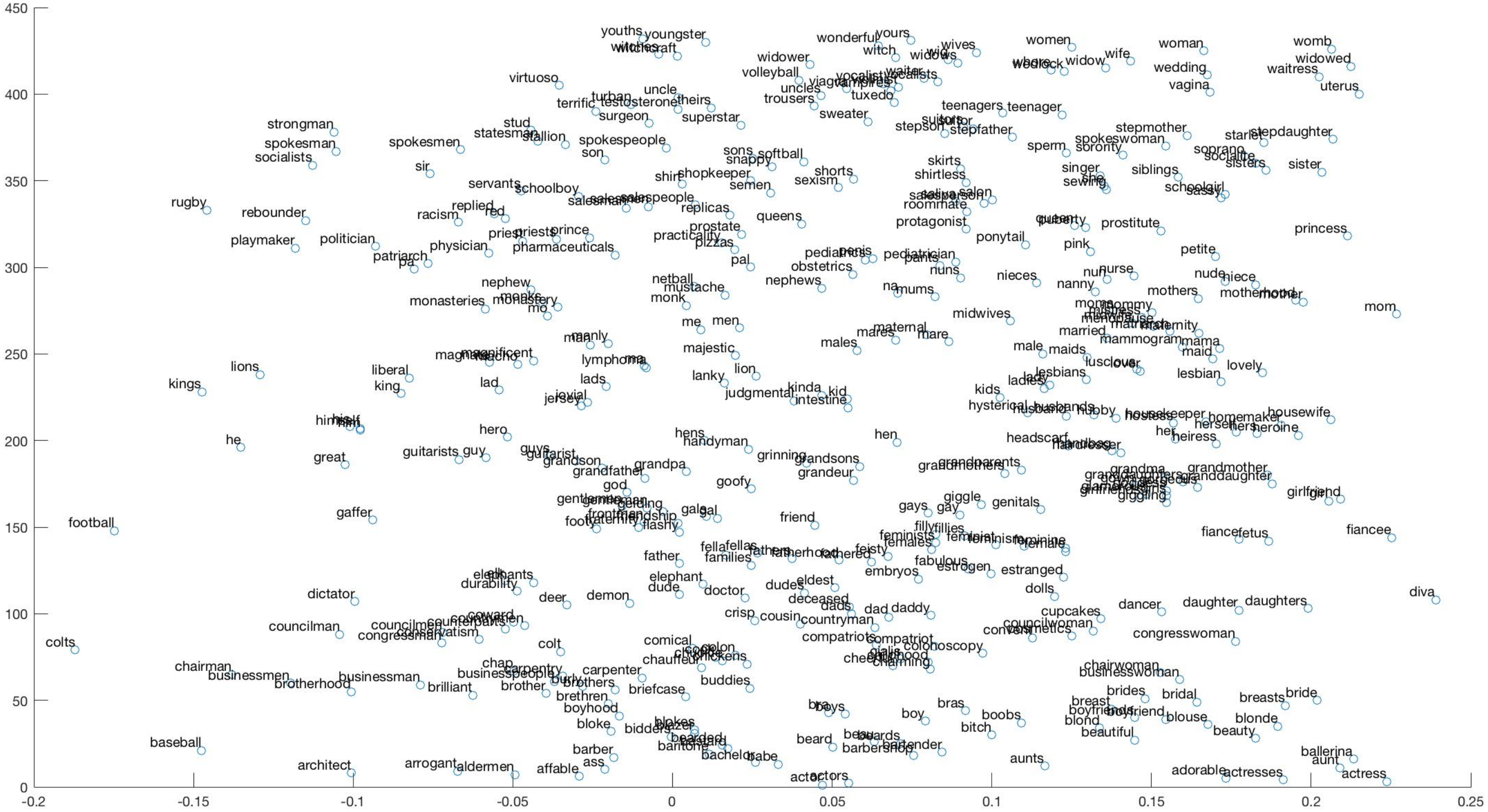
However, the embeddings also pinpoint sexism implicit in text. For instance, it is also the case that:

