

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 [redacted] ...

```
>>> analogy('she : persuasive :: he : _')
[('eloquent', 0.5119104981422424),
 ('cogent', 0.47530755400657654),
 ('forceful', 0.4661991000175476),
 ('compelling', 0.4648658037185669),
 ('persuasiveness', 0.44693121314048767),
 ('astute_tactician', 0.4409894347190857),
 ('astute', 0.4384106397628784),
 ('persuasive_argument', 0.43724218010902405),
 ('persuasively', 0.4368932843208313),
 ('politically_adroit', 0.4337315857410431)]
>>> analogy('he : persuasive :: she : _')
[('unpersuasive', 0.4880642592906952),
 ('seductive', 0.47412118315696716),
 ('motherly', 0.470891535282135),
 ('empathetic', 0.463296502828598),
 ('compelling', 0.4603765606880188),
 ('womanly', 0.45796793699264526),
 ('ditzy', 0.448554664850235),
 ('Renee_Elise_Goldsberry', 0.44458216428756714),
 ('manipulative', 0.4431408643722534),
 ('kittenish', 0.44061100482940674)]
>>> 
```



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 [REDACTED]	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



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

Google Sentiment API

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 [REDACTED] is lazy

✓ Translated

黑老外还是迟到

The [REDACTED] still late

✓ Translated

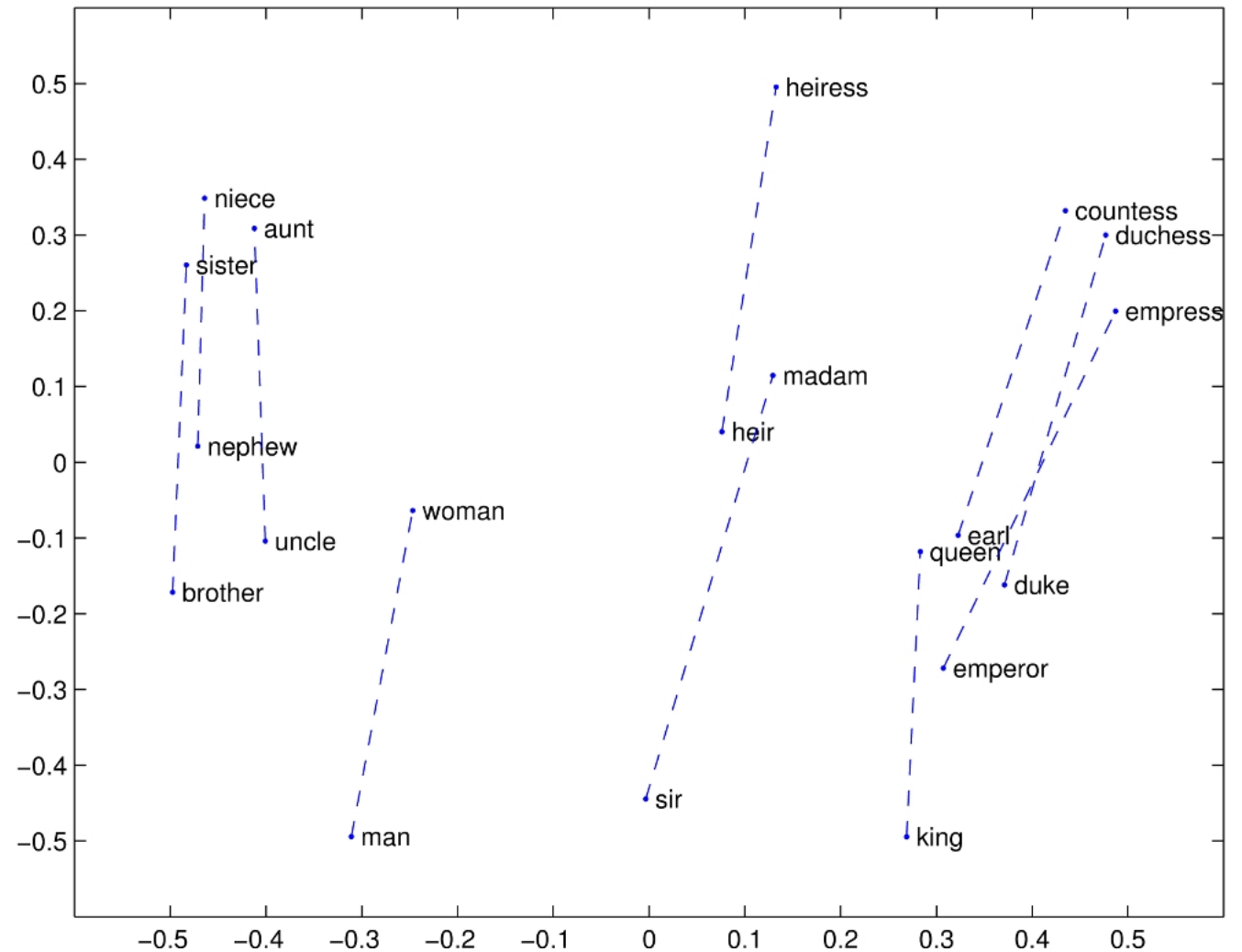
2:42 PM

黑老外是盗贼

A [REDACTED] is a thief

✓ Translated

Word Embeddings



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-occurrence counts between words
- Embeddings are affected by word frequencies in the training corpora
- Embeddings allow you to compare semantic relationships between words (quantitatively using cosine similarity or by taking inner product)
- Amplify Biases found in the text data they are trained on

Why it happens

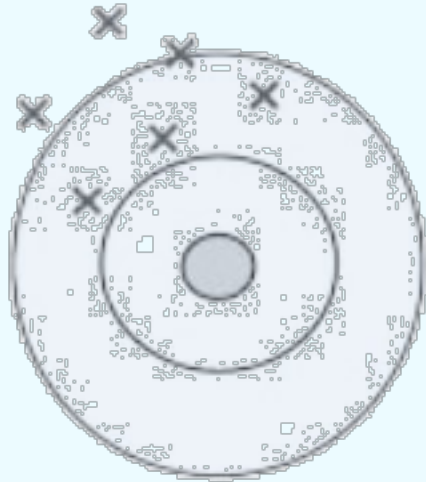
When it happens

What you can do

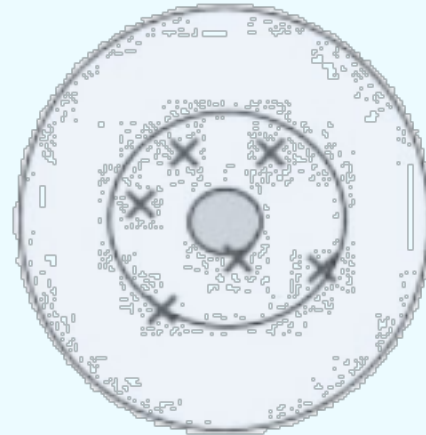
Time for
research.

Prejudice:

