CS 4873: Computing, Society & Professionalism

Munmun De Choudhury | Assistant Professor | School of Interactive Computing

Week 15: Algorithmic Bias and Fairness

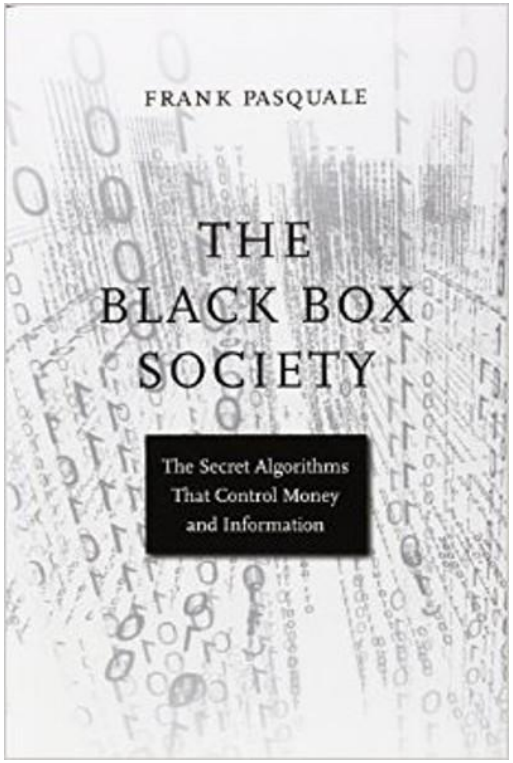
April 12, 2020

Final exam review –
no in-class meeting

XKCD (2017)