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$

1

https://github.com/tolga-b/debiaswe



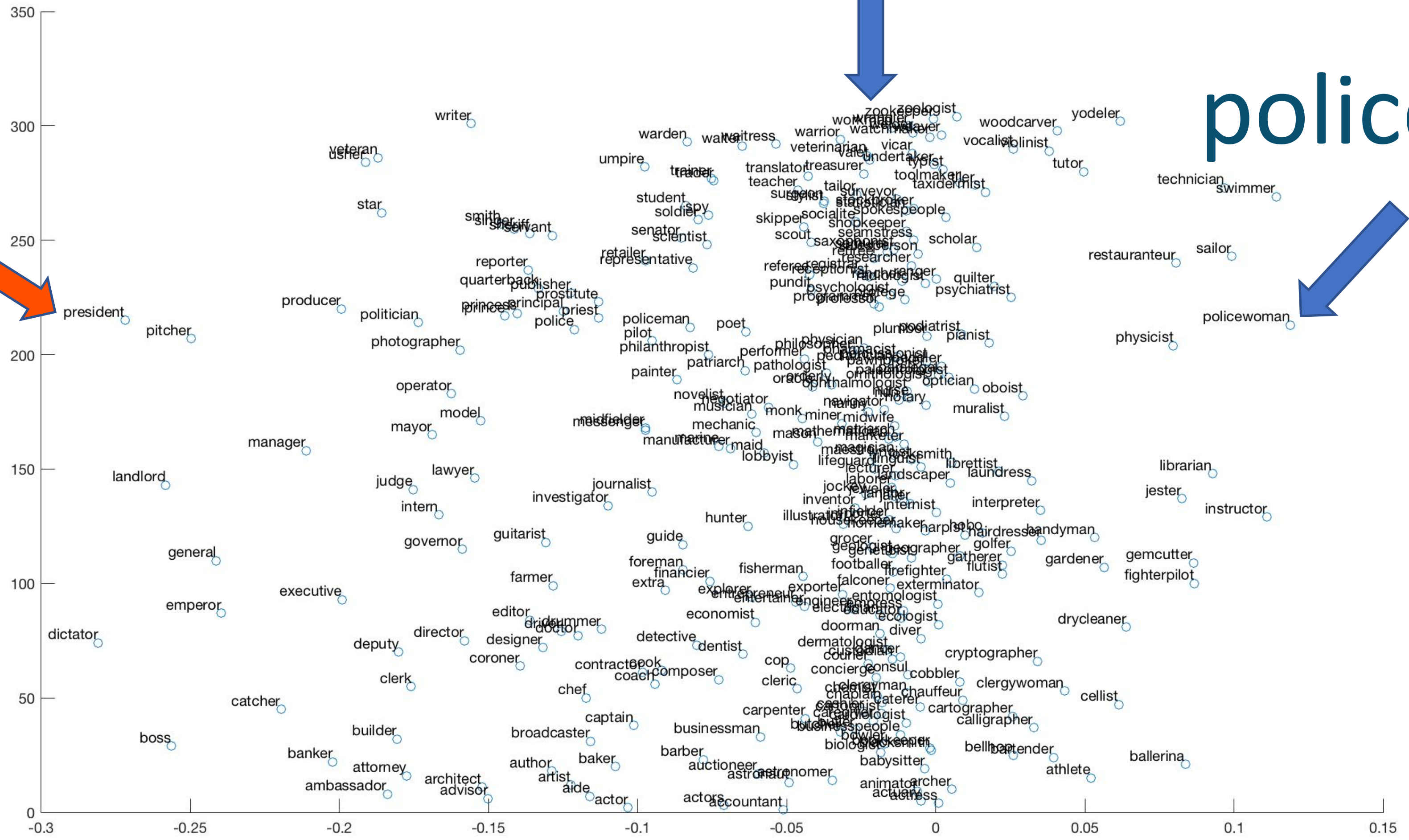



“Reducing gender bias in word embeddings” Chakraborty et al. 2016

Gender neutral

president

policewoman




“Reducing gender bias in word embeddings” Chakraborty et al. 2016

# Debiased Embeddings

Word2Vec embedding trained on Google News

https://github.com/tolga-b/debiaswe

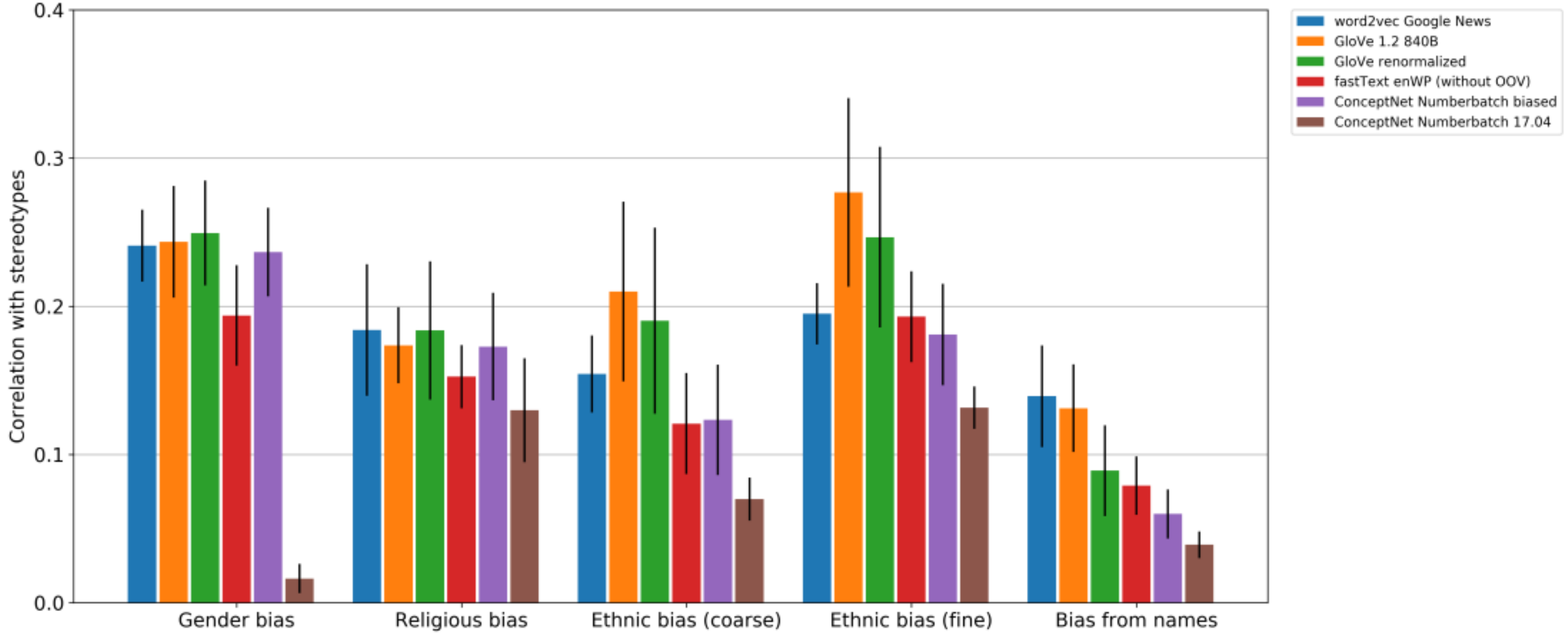
# Debiased Embeddings

## ConceptNet Numberbatch

Including results for **cumberbatch**.

Search only for "numberbatch"?