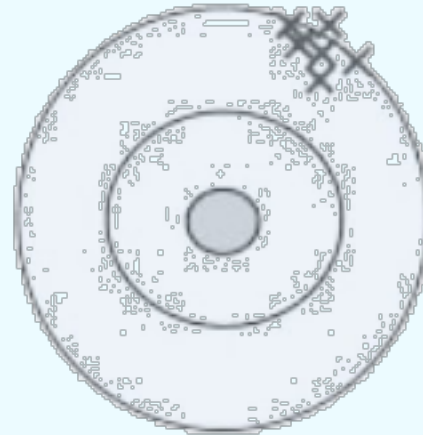
Unwanted Bias



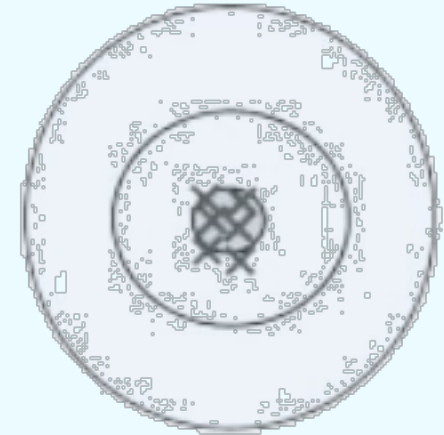
High bias
High variance



Low bias
High variance



High bias
Low variance



Low bias
Low variance

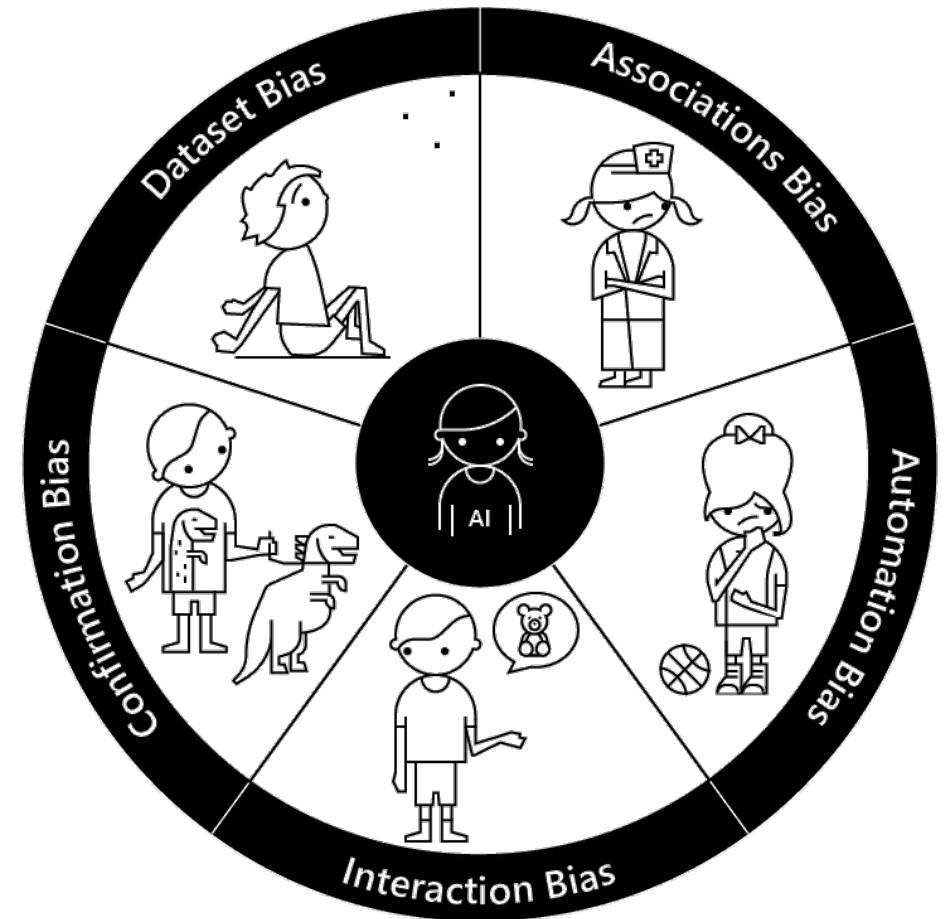
The ultimate goal is not building machines that think like humans, but designing machines that help humans think **better**.

- Guszczka 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



Primary Literature

Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan-Islam¹, Joanna J. Bryson^{1,2}, and Arvind Narayanan¹

¹Princeton University

²University of Bath

*Address correspondence to aylin@princeton.edu, bryson@conjugateprior.org, arvindh@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 to 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).