https://www.xkcd.com/1838/

Summary



Algorithms are "black boxes" protected by

- Industrial secrecy
- Legal protections
- Intentional obfuscation

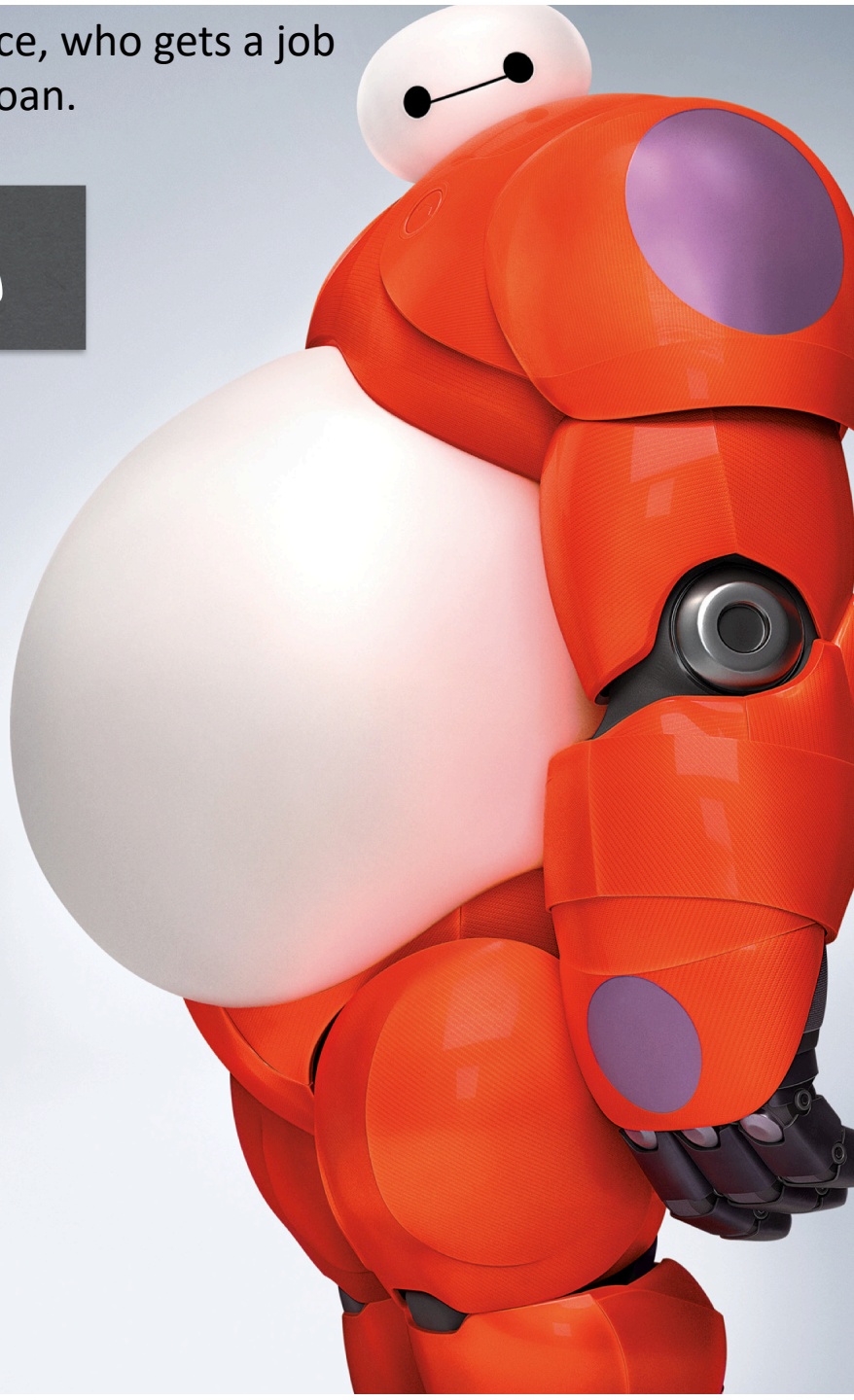
Discrimination becomes invisible

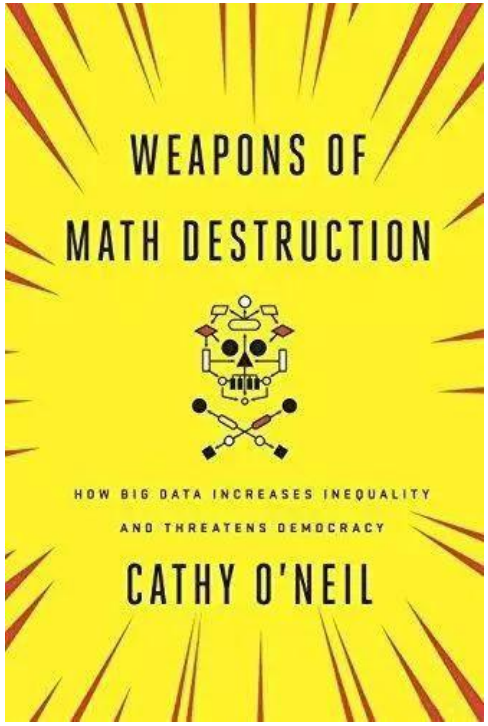
Mitigation becomes impossible

F. Pasquale (2015): The Black Box Society. Harvard University Press.

Proprietary algorithms are used to decide, for instance, who gets a job interview, who gets granted parole, and who gets a loan.

Human(bias) and Algorithms





Cathy O'Neil, a mathematician and the author of *Weapons of Math Destruction*, a book that highlights the risk of algorithmic bias in many contexts, says people are often too willing to trust in mathematical models because they believe it will remove human bias.

Two areas of concern: data and algorithms

Data inputs:

- Poorly selected (e.g., observe only car trips, not bicycle trips)
- Incomplete, incorrect, or outdated
- Selected with bias (e.g., smartphone users)
- Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")

Algorithmic processing:

- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Executive Office of the US President (May 2016): "Big Data: A Report on Algorithmic Systems,Opportunity,and Civil Rights"

