How Computer Algorithms Expose Our Hidden Biases

Revisited

Victor Zimmermann

KÖNIGSWEG

PyData SW

White man explains racism.

Personal primer.

I am not the most qualified person to talk about this subject.

- I am a Caucasian man from a somewhat privileged background.
- I have not experienced discrimination first-hand.
- My involvement with the topics discussed here have so far been limited to academic discourse.

As a result, I can not accurately speak on the true **impact** of bias in machine learning.
This omission should not be taken as a lack of importance.
Every point made here should be considered within the context of the millions of people affected by algorithmic decision making every day.

Why bias?

Then why talk about bias in machine learning at all?

- Biased algorithms are an **intrinsic** machine learning problem.
- There is a major **awareness** gap between researchers, developers and the general public.
- Imbalance of research on debiasing versus bias agnostic systems.

Terminology: Bias

Bias (Hardt, et al. 2016)
Inconsistent behaviour of a system towards input from different demographic groups.

Disparate Treatment
Different treatment because of some **protected attribute**, i.e. driven by discriminatory intent.

Disparate Impact
"Practices that are fair in form, but discriminatory in operation."
Griggs v. Duke Power Co., 401 U.S. 424 (1971)

Bias encoded.

Black TV viewers accuse 'creepy and racist' Netflix of targeting its adverts of films and shows to them by ethnicity

- Streaming giant accused of false advertising its content to entice black people
- An example is Love Actually, starring Hugh Grant and Emma Thompson as leads
- But Netflix used image of Chiwetel Ejiofor to make it look like it's primarily a love story about the black actor

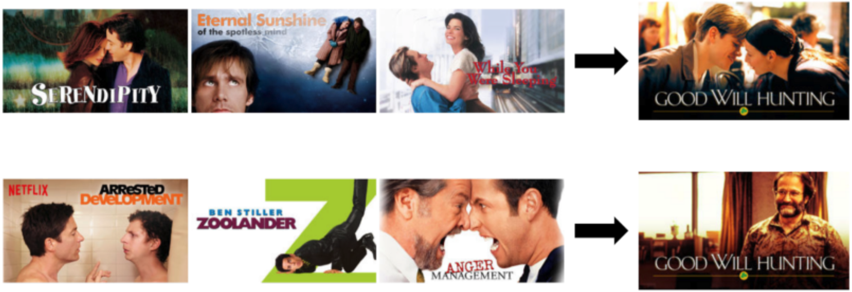
By MICHAEL POWELL FOR THE MAIL ON SUNDAY

PUBLISHED: 01:19 GMT, 21 October 2018 | UPDATED: 02:56 GMT, 21 October 2018

Film and television giant **Netflix** was embroiled in a racism row last night following

Netflix' artwork selection algorithm

"If the artwork representing a title captures something compelling to you, then it acts as a gateway into that title and gives you some visual "evidence" for why the title might be good for you." [Cha+17]



Netflix’ response

“We don’t ask members for their race, gender or ethnicity so we cannot use this information to personalise their individual Netflix experience. The only information we use is a member’s viewing history.” [Iqb18]

Terminology: Proxies and masking

Proxy

A variable that is **highly dependent** on the protected attribute.

Masking

Intentionally using proxies to mask discriminatory intent.

In practice, there should not be made a distinction between disparate treatment and impact, as discriminatory proxies can be selected with malicious intent.

Spoiler: All human language is biased.

Bias in not necessarily **performance based**. [Tan90][GMS98]
Instead it can also be encoded in **orthography**, **lexicography** or **grammar** of a language.

- Asymmetrically marked gender (generic masculine, e.g. actor vs actress)
- Quantity of gendered insults [1] [Sta77]
- Naming conventions (e.g. Chastity vs. Bob) [Swe13]

[1]Wikipedia lists 22 misogynistic and 5 misandric slurs.

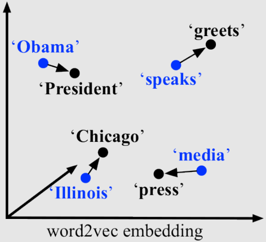
Debiasing word embeddings

What are word embeddings?