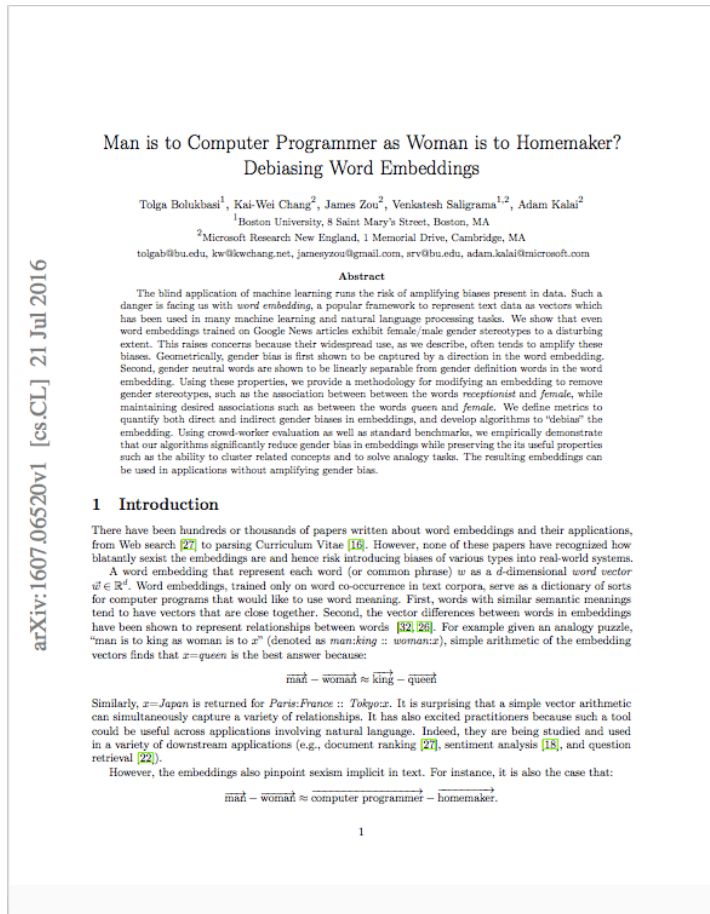
Most experts and commentators recommend that AI should always be applied transparently, and certainly without prejudice. Both the code of the algorithm and the process for applying it must be open to the public. Transparency should allow courts, companies, citizen watchdogs, and others to understand, monitor, and suggest improvements to algorithms (Oswald and Graos, 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).

