CS 3001-C: Computing, Society, and Professionalism

Munmun De Choudhury | Associate Professor | School of Interactive Computing

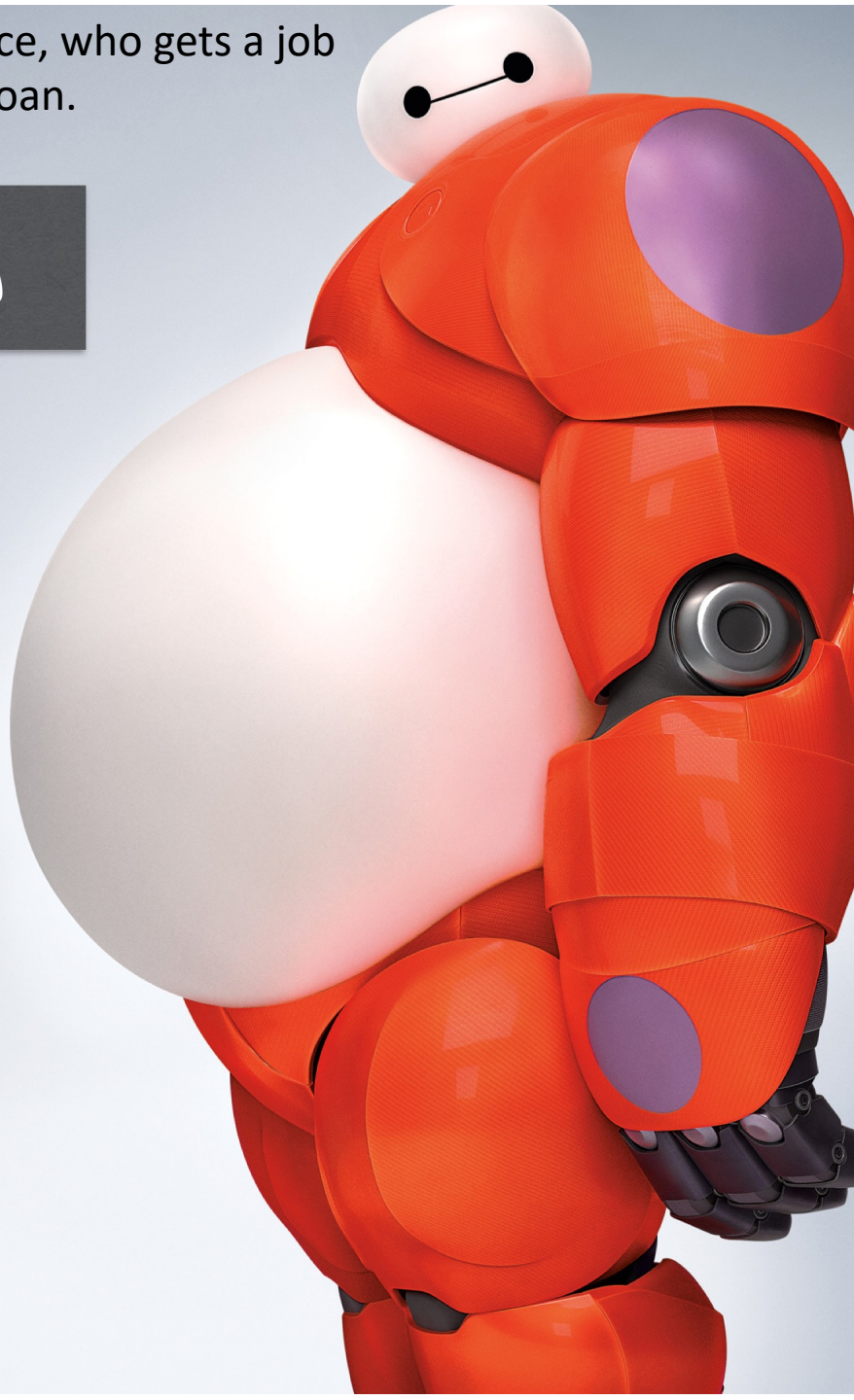
Week 14: Algorithmic Bias and Fairness

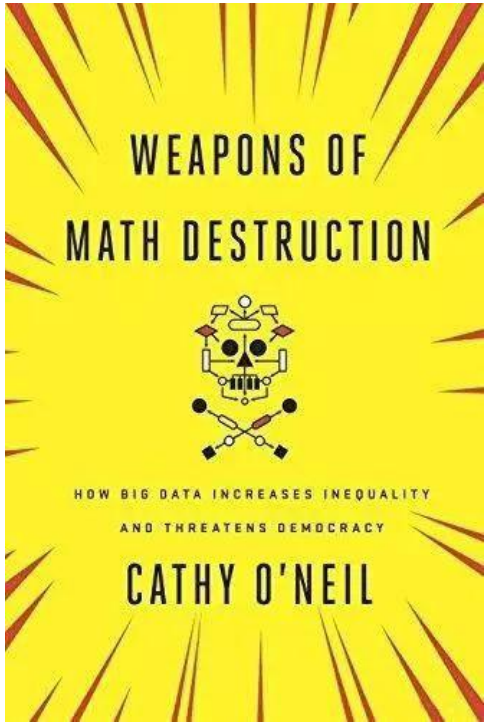
April 12, 2022


Machine Learning is Everywhere

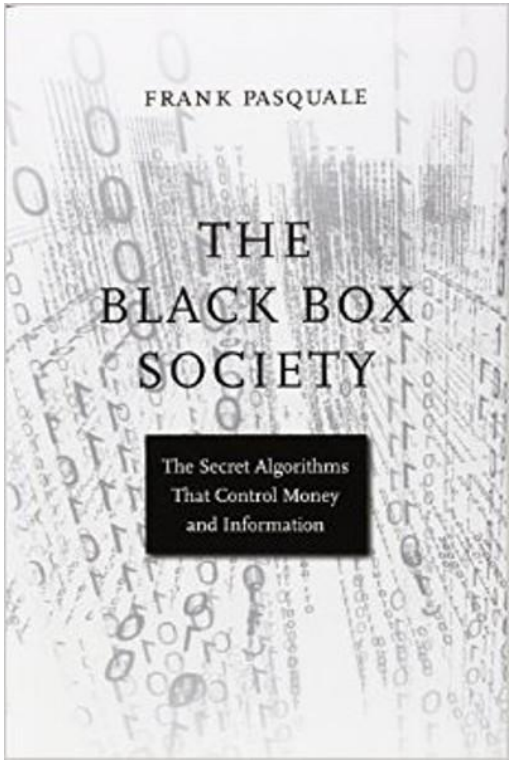
Proprietary algorithms are used to decide, for instance, who gets a job interview, who gets granted parole, and who gets a loan.

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.



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.

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

The ethical challenges

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)

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

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

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 discrimina-

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"

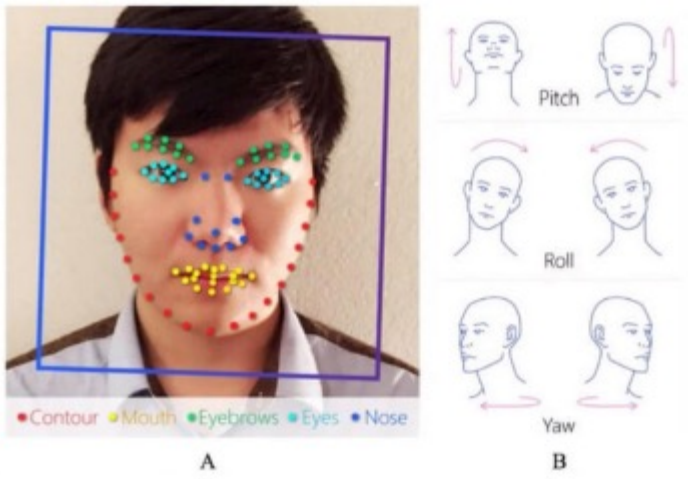
M. Kay, C. Matuszek, S. Munson (2015): Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. CHI'15.

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.

sociological science

Gaydar and the Fallacy of Decontextualized Measurement

Andrew Gelman,[a] Greggor Mattson,[b] Daniel Simpson[c]

a) Columbia University; b) Oberlin College; c) University of Toronto

Abstract: Recent media coverage of studies about "gaydar," the supposed ability to detect another's sexual orientation through visual cues, reveal problems in which the ideals of scientific precision strip the context from intrinsically social phenomena. This fallacy of objective measurement, as w term it, leads to nonsensical claims based on the predictive accuracy of statistical significance. We interrogate these gaydar studies' assumption that there is some sort of pure biological measu of perception of sexual orientation. Instead, we argue that the concept of gaydar inherently exist within a social context and that this should be recognized when studying it. We use this case as an example of a more general concern about illusory precision in the measurement of social phenomena