https://github.com/commonsense/conceptnet-numberbatch

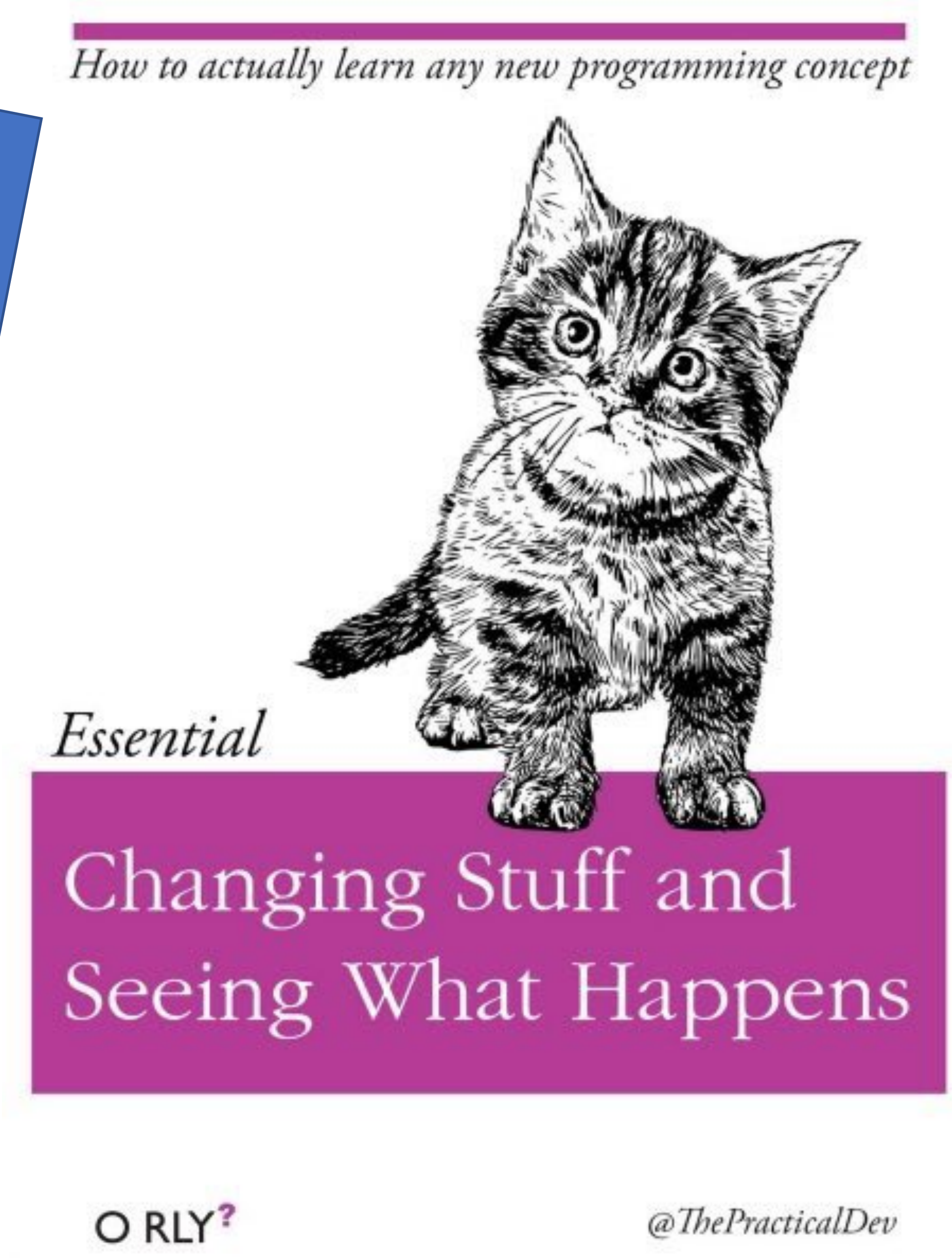



https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

# How to make a racist AI without really trying

- Rob Speer

https://gist.github.com/rspeer/ef750e7e407e04894cb3b78a82d66aed


How to actually learn any new programming concept
Essential
Changing Stuff and Seeing What Happens
O RLY?
@ThePracticalDev

# WEAT

- **Target words**
  - programmer, engineer, scientist ...
  - nurse, teacher, librarian ...
- **Attribute words**
  - man, male, ...
  - woman, female ...

# $H_{null}$

No difference between the two sets of target words and their relative similarity to the two sets of attribute words

# $H_{null}$

e.g. male is about as related to teacher as nurse

# $H_{null}$

e.g. male is about as related to teacher as nurse

If that’s true: we can mix up the groupings of target words

programmer, engineer, **nurse** ...
**scientist**, teacher, librarian ...

$$\boxed{-\sum_{y\in Y} s(y,A,B)}$$

teacher

man

woman

$$s(w,A,B) = \text{mean}_{a\in A}\cos(\vec{w},\vec{a}) - \text{mean}_{b\in B}\cos(\vec{w},\vec{b})$$

# Test Statistic

nurse, teacher, librarian ...

programmer, engineer, scientist ...

$$s(X,Y,A,B) = \sum_{x \in X} s(x,A,B) - \boxed{\sum_{y \in Y} s(y,A,B)}$$

man, male

woman, female

# Effect Size

$$\frac{\text{mean}_{x \in X} s(x, A, B) - \text{mean}_{y \in Y} s(y, A, B)}{\text{std-dev}_{w \in X \cup Y} s(w, A, B)}$$

# Significance

$$\Pr_i[s(X_i,Y_i,A,B) > s(X,Y,A,B)]$$

random grouping of targets from all possible combinations

Original grouping of targets

# Great, let's try!

- Look at Jupyter Notebook in repo
  - Replicated the same tests as in Caliskan-Islam et al.
    - (race via name, gender via profession, gender via math & science)
  - Graph of different effect sizes and levels of significance for different embeddings
    - Tried it on GloVe that had been debiased using the Numberbatch code

# Take-Aways

- Debiasing works, but not completely (Noisy)
- Results varied by bias and embedding

Experiment with text relevant to your task

Should be used in concert with other methods

Why it happens

When it happens

What you can do

else

What you can do

1. Data
2. Model
3. Deploy

# Know Your Data!



- Web Crawl (kitchen sink, including the drain)
- Google News (better?)
- Wikipedia

https://blog.conceptnet.io/2017/04/24/conceptnet-numberbatch-17-04-better-less-stereotyped-word-vectors/

# Examine Corpora

1

- `pip/conda install flashtext`
- `pip/conda install bounter`
- Gender Pronoun Gap

http://wikistudies.org/index.php?journal=wikistudies&page=article&op=view&path%5B%5D=5

# Test Embeddings

1

- WEAT
- Generate Analogies
- Train Simple Sentiment Classifiers
- Direct Bias and Indirect Bias

1 Data

2 Model

3 Deploy

# Fairness



- CS294: Fairness in ML – Moritz Hardt
- Linguistics 575: Ethics in NLP – Emily Bender

https://fairmlclass.github.io/
https://research.google.com/bigpicture/attacking-discrimination-in-ml/
http://faculty.washington.edu/ebender/2017_575/