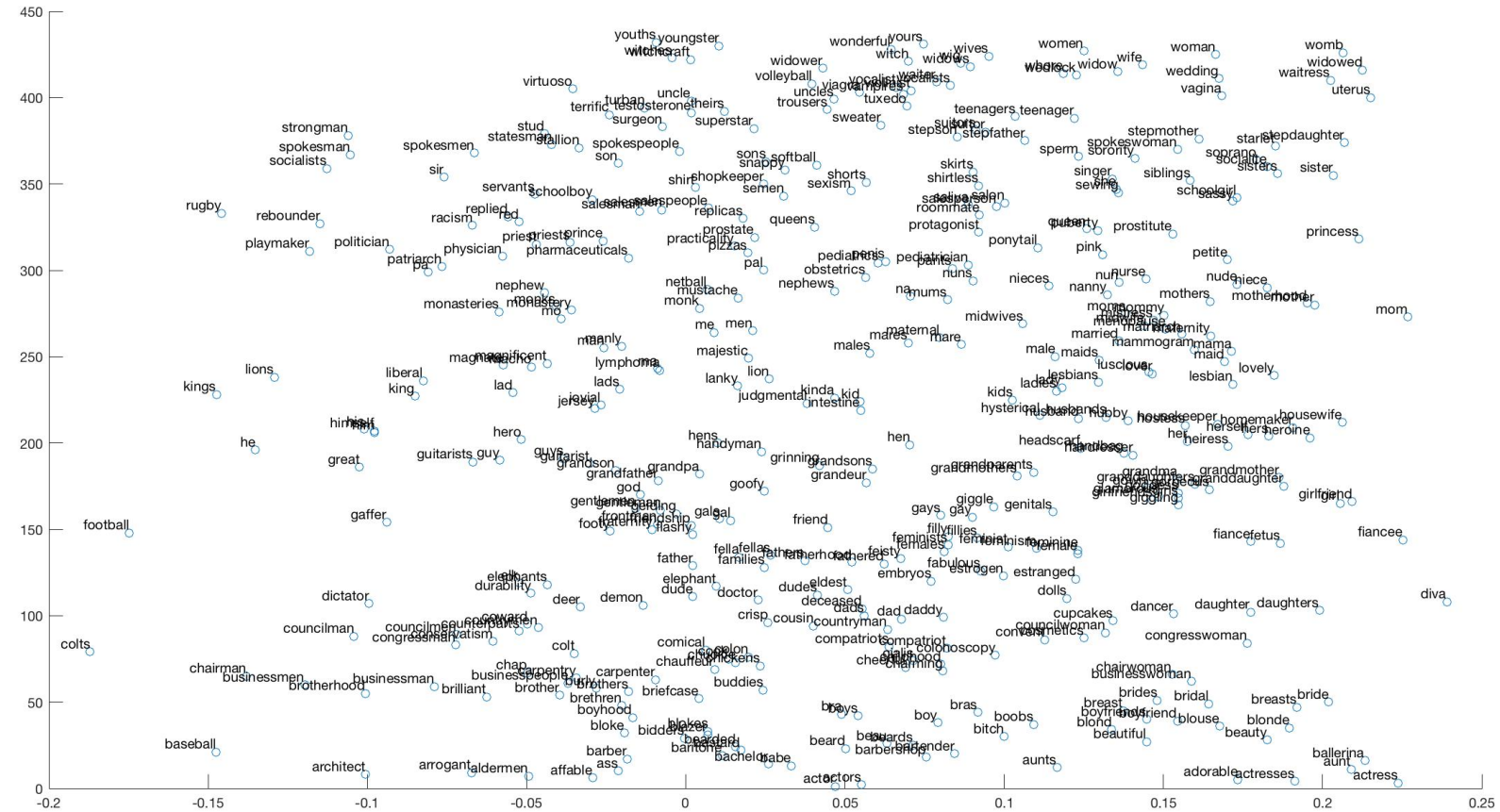
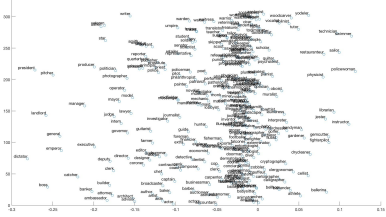
Here we show that while all of these strategies might be helpful and even necessary, they could not be sufficient. We document machine prejudice that derives so fundamentally from human culture that it is not possible to eliminate it through strategies such as the above. We demonstrate here for the first 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 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