Condensed mathematical representations of collocations. [Mik+13]

CHICAGO – Former President Barack Obama campaigned in Chicago and northwest Indiana on Sunday, just days ahead of Tuesday's midterm elections.
Obama spoke Sunday afternoon at a get-out-the-vote rally in Gary, Indiana, supporting Democrat U.S. Sen. Joe Donnelly. The rally ended at about 3 p.m. and then spoke a rally at ... $\Rightarrow$

$$\overrightarrow{Obama}(0.2, 0.6, \ldots)$$
$$\overrightarrow{speaks}(0.1, 0.8, \ldots)$$
$$\overrightarrow{Chicago}(0.3, 0.2, \ldots) \Rightarrow$$
$$\overrightarrow{press}(0.0, 0.5, \ldots)$$
$$\vdots$$



$$\overrightarrow{Berlin} - \overrightarrow{Germany} + \overrightarrow{France} = \overrightarrow{Paris}$$
$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{queen}$$
$$\overrightarrow{programmer} - \overrightarrow{man} + \overrightarrow{woman} = \overrightarrow{homemaker}$$

What are word embeddings used for?

- Similarity measures [Kus+15]
- Machine translation [Zou+13]
- Sentence classification [Kim14]
- Part-of-speech-tagging [SZ14][RRZ18]
- Dependency parsing [CM14]
- Semantic modelling [Fu+14]
- Coreference resolution [Lee+17]

Basically the entire field of Computational Linguistics.

What if we just remove gender?

- Take "good" analogies,
  e.g. man-woman, he-she, king-queen, etc.
- Extract some average "gender vector" from their embeddings.
- Substract this new vector from all **other** relations.

Debiasing word embeddings

Word sets W, defining subsets $D_1, D_2, ..., D_n \subset W$, embedding $\{w \in \mathbb{R}^d\}_{w \in W}$, integer parameter $k \geq 1$, with

$$\mu_i := \sum_{w \in D_i} w / |D_i|$$

being the means of the defining subsets.

Bias subspace B consists of the first k rows of SVD(C), where

$$C := \sum_{i=1}^{n} \sum_{w \in D_i} (w - \mu_i)^T (w - \mu_i) / |D_i|.$$

Words to neutralise $N \in W$, family of equality sets $\mathcal{E} := \{E_1, E_2, ..., E_m\}$, $E_i \subseteq W$, with reembedded words $w \in N$ defined as

$$w := (w - w_B) / |w - w_B|$$

. For each set $E \in \mathcal{E}$, let

$$\mu := \sum_{w \in E} w / |E|$$

$$v := \mu - \mu_B$$

For each $w \in E, w := v + \sqrt{1 - |v|^2} \dfrac{w_B - \mu_B}{|w_B - \mu_B|}$

- This method leads to very little performance loss.
- It has been shown to work for different kinds of biases.
- It does nothing for downstream tasks.

"However, we argue that this removal is superficial. [...] The actual effect is mostly hiding the bias, not removing it. [...] Existing bias removal techniques are insufficient, and should not be trusted for providing gender-neutral modeling." [GG19]

Bias enhanced.

Why are machine learning techniques so vulnerable to exhibit biases?

"Traditional forms of data analysis [...] simply return records or **summary statistics** in response to a specific query [...]. [Machine learning] automates the process of discovering useful **patterns**, revealing regularities upon which subsequent decision making can rely." [BS16]

In contrast to trained statisticians, machine learning systems have no concept of **causality**.

Terminology: Bias in machine learning

Target variable

The attribute our system wants to measure, i.e. credit-worthiness, probability of recidivism, being a good student.

Class labels

Mutually exclusive, classifiable categories, i.e. default rates, arrests within two years, test scores.

Protected attribute

E.g. race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status.

Terminology: Equalised odds and equal opportunity

Equal odds

A system is said to satisfy equal odds if the predictor and the protected attribute are independent conditional on the target variable.

Equal opportunity

A system is said to satisfy equal opportunity if the predictor and the protected attribute are independent conditional on the target variable **for beneficial class labels**.

Bias in health risk assessment [Obe+19]

- A widely used algorithm to determine health needs of patients exhibits racial bias.
- At a given risk score, Black patients are considerably sicker than White patients.
- Health as target variable not directly accessible.
- Instead, the system minimises the projected health cost of a patient.
- As a result of historical, cultural and institutional racism, Black patients receive treatment later than White patients and are more often misdiagnosed.
- The target variable is (largely) independent on race, the class label (risk score) is dependent on race through the proxy of health costs.

