

*Responsible AI: free from
(human) bias, unfairness and
discrimination*

Natalia Criado Pacheco



Gender Discrimination Persists

- Only a **quarter of board chairs, presidents and chief executives in the UK are women**
- **A male graduate could expect to earn 20% more** on average, than a female graduate. The gap was wider for non-degree holders at 23%
- There are 2.11 million men and 5.85 million women in part-time employment
- **Women are responsible for 74%** of the time spent on **childcare**
- Only 28 per cent of top business people are women
- Sexual discrimination continued to be the most frequent type of discrimination claim received by tribunals during

UK gender inequality as bad as 10 years ago, EU league table shows

Britain ranks alongside Slovakia and the Czech Republic in gender equality index as having made no progress in range of fields



<https://www.theguardian.com/inequality/2017/oct/11/uk-no-further-forward-on-tackling-gender-inequality-eu-league-table-shows>

Just a few examples...

Vague guidance won't stop women being forced to wear heels and makeup

Anna Macey

The government's bland advice on sexist dress codes misses an opportunity to take on this serious problem

Mon 21 May 2018 13:39 BST



102



<https://www.theguardian.com/commentisfree/2018/may/21/women-heels-makeup-guidance-sexist-dress-codes>

Just a few examples...

Revealed: MPs' gender pay gap



Male Conservative MPs earn an average of £7,789 more than female Tory MPs
ALAMY

Male MPs earn 10.4% more on average than their female counterparts despite identical basic salaries, exposing the way the gender pay gap in British society runs right to the heart of government.

<https://www.thetimes.co.uk/article/revealed-mps-gender-pay-gap-glzhtb0g7>

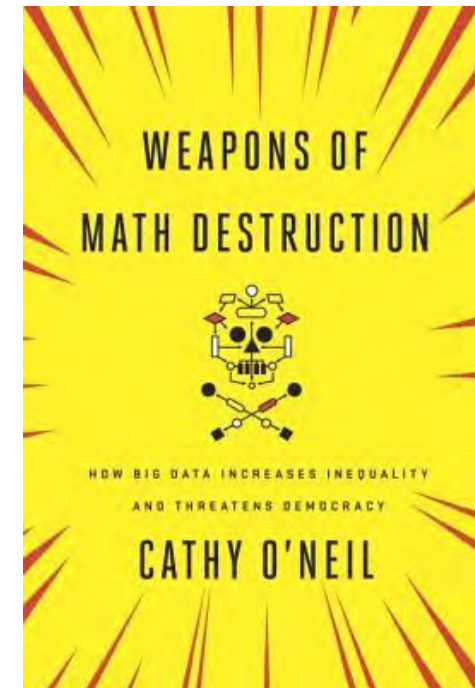
Algorithms and Decision Making

- **More and more decisions** are delegated to/informed by **algorithms**
 - From the jobs we apply for, the products we buy, to the news we read and the persons we date
- These algorithms improve automatically through experience:
 - Data mining find patterns in a given dataset
 - Identifying associations between candidates characteristics and hiring decisions
 - Supervised algorithms learn a general rule that maps inputs to outputs
 - Selecting good candidates for a job
- Algorithms are perceived as **faultless**, not having most of the shortcomings that we humans have
 - Tiredness, lack of stamina, lack personal prejudices
- Their decisions may be **less scrutinized** than the ones made by their human counterparts

Algorithms are not free from discrimination:

Digital Discrimination

- However, algorithms are likely to **inherit prejudices** of:
 - Programmers
 - Previous decisions
 - Users
 - Society
- They have the potential to
 - Discriminate more *consistently* and *systematically* and at a *larger scale*
 - *Perpetuate* discrimination



Underrepresentation in Search Results: CEO

2015



2018



Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations

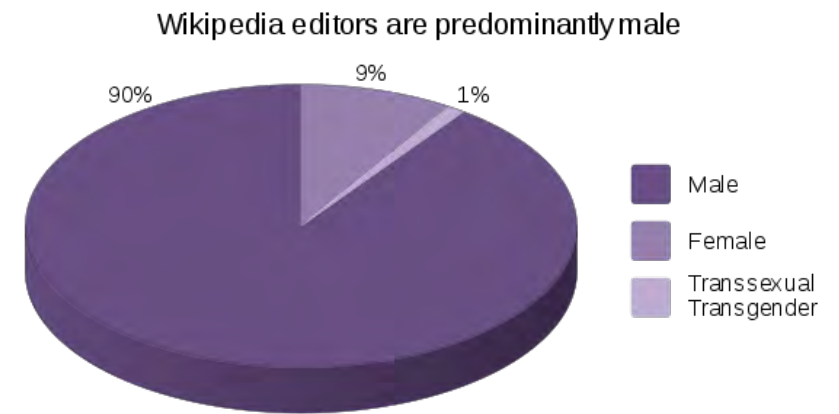
Stereotypes in Search Results: Doctor



Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations

Gender Discrimination in Wikipedia

- Family-, gender-, and relationship-related topics being more present in biographies about women
- Positive terms being more frequent in the biographies of men and negative terms more frequent in the biographies of women
- Structural differences in terms of meta-data and hyperlinks, which have consequences for searching activities