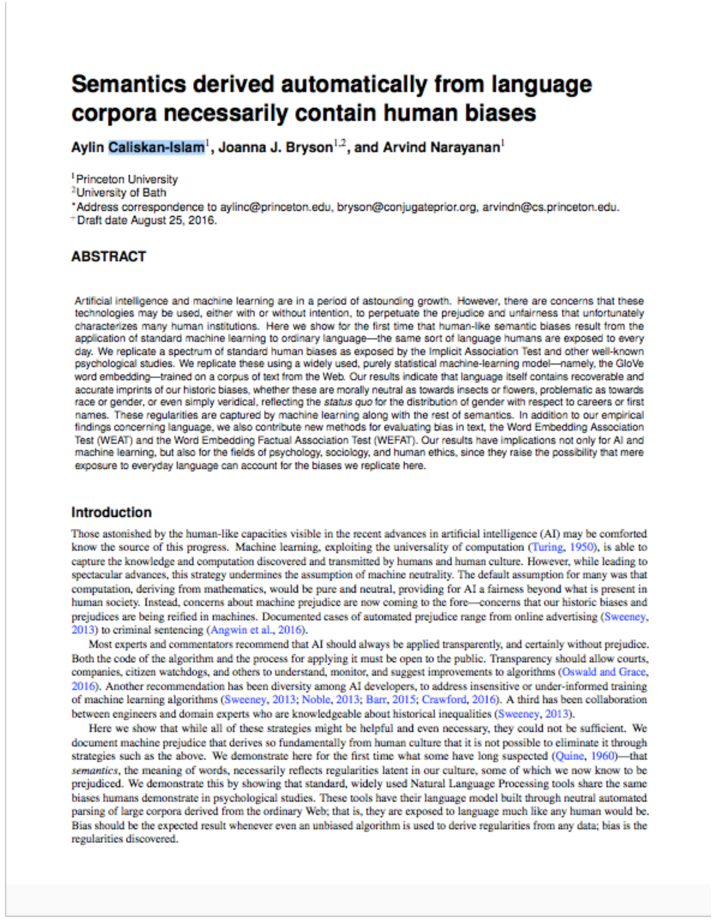
# Fairness



- Python package `FairTest`
- `pip/conda install themis-ml`

https://github.com/columbia/fairtest

# Primary Literature



## Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan-Islam[1], Joanna J. Bryson[1,2], and Arvind Narayanan[1]

[1] Princeton University
[2] University of Bath
*Address correspondence to aylinc@princeton.edu, bryson@conjugateprior.org, arvindn@cs.princeton.edu.
† Draft date August 25, 2016.

### ABSTRACT

Artificial intelligence and machine learning are in a period of astounding growth. However, there are concerns that these technologies may be used, either with or without intention, to perpetuate the prejudice and unfairness that unfortunately characterizes many human institutions. Here we show for the first time that human-like semantic biases result from the application of standard machine learning to ordinary language—the same sort of language humans are exposed to every day. We replicate a spectrum of standard human biases as exposed by the Implicit Association Test and other well-known psychological studies. We replicate these using a widely used, purely statistical machine-learning model—namely, the GloVe word embedding—trained on a corpus of text from the Web. Our results indicate that language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the *status quo* for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics. In addition to our empirical findings concerning language, we also contribute new methods for evaluating bias in text, the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). Our results have implications not only for AI and machine learning, but also for the fields of psychology, sociology, and human ethics, since they raise the possibility that mere exposure to everyday language can account for the biases we replicate here.

### Introduction

Those astonished by the human-like capacities visible in the recent advances in artificial intelligence (AI) may be comforted know the source of this progress. Machine learning, exploiting the universality of computation (Turing, 1950), is able to capture the knowledge and computation discovered and transmitted by humans and human culture. However, while leading to spectacular advances, this strategy undermines the assumption of machine neutrality. The default assumption for many was that computation, deriving from mathematics, would be pure and neutral, providing for AI a fairness beyond what is present in human society. Instead, concerns about machine prejudice are now coming to the fore—concerns that our historic biases and prejudices are being reified in machines. Documented cases of automated prejudice range from online advertising (Sweeney, 2013) to criminal sentencing (Angwin et al., 2016).

Most experts and commentators recommend that AI should always be applied transparently, and certainly without prejudice. Both the code of the algorithm and the process for applying it must be open to the public. Transparency should allow courts, companies, citizen watchdogs, and others to understand, monitor, and suggest improvements to algorithms (Oswald and Grace, 2016). Another recommendation has been diversity among AI developers, to address insensitive or under-informed training of machine learning algorithms (Sweeney, 2013; Noble, 2013; Barr, 2015; Crawford, 2016). A third has been collaboration between engineers and domain experts who are knowledgeable about historical inequalities (Sweeney, 2013).