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings Bolukbasi et al. 2016

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





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

Debiased Embeddings

Word2Vec embedding trained on
Google News

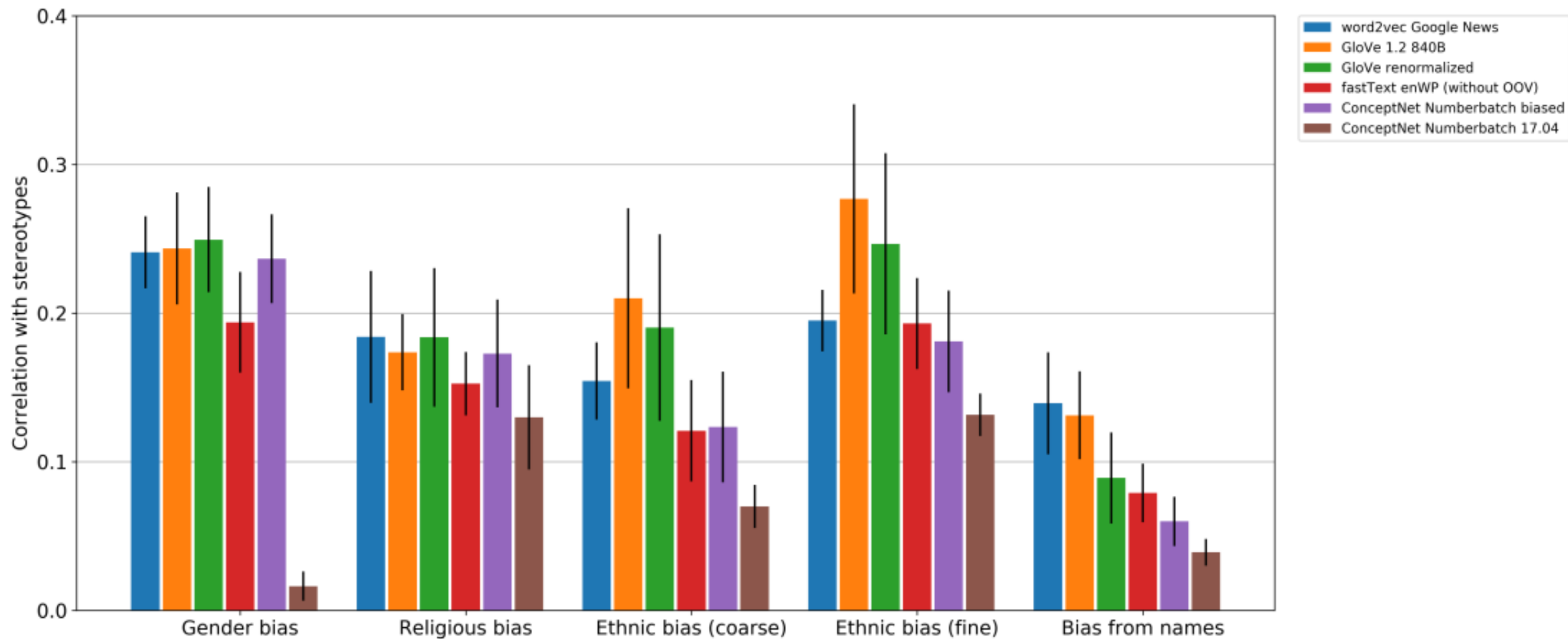
<https://github.com/tolga-b/debiaswe>

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 ...

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

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

programmer, engineer,
scientist ...

nurse, teacher,
librarian ...

man, male

woman, female

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

The diagram illustrates the components of the test statistic formula. Four blue arrows point from descriptive text to the variables in the formula: 'programmer, engineer, scientist ...' points to X, 'nurse, teacher, librarian ...' points to A, 'man, male' points to Y, and 'woman, female' points to B. The second summation term, $\sum_{y \in Y} s(y, A, B)$, is enclosed in an orange rectangular box.

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!

1

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

Examine Corpora

1

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

Test Embeddings

1

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

1

Data

2

Model

3

Deploy

- 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/

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

Primary Literature

Semantics derived automatically from language corpora necessarily contain human biases

Aylin Caliskan-Islam¹, Joanna J. Bryson^{1,2}, and Arvind Narayanan¹

¹Princeton University

²University of Bath

*Address correspondence to ayilinc@princeton.edu, bryson@conjugateprior.org, arvindh@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 to 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 Graos, 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

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

TextExplainer

2

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.