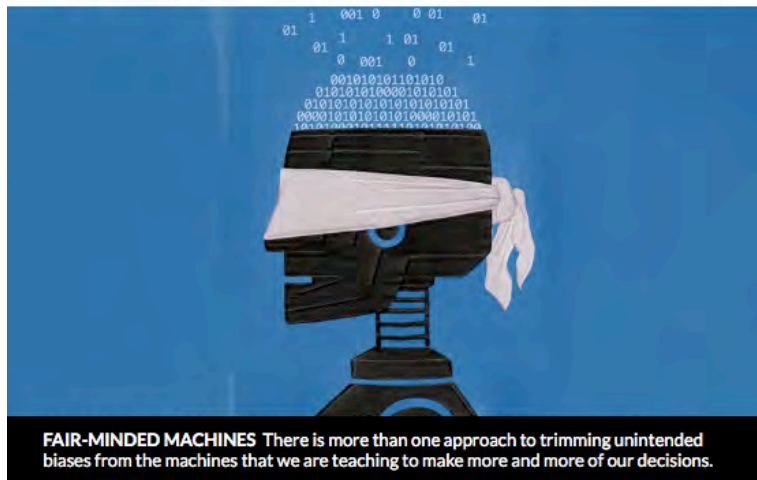
Wagner, C., Garcia, D., Jadidi, M., & Strohmaier, M. (2015). It's a man's Wikipedia? assessing gender inequality in an online encyclopedia.

Gender Discrimination in Ads Targeting

Machines are getting schooled on fairness

Machine-learning programs can introduce biases that may harm job seekers, loan applicants and more

BY MARIA TEMMING 1:00PM, SEPTEMBER 6, 2017

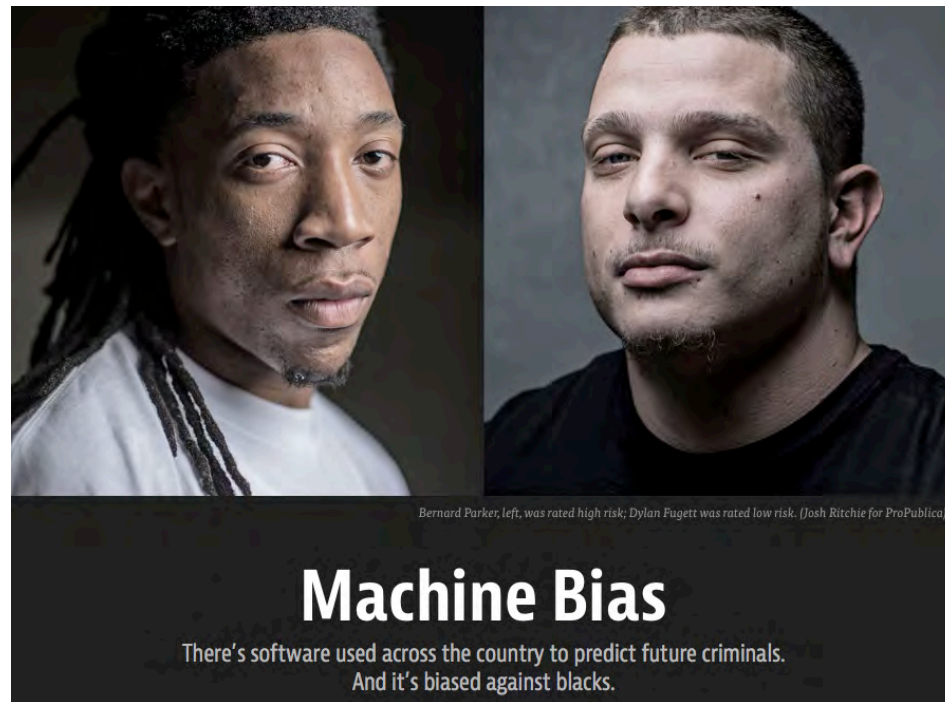


Google's ad service preferentially displayed postings related to high-paying jobs to men

<https://www.sciencenews.org/article/machines-are-getting-schooled-fairness>

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings.

Digital Discrimination in other Domains



<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Causes for Digital Discrimination

- Modelling:
 - Algorithms predict the value of some variable based on some data
 - Some problems there may be an objective and ambiguous definition of that variable
 - Predict age from images
 - For other problems the target variable and its categories are an artefact defined by the designer of the algorithm
 - Classification of potential employees into good/bad
- Training Data
 - Algorithms build a prediction model on the information on the training dataset which can:
 - Reflect existing prejudices of decision makers
 - Underrepresent a particular social group
 - Overrepresents a particular social group
 - Reflects social biases
- Usage
 - A non-discriminatory algorithm can also lead to discrimination when it is used in an situation for which it was not intended

Technological Solutions for Digital Discrimination

- Detection
- Prevention

Detection Metrics

- Direct Discrimination
 - Influence protected attributes and decision
 - Gender > Hiring Decisions
 - Influence between non-protected attributes and protected attributes (Proxies)
 - Career breaks -> Gender
- Disparate Impact
 - 80% rule
 - The selection rate of a protected group should be at least 80% the selection rate of the non-protected group.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems.

Prevention

- Modifying the problem modeling
 - Mask protected attributes or learn non-discriminative representations
- Data Preprocessing
 - Unbalanced datasets
 - Over-sampling, under-sampling, re-sampling
 - Textual data
 - Modify vector representations to remove discriminatory association
- Algorithm Modification
 - Include non-discrimination as a criterion to maximize together with prediction accuracy

Calders, T., Karim, A., Kamiran, F., Ali, W., & Zhang, X. (2013). Controlling attribute effect in linear regression.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness.

Zhang, H., Berg, A. C., Maire, M., & Malik, J. (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition.

Open Issues

- Previous research can quantify and/or limit the extent of bias of an algorithm/dataset
 - Assume that the protected ground are externally given
 - Assume that there is a quantifiable definition of discrimination

What bias count as discrimination?

How much bias is too much?

Discovering and Attesting Digital Discrimination

- A **mutli-disciplinary** team formed by experts in *computer science, social sciences, ethics* and *law*
 - Mainly contribution in ICT
 - Survey Legal-Ethical Fields
 - Investigate Methodologies to understand user expectations
- Create the new transdisciplinary field of **digital discrimination certification**