Common language identification systems use extensive news corpora for training.

+ Big corpora in most languages.
+ Mostly “unbiased” texts.
- Written in main dialect.
- Privileged writing staff.

Problem: African American English is 20% less likely to be classified as English than Standard English. [BO17]

Solution by Blodgett, Green, and O'Connor (2016):

1. Use **US Census** data und **geolocated tweets** to estimate race of user,
2. Train **classifier** to identify "race" of a given tweet, based on **high AA tweets** from first set.

Result:

- Build **new corpus** from high AA tweets.
- (Find out that "Asian" captures all foreign languages and use that fact for classification.)

Word embeddings, language classification and tons of other tasks not mentioned here (e.g. coreference resolution) are fundamental NLP tasks, often performed in preprocessing.

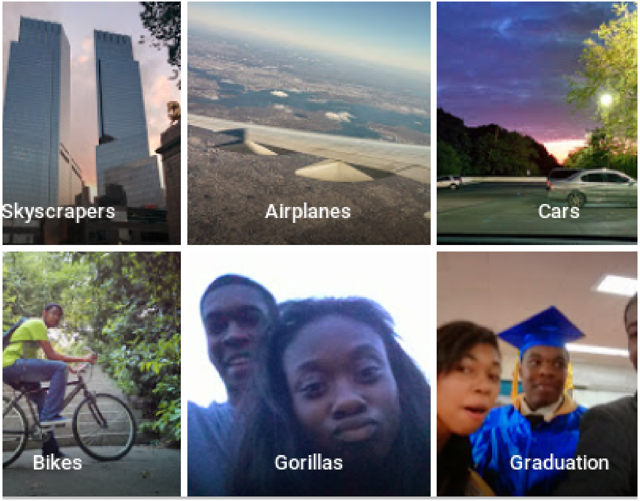
As a result, downstream tasks often not only reproduce bias, they amplify it. [Zha+17]

Using a kind of affirmative action, bias can be reduced after the fact [HPS16], but only with substantial performance loss and access to the protected attribute.

Bias enabled.

Google's image recognition controversy

Google automatically labels pictures according to their content.



Source: @jackyalcine on Twitter

Their solution:

GOOGLE \ TECH \ ARTIFICIAL INTELLIGENCE

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

51

Nearly three years after the company was called out, it hasn't gone beyond a quick workaround

By James Vincent | @jjvincent | Jan 12, 2018, 10:35am EST

SHARE

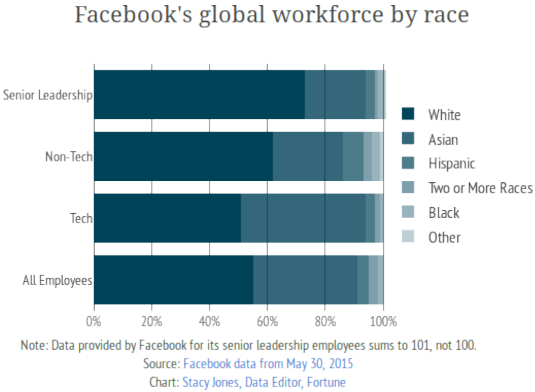
Source: `www.theverge.com` (visited on 2018-11-06)

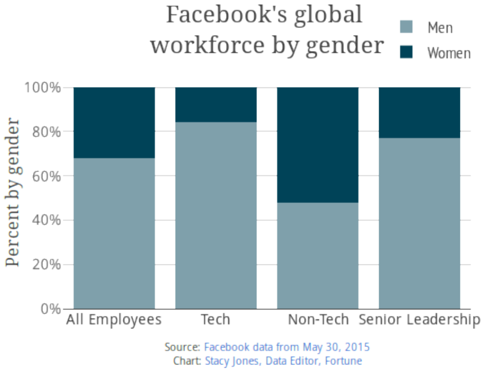
Actual quote from an actual Facebook employee

"We started out of a college dorm. I mean, c'mon, we're Facebook. We never wanted to deal with this shit." [Sha16]

Possible cause of this apathy:

(Don’t quote me on this.)





- Recent headlines show even internet giants are largely unaware of the issue.
- Even though the user base is diverse, the people writing the code are not.
- There are systemic issues that go beyond algorithmic bias, that lead to biased systems that should not have passed the testing stage.

