# Data Science as Decision Tool: Algorithmic Fairness

## Methods for Detecting Automated Discrimination

Caitlin Kuhlman PhD Candidate
Computer Science Department, Worcester Polytechnic Institute

WPI

WOMEN IN DATA SCIENCE
CENTRAL MASSACHUSETTS
@ WPI

https://www.widscentralmass.org/

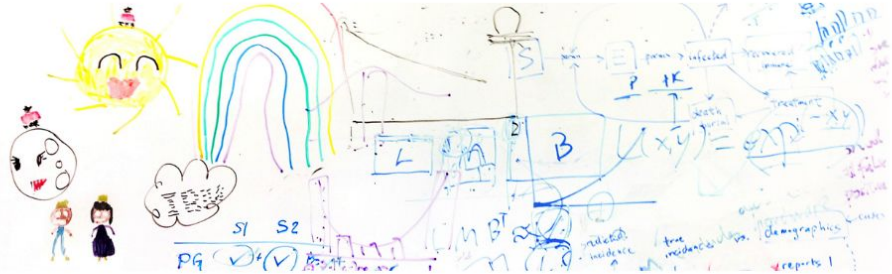
https://www.widsconference.org/

IBM

IBM Research  Research areas  Work with us  About us  Blog

Science for Social Good  feedback

Overview  Apply  Our Team  2017 Fellows  2016 Fellows & Mentors  2016 Projects



Mentoring, giving back, making a difference -- with science.

Promoting scientific advances by addressing untapped social and humanitarian problems.

https://www.research.ibm.com/science-for-social-good/

https://dssg.uchicago.edu/

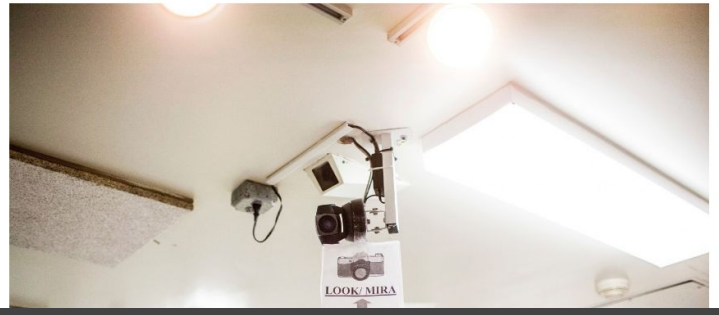
https://escience.washington.edu/dssg/

https://ptc.gatech.edu/dssg

The Washington Post
Democracy Dies in Darkness

Business

# Racial profiling, by a computer? Police facial-ID tech raises civil rights concerns.




The New York Times

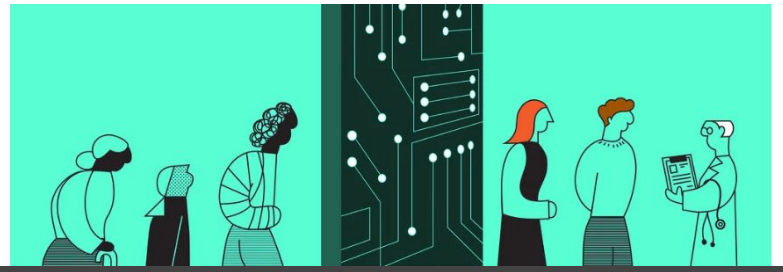
Opinion

# A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

By Dhruv Khullar

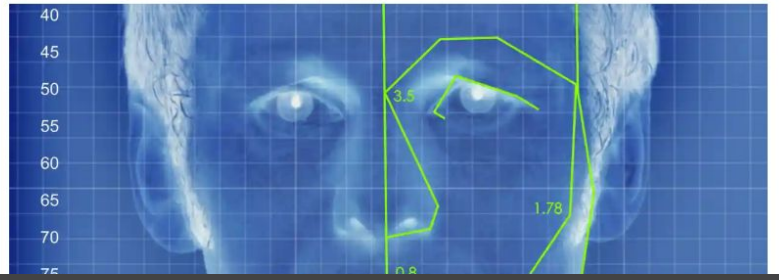
Dr. Khullar is an assistant professor of health care policy and research.