1

Data

2

Model

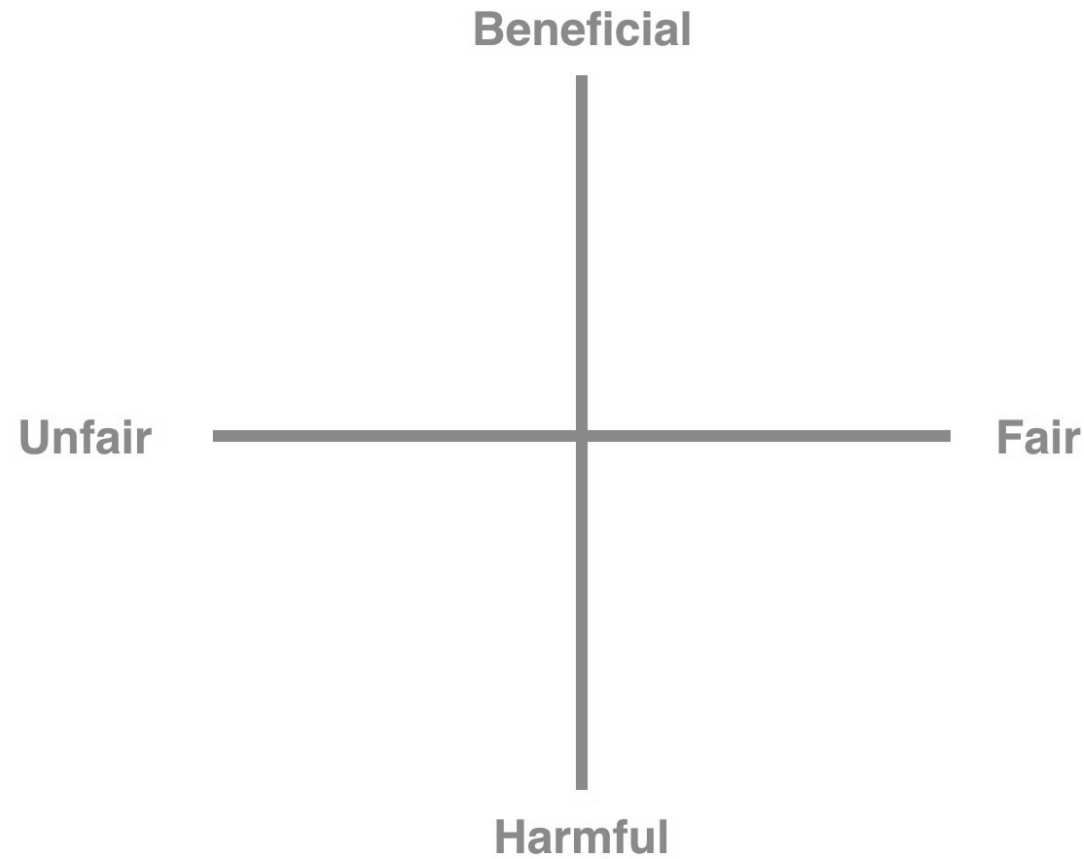
3

Deploy

- “Ethics for Powerful Algorithms”-
Abe Gong
- AI Now Institute 2017 Report

Ethics Review

3



Ethics Review

3

True positive	False positive, Type I error
False negative, Type II error	True negative



“Toxic” vs “Needs Review”