Judiciary use of COMPAS scores

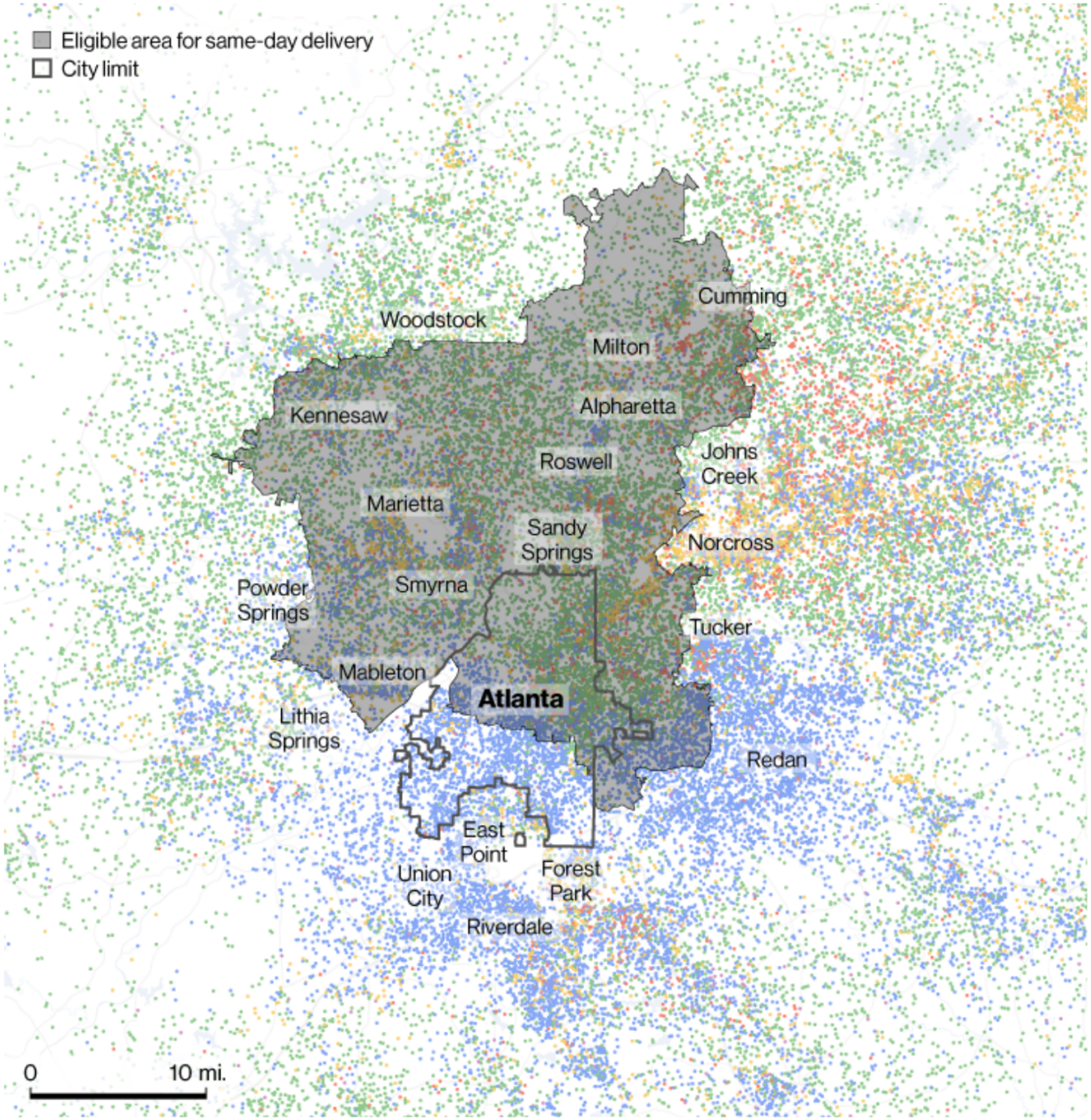
COMPAS (Correctional Offender Management Profiling for Alternative Sanctions): 137-questions questionnaire and predictive model for "risk of recidivism"

Prediction accuracy of recidivism for blacks and whites is about 60%, but ...