Search jobs | Sign in | Search | The Guardian | US edition

# New AI can guess whether you're gay or straight from a photograph

**An algorithm deduced the sexuality of people on a dating site with up to 91% accuracy, raising tricky ethical questions**



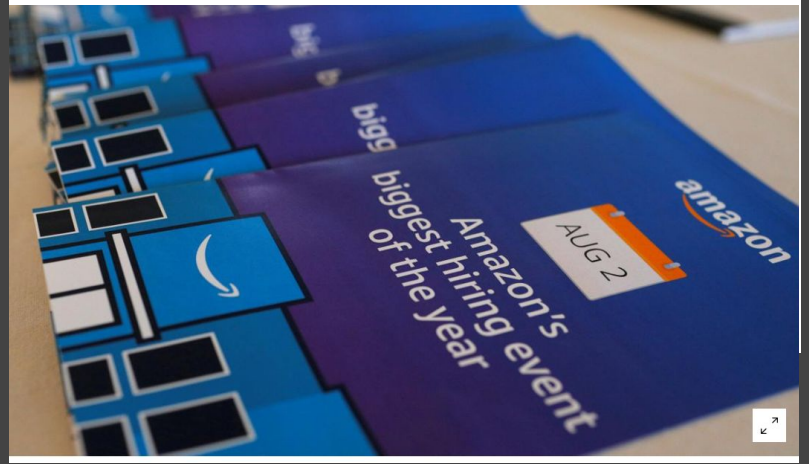

SundayReview | OPINION

# Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

REUTERS World Business Markets Politics TV

# Amazon scraps secret AI recruiting tool that showed bias against women

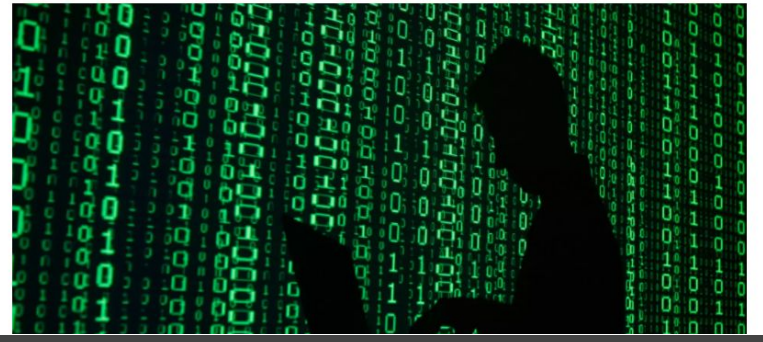



The Telegraph HOME | NEWS | SPO

UK | World | Politics | Science | Education | Health | Brexit | Royals | Investigations

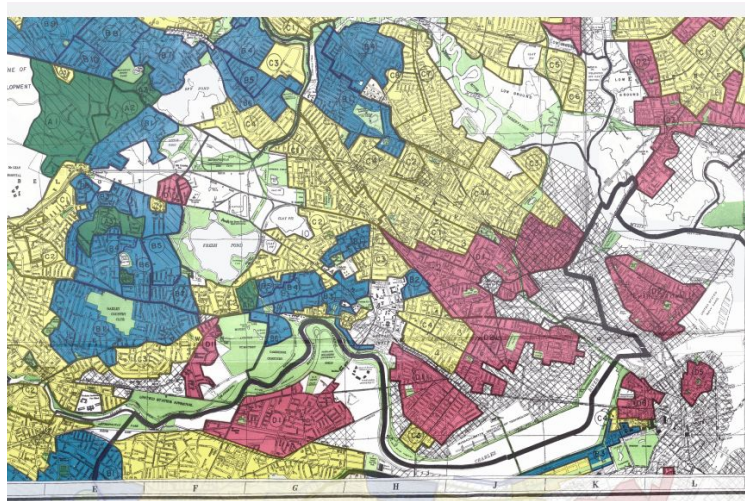
News

# AI robots are sexist and racist, experts warn



# Who Cares?

# Who Cares?



## Regulated Domains:

- Credit (Equal Opportunity Act)
- Housing (Fair Housing Act)
- Education, Employment, Public Accommodation (Civil Rights Act)

## Protected Attributes:

- Race, Religion, National Origin, Familial Status. . . (Civil Rights Act)
- Sex (Equal Pay Act, Civil Rights Act)
- Age (Age Discrimination in Employment Act)
- Disability (Americans with Disabilities Act)
- Genetic Information (Genetic Information Nondiscrimination Act)
- . . .