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

Last Mile Filters

3

```
pip install wordfilter profanity
```

```
pip install profanityfilter
```

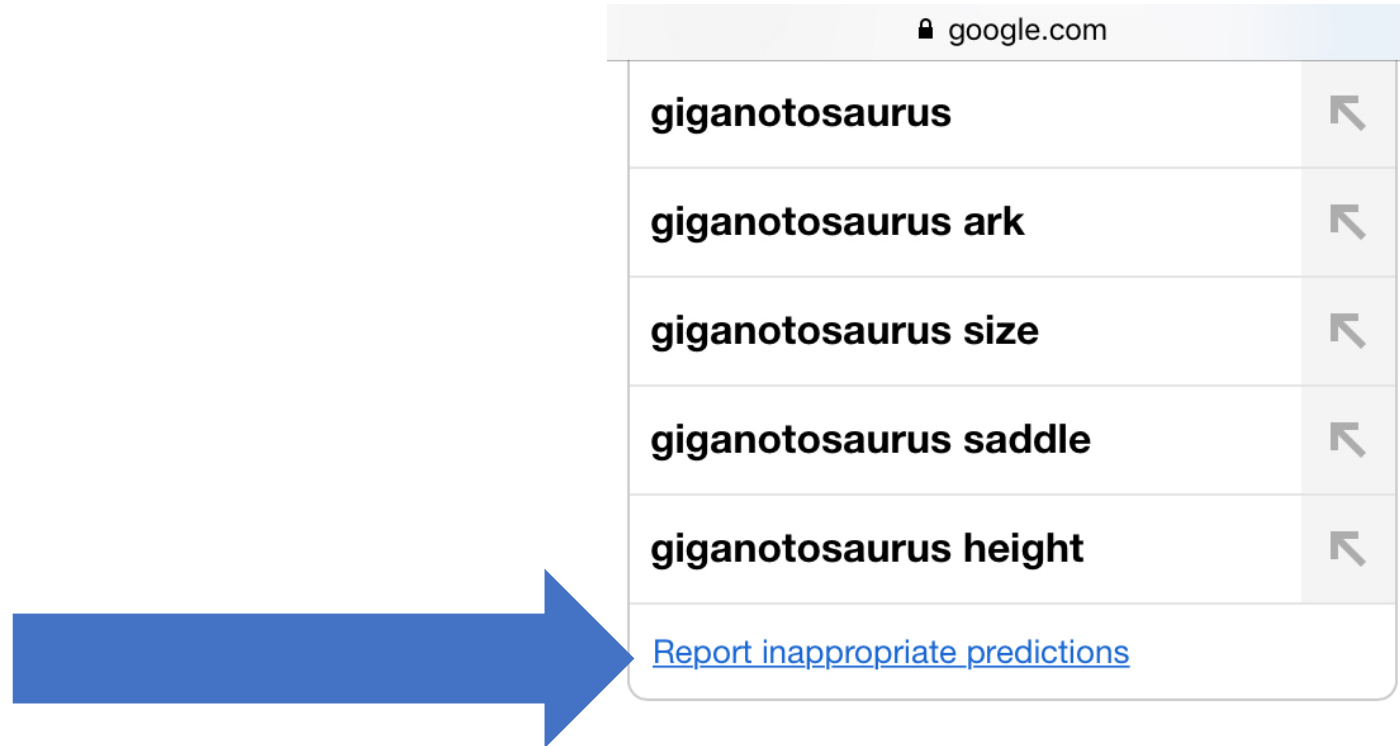
Last Mile Filters

3

- 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

3



Accountability

3

GDPR May 25, 2018

Disparate Impact

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 scholarly pubs and principles & 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>

Questions

github.com/mikecunha/pydata2017

@almostmike

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.

”

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”

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.

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.

”

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.

”

AI Now 2017 Report Reco's

“

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

”

AI Now 2017 Report Reco's

“

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

”

AI Now 2017 Report Reco's

“

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

”

Checklist

- Examine source data
 - Explore corpora with flashtext
 - Gender Pronoun Gap
- Assess embeddings (WEAT, WEFAT)
- Have you thought about “Do No Harm”?
 - What’s the worst that could happen? Can you trade off accuracy to prevent it?
- Test fairness of model output
 - Fairtest
 - Parity?
- Transparency: Can model output for a given input be explained
 - LIME
- Marketed appropriately (“toxic” vs “likely to need review”)
 - Avoid the “7 deadly sins of AI predictions”

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>

Quotes

- "If you do not intentionally include, you unintentionally exclude" -
[@carmatocity](#)

Quotes

- "Machine Learning is effectively, wizard magic from space. The public does not have a good mental model for what it is" - [@mtrc](#)
- "80% of modern AI is marketing" - [@mtrc](#)

Quotes

- “The role of computers in society has changed. Algorithms started out as recipes for automating tasks like sorting numbers. Today algorithmic systems make decisions about people's lives — screening job applicants, personalized health, scoring loan applications, criminal risk prediction, and many others — using deeply personal data.

So it's not enough to ask if code executes correctly. We also need to ask if it makes society better or worse. This of course involves judgment and isn't a pure computer science problem.”

- Arvind Narayanan