Conclusions.

Main takeaways

- Bias is a defining problem of machine learning.
- Neglecting to address bias crosses into unethical behaviour as soon as peoples lives are affected.
- Try to think about bias every step of the way.
- Diverse staffing makes a difference.
- Training data makes a difference.
- Awareness makes a difference.

Thank you!

Slides, resources and contact info:

vz@koenigsweg.com

axtimhaus.eu

gitlab.com/axtimhaus

@dieaxtimhaus

linkedin.com/in/viczim/

Further reading

Getting Specific About Algorithmic Bias - Rachel Thomas at PyBay 2019

Big Data’s Disparate Impact - Solon Barocas, Andrew D. Selbst

Ethical Implications of Bias in Machine Learning - Adrienne Yapo, Joseph Weiss

AI Fairness 360 Open Source Toolkit - IBM Research Trusted AI

References

[BGO16] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. "Demographic dialectal variation in social media: A case study of African-American English". In: *arXiv preprint arXiv:1608.08868* (2016).

[BO17] Su Lin Blodgett and Brendan O'Connor. "Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English". In: *arXiv preprint arXiv:1707.00061* (2017).

[BS16] Solon Barocas and Andrew D Selbst. "Big data's disparate impact". In: *Calif. L. Rev.* 104 (2016), p. 671.

[Cha+17] Ashok Chandrashekar, Fernando Amat, Justin Basilico, and Tony Jebara. "Artwork Personalization at Netflix". In: *Netflix Techblog* (Dec. 7, 2017). URL: *https://medium.com/netflix-techblog/artwork-personalization-c589f074ad76* (visited on 11/05/2018).

[CM14] Danqi Chen and Christopher Manning. "A fast and accurate dependency parser using neural networks". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 740–750.

[Fu+14] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. ''Learning semantic hierarchies via word embeddings''. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014, pp. 1199–1209.

[GG19] Hila Gonen and Yoav Goldberg. ''Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them''. In: *Proceedings of NAACL-HLT*. 2019, pp. 609–614.

[GMS98] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. ''Measuring individual differences in implicit cognition: the implicit association test.''. In: *Journal of personality and social psychology* 74.6 (1998), p. 1464.

[HPS16] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of Opportunity in Supervised Learning". In: *arXiv preprint arXiv:1610.02413* (2016).

[Iqb18] Nosheen Iqbal. "Film fans see red over Netflix 'targeted' posters for black viewers". In: *The Guardian* (Oct. 20, 2018). URL: *https://www.theguardian.com/media/2018/oct/20/netflix-film-black-viewers-personalised-marketing-target* (visited on 11/05/2018).

[Kim14] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[Kus+15] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. "From word embeddings to document distances". In: *International Conference on Machine Learning*. 2015, pp. 957–966.

[Lee+17] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. "End-to-end neural coreference resolution". In: *arXiv preprint arXiv:1707.07045* (2017).

[Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[Obe+19] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[RRZ18] Ines Rehbein, Josef Ruppenhofer, and Victor Zimmermann. "A harmonised testsuite for POS tagging of German social media data". In: (2018).

[Sha16] Aarti Shahani. "From Hate Speech To Fake News: The Content Crisis Facing Mark Zuckerberg". In: *NPR* (Nov. 17, 2016). URL: *https://www.npr.org/sections/alltechconsidered/2016/11/17/495827410/from-hate-speech-to-fake-news-the-content-crisis-facing-mark-zuckerberg?t=1542640881872* (visited on 11/19/2018).

[Sta77] Julia Penelope Stanley. "Paradigmatic woman: The prostitute". In: *Papers in language variation* (1977), pp. 303–321.

[Swe13] Latanya Sweeney. "Discrimination in online ad delivery". In: *Queue* 11.3 (2013), p. 10.

[SZ14] Cicero D Santos and Bianca Zadrozny. "Learning character-level representations for part-of-speech tagging". In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014, pp. 1818–1826.

[Tan90] Dali Tan. "Sexism in the Chinese Language". In: *NWSA Journal* 2.4 (1990), pp. 635–639.

[Zha+17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". In: (2017). arXiv: *1707.09457*. URL: *http://arxiv.org/abs/1707.09457*.

[Zou+13] Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. "Bilingual word embeddings for phrase-based machine translation". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013, pp. 1393–1398.