*Extends to marketing and advertising

NIPS Tutorial, Hardt, Moritz 2017 http://fairml.how/tutorial
https://dsl.richmond.edu/panorama/redlining

# Who Cares?

**Standards:**

- Proposed IEEE standard 7003 “Algorithmic Bias Considerations”

**New Regulations:**

- General Data Protection Regulation (GDPR)
  - Right to Explanation
  - Washington Privacy act
  - California Consumer Privacy Act

https://standards.ieee.org/develop/project/7003.html
Goodman, Bryce, and Seth Flaxman. European Union regulations on algorithmic decision-making and a" right to explanation". *arXiv:1606.08813* (2016).
https://fatconference.org/

# Who Cares?

AAAI / ACM conference on
ARTIFICIAL INTELLIGENCE, ETHICS, AND SOCIETY

## Conferences and Workshops

FAT* Conferences

AI Society and Ethics

FATES@WebConference - Feb 10th!!

FATRec @RecSys

AI Ethics Workshop @NeurIPS

. . .


https://fatconference.org/index.html
http://www.aies-conference.com/

# Overview

Algorithmic Fairness studies ways in which automated systems may perpetuate or exacerbate unfair bias and inequity.

We will investigate 3 examples of ways unfair bias can enter a data mining pipeline:

1. Historical Bias
2. Data Collection
3. Validation Techniques

# Learning Objectives

1. Recognize ways in which unfair bias might be introduced into a data mining pipeline.
2. Perform analysis to verify whether a predictive model is fair.
3. Discuss what data mining outcomes we should consider beyond accuracy.
4. Get hands-on experience with open tools for conducting transparent, reproducible research.

# Reproducible Research

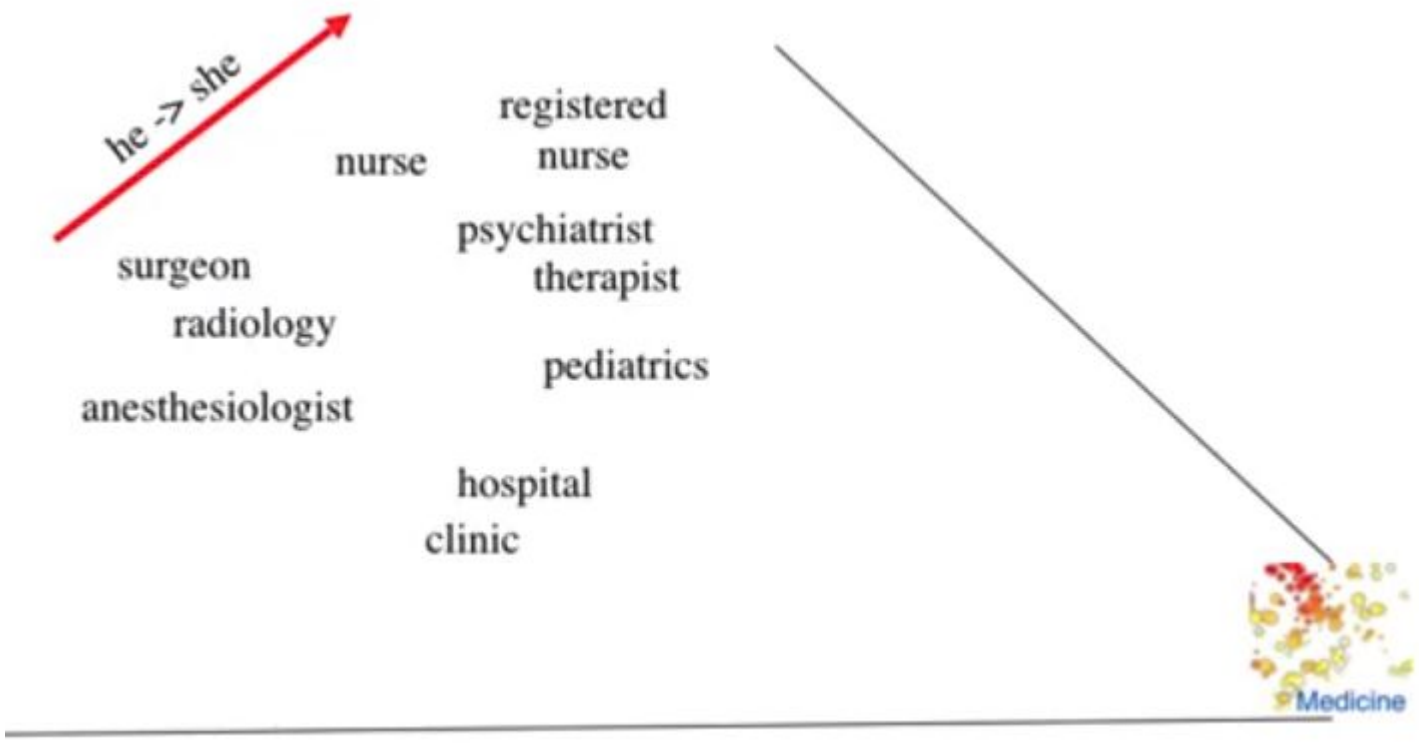
- Github
  - https://github.com/caitlinkuhlman/bpdmtutorial
- Jupyter notebooks
  - https://www.anaconda.com
- Python libraries
  - https://scikit-learn.org
  - https://pandas.pydata.org