- Blacks that did not reoffend
  were classified as high risk twice as much as whites that did not reoffend
- Whites who did reoffend
  were classified as low risk twice as much as blacks who did reoffend

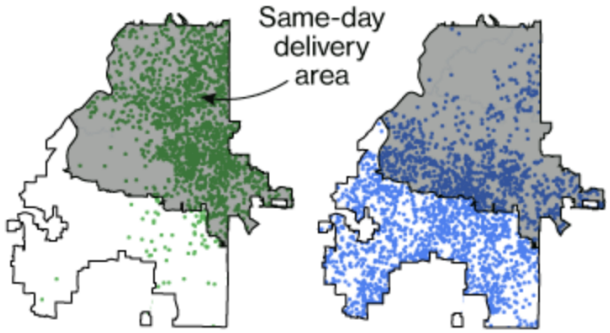
Pro Publica, May 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

All | White | Black | Hispanic | Asian | Other | Labels

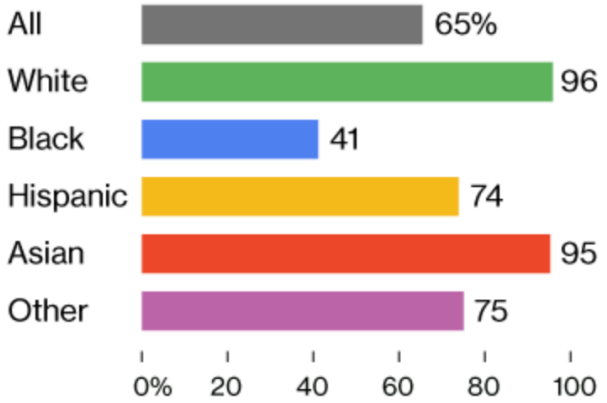


The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.

White residents **Black residents**



Percentage of residents living in ZIP codes with same-day delivery



Population percentages are based on American Community Survey estimates and have a 90% confidence interval.

The ethical challenges

- Algorithmic bias is shaping up to be a major societal issue at a critical moment in the evolution of machine learning and AI.
- If the bias lurking inside the algorithms that make ever-more-important decisions goes unrecognized and unchecked, it could have serious negative consequences, especially for marginalized communities and minorities.

Some case studies of algorithmic bias

American Economic Journal: Applied Economics 2017, 9(2): 1–22
https://doi.org/10.1257/app.20160213

Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment†

By Benjamin Edelman, Michael Luca, and Dan Svirsky*

In an experiment on Airbnb, we find that applications from guests with distinctively African American names are 16 percent less likely to be accepted relative to identical guests with distinctively white names. Discrimination occurs among landlords of all sizes, including small landlords sharing the property and larger landlords with multiple properties. It is most pronounced among hosts who have never had an African American guest, suggesting only a subset of hosts discriminate. While rental markets have achieved significant reductions in discrimination in recent decades, our results suggest that Airbnb's current design choices facilitate discrimination and raise the possibility of erasing some of these civil rights gains.
(*JEL* C93, J15, L83)

Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment

- Experimental study on Airbnb showing that applications from guests with distinctively African-American names are 16% less likely to be accepted relative to identical guests with distinctively White names.
- Discrimination occurs among landlords of all sizes, including small landlords sharing the property and larger landlords with multiple properties.
- **Both African-American and White hosts discriminate against African-American guests; both male and female hosts discriminate; both male and female African-American guests are discriminated against.**
- Airbnb's current design choices facilitate discrimination and raise the possibility of erasing some of these civil rights gains.

REPORTS | PSYCHOLOGY

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]
+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Article | Figures & Data | Info & Metrics | eLetters | PDF

Machines learn what people know implicitly

AlphaGo has demonstrated that a machine can learn how to do things that people spend many years of concentrated study learning, and it can rapidly learn how to do them better than any human can. Caliskan *et al.* now show that machines can learn word associations from written texts and that these associations mirror those learned by humans, as measured by the Implicit Association Test (IAT) (see the Perspective by Greenwald). Why does this matter? Because the IAT has predictive value in uncovering the association between concepts, such as pleasantness and flowers or unpleasantness and insects. It can also tease out attitudes and beliefs—for example, associations between female names and family or male names and career. Such biases may not be expressed explicitly, yet they can prove influential in behavior.