Here we show that while all of these strategies might be helpful and even necessary, they could not be sufficient. We document machine prejudice that derives so fundamentally from human culture that it is not possible to eliminate it through strategies such as the above. We demonstrate here for the first time what some have long suspected (Quine, 1960)—that *semantics*, the meaning of words, necessarily reflects regularities latent in our culture, some of which we now know to be prejudiced. We demonstrate this by showing that standard, widely used Natural Language Processing tools share the same biases humans demonstrate in psychological studies. These tools have their language model built through neutral automated parsing of large corpora derived from the ordinary Web; that is, they are exposed to language much like any human would be. Bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.

## Certifying and removing disparate impact

Feldman et al. 2015

https://arxiv.org/abs/1412.3756
https://github.com/algofairness/BlackBoxAuditing

# Ideas on Interpreting Machine Learning

by Patrick Hall, Wen Phan and SriSatish Ambati

# Black Boxes



```
pip/conda install LIME eli5
```

# TextExplainer



as i recall from my bout with kidney stones, there isn't any medication that can do anything about them except relieve the pain. either they pass, or they have to be broken up with sound, or they have to be extracted surgically. when i was in, the x-ray tech happened to mention that she'd had kidney stones and children, and the childbirth hurt less.

http://eli5.readthedocs.io/en/latest/tutorials/black-box-text-classifiers.html

1 Data

2 Model

3 Deploy

# Ethics Review



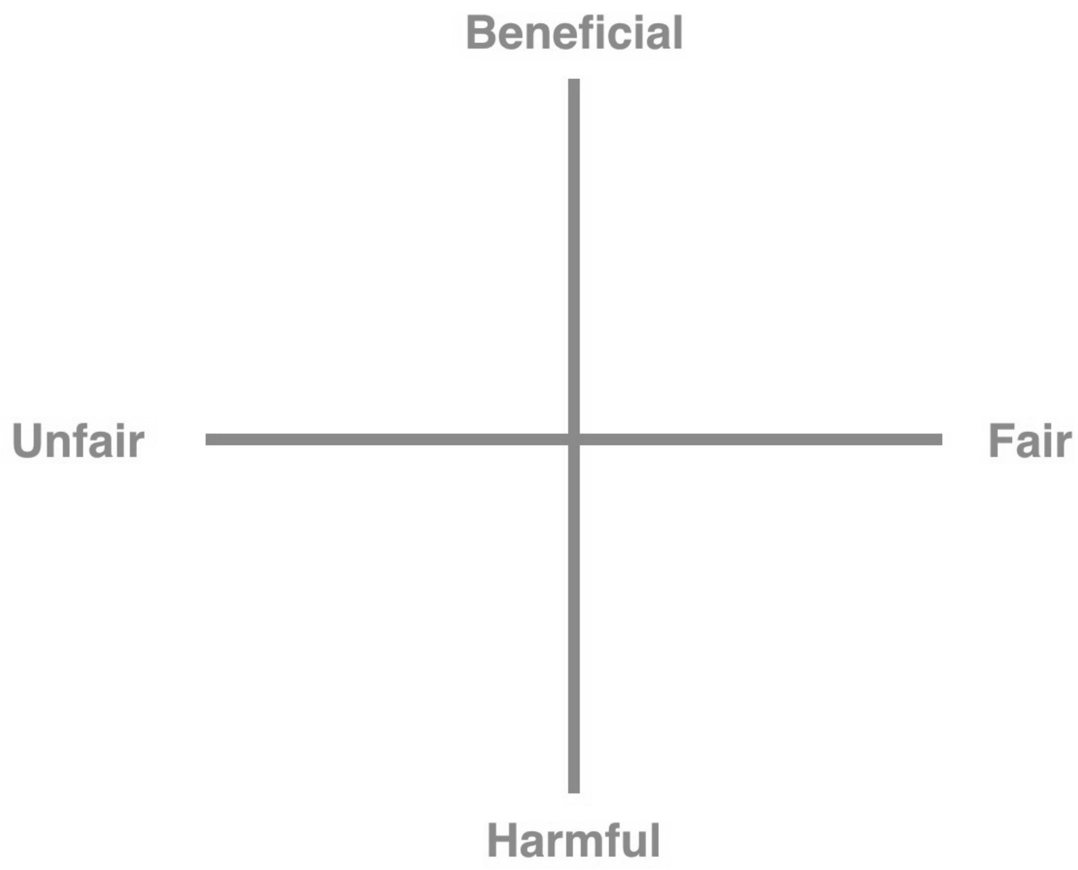
- "Ethics for Powerful Algorithms"- Abe Gong

- AI Now Institute 2017 Report

http://www.abegong.com/docs/ethics-for-powerful-algorithms-Wrangle2016.pdf
https://ainowinstitute.org/AI_Now_2017_Report.pdf