# Example 1: Historical Bias

# Historical Bias: Gender Steotypes in NLP

| he: ___ | she: __ |
|---|---|
| uncle | aunt |
| lion | lioness |
| surgeon | nurse |
| architect | interior designer |
| beer | cocktail |
| professor | associate professor |

… many more




Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in Neural Information Processing Systems*. 2016.
https://www.youtube.com/watch?v=aZariglogXc

# Example 2: Data Collection

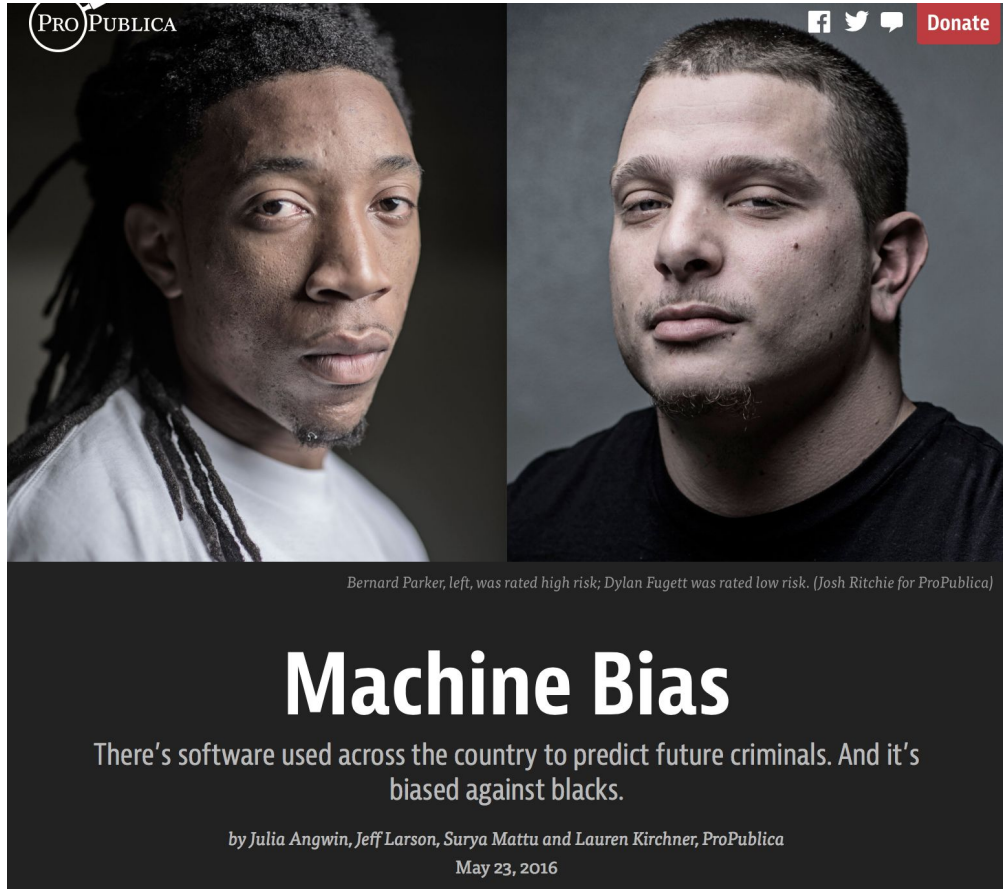
# Data Collection: Bias in Computer Vision