Science, this issue p. 183; see also p. 133

Science
Vol 356, Issue 6334
14 April 2017

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

ARTICLE TOOLS
Email
Print
Alerts
Citation tools
Download Powerpoint
Save to my folders
Request Permissions
Share

Major Finding

- Word embedding models
- The paper shows that some more troubling implicit biases seen in human psychology experiments are also readily acquired by algorithms. The words “female” and “woman” were more closely associated with arts and humanities occupations and with the home, while “male” and “man” were closer to math and engineering professions.
- And the AI system was more likely to associate European American names with pleasant words such as “gift” or “happy”, while African American names were more commonly associated with unpleasant words.

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations

Matthew Kay
Computer Science & Engineering | dub, University of Washington
mjskay@uw.edu

Cynthia Matuszek
Computer Science & Electrical Engineering, University of Maryland Baltimore County
cmat@umbc.edu

Sean A. Munson
Human-Centered Design & Engineering | dub, University of Washington
smunson@uw.edu

ABSTRACT
Information environments have the power to affect people's perceptions and behaviors. In this paper, we present the results of studies in which we characterize the gender bias present in image search results for a variety of occupations. We experimentally evaluate the effects of bias in image search results on the images people choose to represent those careers and on people's perceptions of the prevalence of men and women in each occupation. We find evidence for both stereotype exaggeration and systematic underrepresentation of women in search results. We also find that people rate search results higher when they are consistent with stereotypes for a career, and shifting the representation of gender in image search results can shift people's perceptions about real-world distributions. We also discuss tensions between desires for high-quality results and broader

tional choices, opportunities, and compensation [20,26]. Stereotypes of many careers as gender-segregated serve to reinforce gender sorting into different careers and unequal compensation for men and women in the same career. Cultivation theory, traditionally studied in the context of television, contends that both the prevalence and characteristics of media portrayals can develop, reinforce, or challenge viewers' stereotypes [29].

Inequality in the representation of women and minorities, and the role of online information sources in portraying and perpetuating it, have not gone unnoticed in the technology community. This past spring, Getty Images and LeanIn.org announced an initiative to increase the diversity of working women portrayed in the stock images and to improve how they are depicted [27]. A recent study identified discrimination in online advertising delivery when searching for

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations

- Algorithms can be biased in how they represent the world.
- The information people access affects their understanding of the world around them and the decisions they make: biased information can affect both how people treat others and how they evaluate their own choices or opportunities.
- The paper experimentally evaluates the effects of bias in image search results on the images people choose to represent those careers and on people's perceptions of the prevalence of men and women in each occupation.

Findings

- Stereotype exaggeration: Results for many occupations exhibit a slight exaggeration of gender ratios according to stereotype: e.g., male-dominated professions tend to have even more men in their results
- Systematic over-/under- representation: Search results also exhibit a slight under-representation of women in images, such that an occupation with 50% women would be expected to have about 45% women in the results on average.

Findings

- Qualitative differential representation: Image search results also exhibit biases in how genders are depicted: **those matching the gender stereotype of a profession tend to be portrayed as more professional-looking** and less inappropriate-looking.
- Perceptions of occupations in search results: We find that people's existing perceptions of gender ratios in occupations are quite accurate, but that manipulated search results can have a small but significant effect on perceptions, shifting estimations on average ~7%.

On the web: race and gender stereotypes reinforced

- Results for "CEO" in Google Images: 11% female, US 27% female CEOs
  - Also in Google Images, "doctors" are mostly male, "nurses" are mostly female
- Google search results for professional vs. unprofessional hairstyles for work

Image results: "Unprofessional hair for work"

Image results: "Professional hair for work"

M. Kay, C. Matuszek, S. Munson (2015): Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. CHI'15.

Scholarly criticism of bias due to a lack of algorithmic transparency

- Joanna Bryson, a computer scientist at the University of Bath and a co-author, said: “A lot of people are saying this is showing that AI is prejudiced. No. This is showing we’re prejudiced and that AI is learning it.”
- But Bryson warned that AI has the potential to reinforce existing biases because, unlike humans, algorithms may be unequipped to consciously counteract learned biases. “A danger would be if you had an AI system that didn’t have an explicit part that was driven by moral ideas, that would be bad,” she said.