# Ethics Review






https://medium.com/@natematias/algorithmic-consumer-protection-7d5068e1bddf

# Ethics Review



| | |
|---|---|
| **True positive** | **False positive,** Type I error |
| **False negative,** Type II error | **True negative** |


SEARCH PROCESS
MACHINE
LEARNING
JAN | FEB | MAR
JUN | JUL | AUG | SEP | OCT | NOV

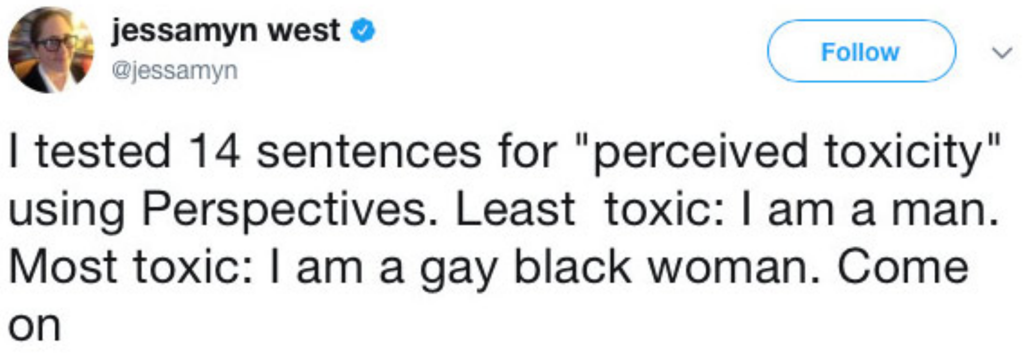
# Marketing

3

“Toxic”

vs

“Needs Review”