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on Fairness, Accountability and Transparency*. 2018.
http://gendershades.org/overview.html

# Example 3: Validation Techniques

# Validation Techniques: Risk Scoring




| Fairness Criteria |
| --- |
| Calibration |
| Balance for Negative Class |
| Balance for Positive Class |

NORTHPOINTE

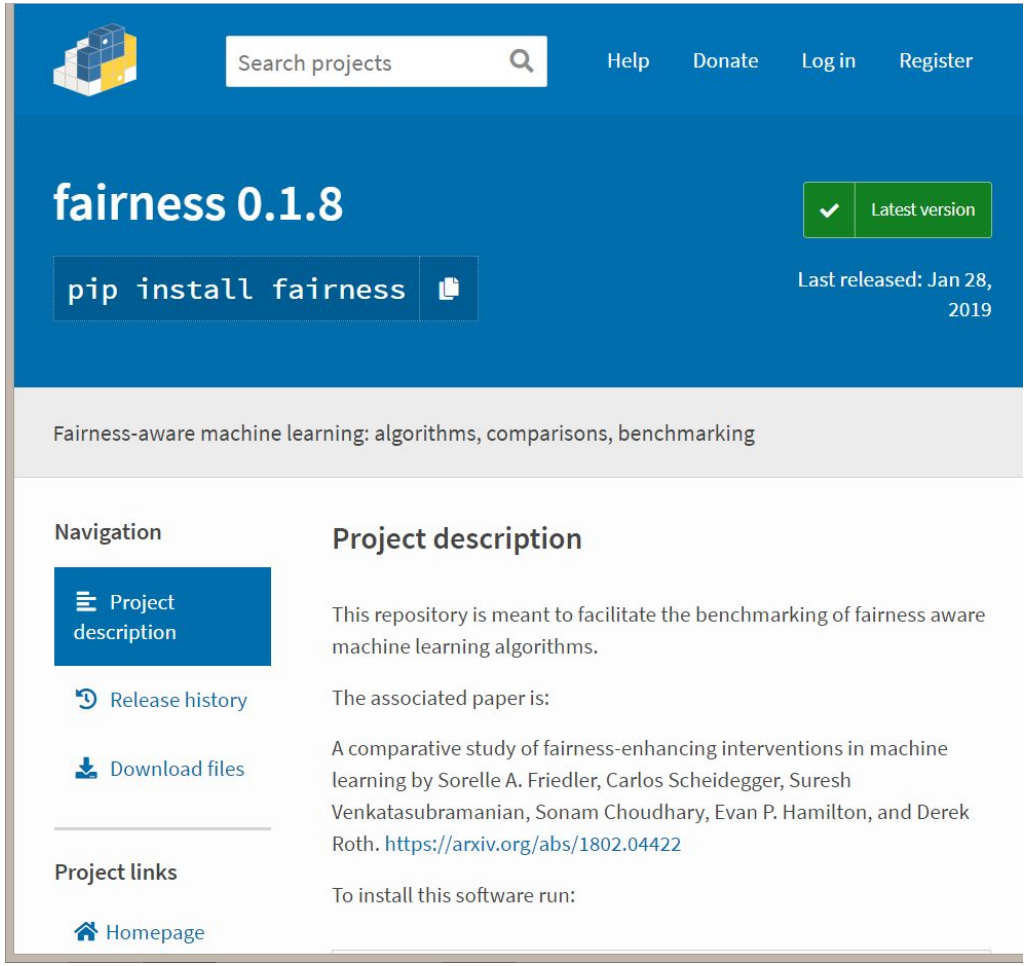