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

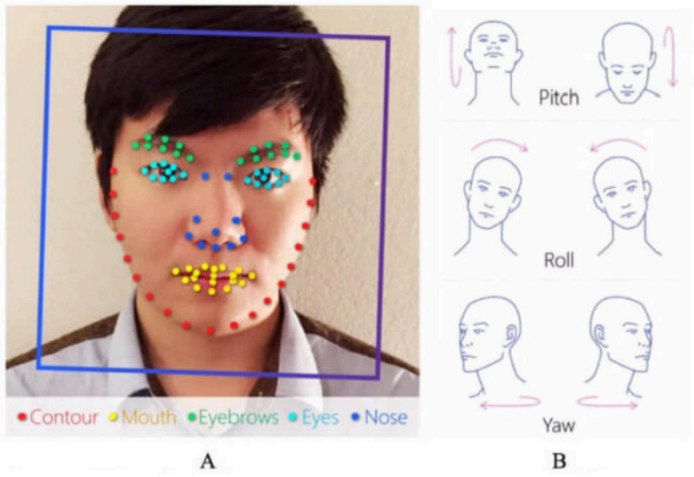
- Authors used deep neural networks to extract features from 35,326 facial images.
  - Images scraped from public profiles posted on a U.S. dating website
- These features were entered into a logistic regression aimed at classifying sexual orientation.
- Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women.
- The authors claimed that their findings therefore provided “strong support” for the idea that sexual orientation stems from hormone exposure in the womb

SCIENCE

The Study Claiming AI Can Tell If You're Gay or Straight Is Now Under Ethical Review

By Lisa Ryan @lisarya

SEPTEMBER 12, 2017
6:21 PM



An image from the study. Photo: Journal of Personality and Social Psychology/Stanford University

A recent Stanford University study published in the *Journal of Personality and Social Psychology* claimed artificial intelligence can figure out if a person is gay or straight by analyzing pictures of their faces. However, the Outline reports the study was met with "immediate backlash" from the AI community, academics, and LGBTQ advocates alike — and the paper is now under ethical review.

Some argued that the study is just the latest example of a disturbing technology-fueled revival of physiognomy, the long discredited notion that personality traits can be revealed by measuring the size and shape of a person's eyes, nose and face.

Class Discussion:

What kind of biases can this sexual orientation detector that uses facial images introduce in platforms that rely on profiling users, for example, for ad placement?

Artificial intelligence

DeepMind's new AI ethics unit is the company's next big move

Google-owned DeepMind has announced the formation of a major new AI research unit comprised of full-time staff and external advisors

By **JAMES TEMPERTON**

Wednesday 4 October 2017

Google DeepMind

Job Openings

Artificial Intelligence/FutureTech Investigative Reporter

New York, NY

Posted 30+ Days Ago

Full time

REQ-001480

Apply | Apply with LinkedIn

Job Description

Investigate how algorithms, artificial intelligence, robots and technology are influencing our lives, our businesses, our privacy and the future.

This deeply-informed reporter will be able to understand and explain complex technologies while investigating the people and companies behind them. They will be expected to discover and cultivate sources and contacts and to break ground reporting on issues that many companies would rather go uncovered. They will also be comfortable with - and even capable of - a variety of computer-assisted reporting techniques. The reporter will work on a small team and be interested in telling stories through multiple mediums including interactive graphics, virtual reality, audio, video and of course the written word.

Location is flexible.

This is a guild position.

To apply:

New York Times

About Us

Help shape the future Times

This is an important moment to v
organization, we're taking advant
landscape to pioneer a new era o
original reporting at our core, we'
about our reader relationships an
vant offerings and experiences. V

Research Attention

- In 2017, a group of researchers, together with the American Civil Liberties Union, launched an effort to identify and highlight algorithmic bias, called AI Now


MAY
2018
HISTORY
Right to be Forgotten


MAY
2018
Right to Explanation