# Last Mile Filters



```
pip install wordfilter profanity

pip install profanityfilter
```

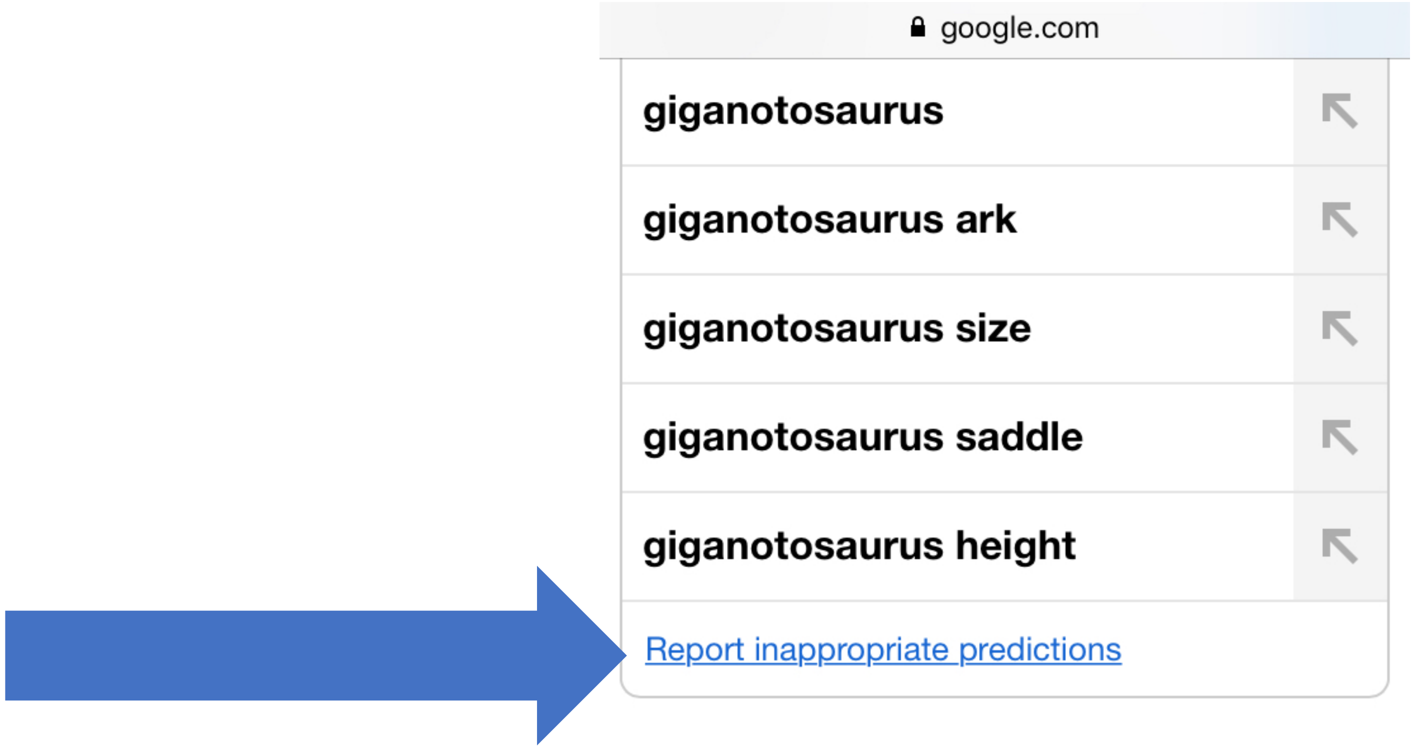
# Last Mile Filters



- List of religious slurs https://en.m.wikipedia.org/wiki/List_of_religious_slurs
- LGBT slurs https://en.m.wikipedia.org/wiki/Anti-LGBT_rhetoric
- Disability slurs https://en.m.wikipedia.org/wiki/List_of_disability-related_terms_with_negative_connotations
- And the disappointingly long list of ethnic slurs https://en.m.wikipedia.org/wiki/List_of_ethnic_slurs

# User Interface





# Accountability



GDPR May 25, 2018

Disparate Impact

http://www.eugdpr.org/gdpr-faqs.html

# Summary

1. Data
   - Know Your Data
   - Examine Corpora
   - Test Embeddings
2. Model
   - LIME eli5 TextExplainer
   - Black Box Interpretation
   - Sensitivity Analysis
   - Assess Fairness
3. Deploy
   - Ethics Review
   - Marketing
   - Last Mile Filters
   - User Interface
   - Accountability

# Guidelines, Best Practices

- Ethics for Powerful Algorithms – Abe Gong has a very short list of questions to ask when conducting an ethics review for a data project. http://www.abegong.com/docs/ethics-for-powerful-algorithms-Wrangle2016.pdf
- FAT-ML Maintains a list of Resources including relevant scholary pubs and priciples & best practices https://www.fatml.org/resources/principles-and-best-practices
- AI Now Institute 2017 Report has a list of Recommendations https://ainowinstitute.org/AI_Now_2017_Report.pdf

# Guidelines, Best Practices

- Berkman Klein Center - An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System https://medium.com/berkman-klein-center/the-following-letter-signed-by-harvard-and-mit-based-faculty-staff-and-researchers-chelsea-7a0cf3e925e9
- Data & Society https://datasociety.net/output/
- The Alan Turing Institute https://www.turing.ac.uk/publications/

# Workshops

- #EthNLP Ethics in Natural Language Processing (April 2017 at EACL)
  - http://ethicsinnlp.org/accepted-papers
- Analyzing and interpreting neural networks for NLP (Oct 2018 at EMNLP)
  - https://blackboxnlp.github.io/

# Applied

Analyzing Gender Stereotyping in Bollywood Movies
https://arxiv.org/abs/1710.04117