PRO PUBLICA

Mutually exclusive measurements of fairness

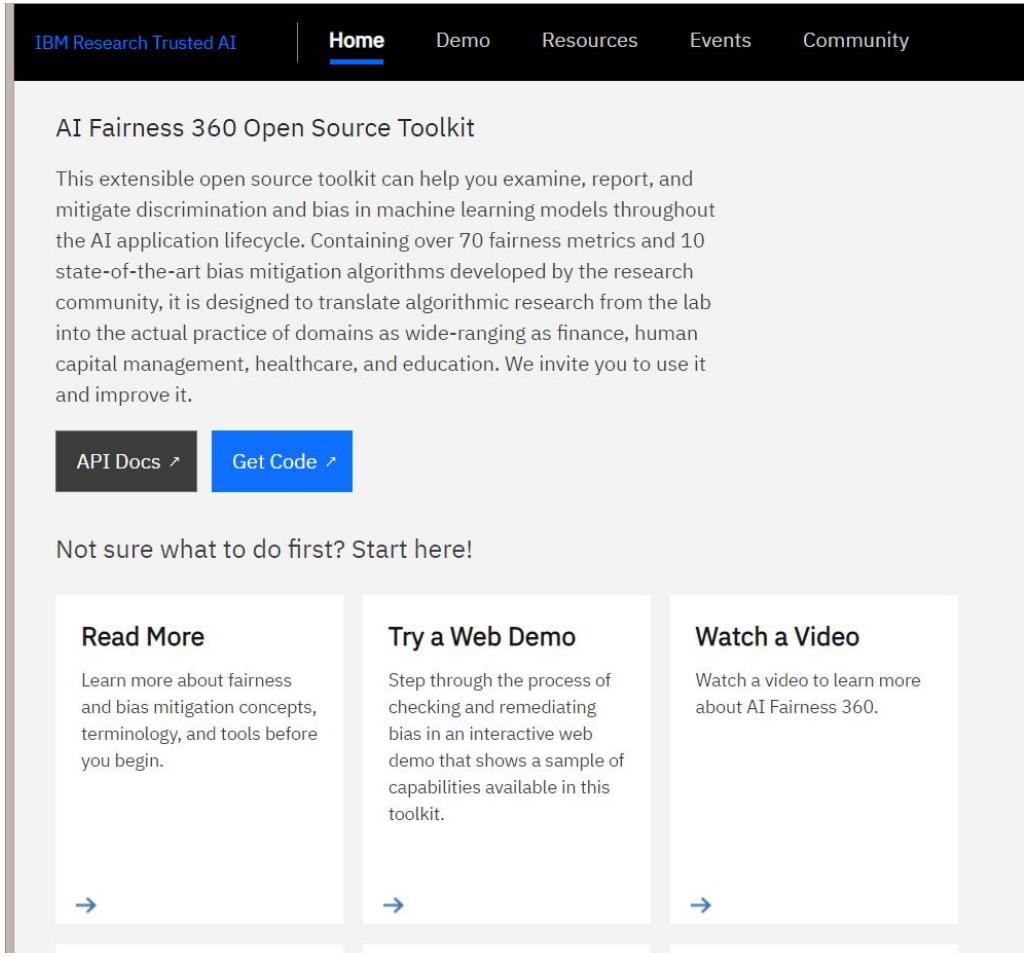

Angwin, Julia, et al. "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *ProPublica, May* 23 (2016).
Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." arXiv preprint arXiv:1703.00056 (2017).
Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807(2016).