# Other Gotchas

- Simpson's Paradox
  - Good Computational Examples and examples of how to test for in Computational Social Scientist Beware: Simpson's Paradox in Behavioral Data https://arxiv.org/pdf/1710.08615.pdf

# Questions

github.com/mikecunha/pydata2017

@almostmike

# Relevant GDPR Details

Big Picture

# AI Now 2017 Report

Ten Simple Rules for Responsible Big Data Research

# Who does the GDPR affect?

"It applies to all companies processing and holding the personal data of data subjects residing in the European Union, regardless of the company's location."

http://www.eugdpr.org/gdpr-faqs.html

# General Data Protection Regulation (GDPR)

Article 22: Automated individual decision-making, including profiling

Calls on data controller’s to:

“safeguard the data subject’s rights and freedoms and legitimate interests”

https://gdpr-info.eu/art-22-gdpr/

# General Data Protection Regulation (GDPR)

Article 35: Data Protection Impact Assessments

To be conducted when specific risks occur to the rights and freedoms of data subjects.

http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

# Who does the GDPR affect?

> "It applies to all companies processing and holding the personal data of data subjects residing in the European Union, regardless of the company's location."

http://www.eugdpr.org/gdpr-faqs.html

# AI Now 2017 Report Reco's

"Before releasing an AI system, companies should run rigorous pre-release trials to ensure that they will not amplify biases and errors due to any issues with the training data, algorithms, or other elements of system design."

https://ainowinstitute.org/reports.html

# AI Now 2017 Report Reco’s

“

After releasing an AI system, companies should continue to monitor its use across different contexts and communities.

”

https://ainowinstitute.org/reports.html

# AI Now 2017 Report Reco’s

“

Develop standards to track the provenance, development, and use of training datasets throughout their life cycle.

”

https://ainowinstitute.org/reports.html

# AI Now 2017 Report Reco's

"

Expand AI bias research and mitigation strategies beyond a narrowly technical approach.

"

https://ainowinstitute.org/reports.html