# More Resources

# Fairness Libraries

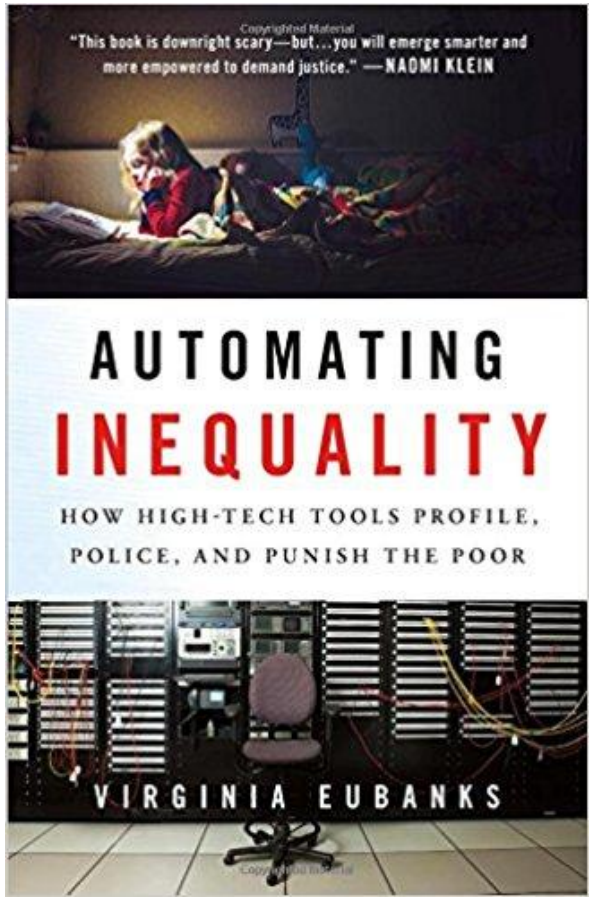


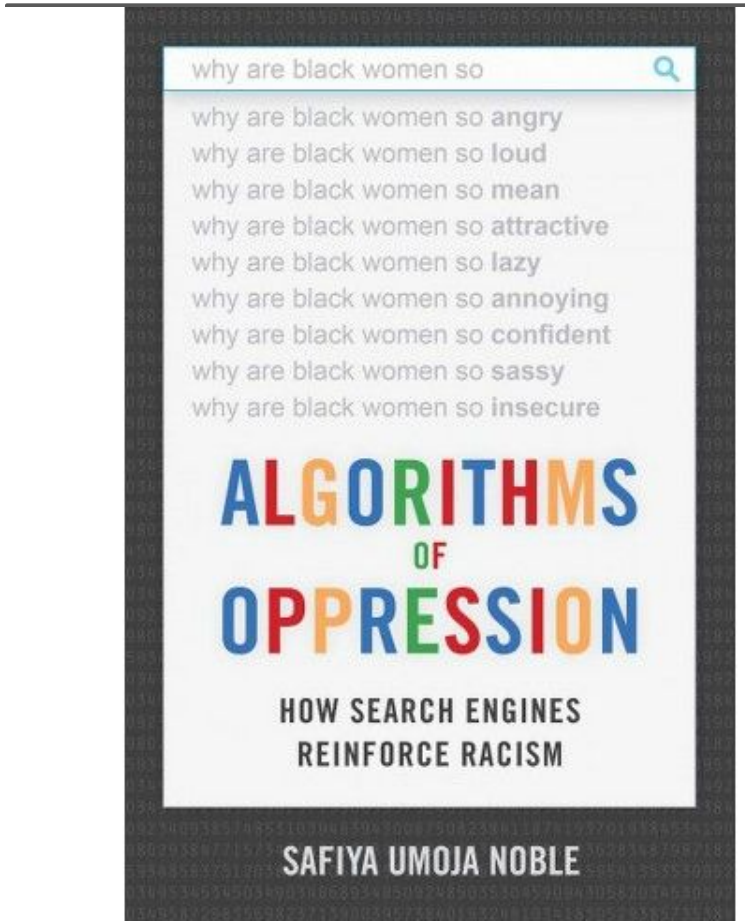

https://pypi.org/project/fairness

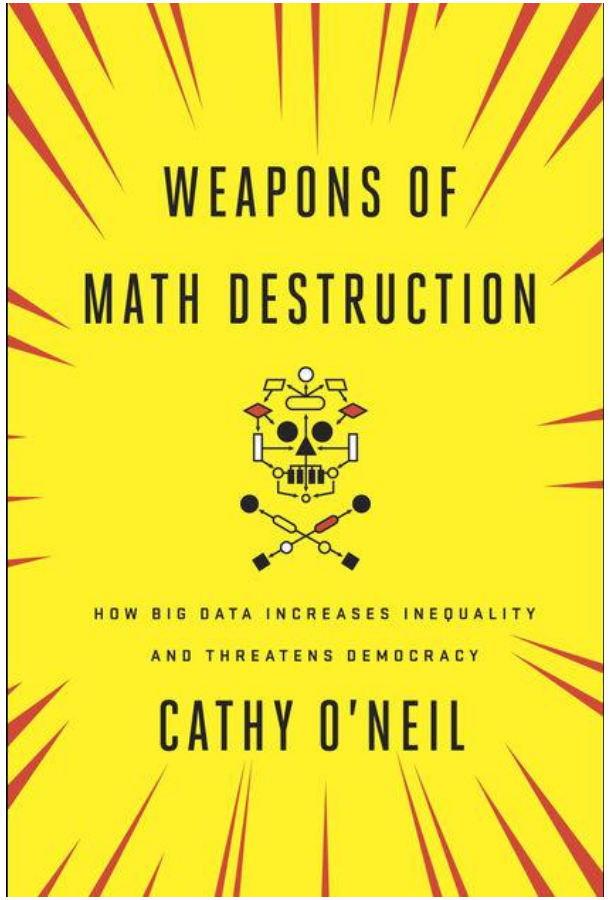



http://aif360.mybluemix.net/

# Books




Eubanks, Virginia. "Automating inequality: how high-tech tools profile, police, and punish the poor." (2018).
O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.
Noble, Safiya Umoja. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

**News Articles**

https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html
http://www.telegraph.co.uk/news/2017/08/24/ai-robots-sexist-racist-experts-warn/
https://www.washingtonpost.com/business/economy/face-recognition-tech/2016/10/17/986929ea-41f0-44a2-b2b9-90b495230dce_story.html?utm_term=.491efe7bec56
https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de
https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses
https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html
https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

Crawford, Kate. "Artificial intelligence's white guy problem." The New York Times (2016).

Angwin, Julia, et al. "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." *ProPublica, May* 23 (2016).

**Resources**

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
Machine Bias Interactive Tutorial: https://github.com/caitlinkuhlman/bpdmtutorial
NIPS Fairness Tutorial: http://fairml.how/tutorial
O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2017.
**https://www.fatml.org/**
**https://fatconference.org/**
Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, et al., "Mapping Inequality," American Panorama, ed. Robert K. Nelson and Edward L. Ayers, accessed February 11, 2018,https://dsl.richmond.edu/panorama/redlining
https://en.wikipedia.org/wiki/Disparate_impact#The_80.25_rule
IEEE Proposed fairness standard https://standards.ieee.org/develop/project/7003.html

**References**


1. Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." *Cal. L. Rev.* 104 (2016): 671.

2.Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware data mining." In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 560-568. ACM, 2008.

3. Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and removing disparate impact." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 259-268. ACM, 2015.

4. Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. "Building classifiers with independency constraints." Data mining workshops, 2009. ICDMW'09. IEEE international conference on. IEEE, 2009.

5. Lum, Kristian, and James Johndrow. "A statistical framework for fair predictive algorithms." arXiv preprint arXiv:1610.08077 (2016).

6. Zemel, Rich, et al. "Learning fair representations." International Conference on Machine Learning. 2013.

7. Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." In Proceedings of the 26th International Conference on World Wide Web, pp. 1171-1180. International World Wide Web Conferences Steering Committee, 2017.

8. Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic decision making and the cost of fairness." arXiv preprint arXiv:1701.08230 (2017).

9. Calmon, Flavio P., Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. "Optimized Data Pre-Processing for Discrimination Prevention." In Advances in Neural Information Processing Systems, 2017.