Automatic Crime Prediction using Events Extracted from Twitter Posts

Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown

Department of Systems and Information Engineering, University of Virginia
{xw4u,msg8u,brown}@virginia.edu

Abstract. Prior work on criminal incident prediction has relied primarily on the historical crime record and various geospatial and demographic information sources. Although promising, these models do not take into account the rich and rapidly expanding social media context that surrounds incidents of interest. This paper presents a preliminary investigation of Twitter-based criminal incident prediction. Our approach is based on the automatic semantic analysis and understanding of natural language Twitter posts, combined with dimensionality reduction via latent Dirichlet allocation and prediction via linear modeling. We tested our model on the task of predicting future hit-and-run crimes. Evaluation results indicate that the model comfortably outperforms a baseline model that predicts hit-and-run incidents uniformly across all days.

1 Introduction

Traditional crime prediction systems (e.g., the one described by Wang and Brown [14]) make extensive use of historical incident patterns as well as layers of in-

Discussion Point:

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

Scholarly criticism of bias due to a lack of algorithmic transparency

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

danah boyd & Kate Crawford

CRITICAL QUESTIONS FOR BIG DATA

Provocations for a cultural, technological, and scholarly phenomenon

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech? Will it transform how we study human communi-

Media, Culture & Society

2.000 Impact Factor
5-Year Impact Factor 1.929
Journal Indexing & Metrics »

Deeper data: a response to boyd and Crawford

Andre Brock

First Published August 24, 2015 | Research Article | Check for updates
https://doi.org/10.1177/0163443715594105

Abstract

Data analysis of any sort is most effective when researchers first take account of the complex ideological processes underlying data's originating impetus, selection bias, and semiotic affordances of the information and communication technologies (ICTs) under examination.

Keywords

Big Data, critical cultural informatics, critical information studies, data and society, digital sociology, social media and society

In 2013, Lois Scheidt and I organized a panel for the International Congress of Qualitative Inquiry titled 'Small data in a big data world' as a response to 'Six Provocations for Big Data'. Our panelists presented incredible work conceptualizing new approaches in an age of 'big data' to qualitative social media research,

Aequitas: A Bias and Fairness Audit Toolkit

Pedro Saleiro PEDROSALEIRO@GMAIL.COM
Benedict Kuester
Loren Hinkson
Jesse London
Abby Stevens
Ari Anisfeld
Kit T. Rodolfa
Rayid Ghani RAYID@UCHICAGO.EDU
Center for Data Science and Public Policy
University of Chicago
Chicago, IL 60637, USA

Editor:

Abstract

Recent work has raised concerns on the risk of unintended bias in AI systems being used nowadays that can affect individuals unfairly based on race, gender or religion, among other possible characteristics. While a lot of bias metrics and fairness definitions have been proposed in recent years, there is no consensus on which metric/definition should be used and there are very few available resources to operationalize them. Therefore, despite recent awareness, auditing for bias and fairness when developing and deploying AI systems is not yet a standard practice. We present Aequitas, an open source bias and fairness audit toolkit that was released in 2018 and it is an intuitive and easy to use addition to the machine

AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias

Rachel K. E. Bellamy [1] **Kuntal Dey** [2] **Michael Hind** [1] **Samuel C. Hoffman** [1] **Stephanie Houde** [1]
Kalapriya Kannan [3] **Pranay Lohia** [3] **Jacquelyn Martino** [1] **Sameep Mehta** [3] **Aleksandra Mojsilovic** [1]
Seema Nagar [3] **Karthikeyan Natesan Ramamurthy** [1] **John Richards** [1] **Diptikalyan Saha** [3] **Prasanna Sattigeri** [1]
Moninder Singh [1] **Kush R. Varshney** [1] **Yunfeng Zhang** [1]

Abstract

Fairness is an increasingly important concern as machine learning models are used to support decision making in high-stakes applications such as mortgage lending, hiring, and prison sentencing. This paper introduces a new open source Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360), released under an Apache v2.0 license (https://github.com/ibm/aif360). The main objectives of this toolkit are to help facilitate the transition of fairness research algorithms to use in an industrial setting and to provide a common framework for fairness researchers to share and evaluate algorithms.

The package includes a comprehensive set of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. It also includes an interactive Web experience (https://aif360.mybluemix.net) that provides a gentle introduction to the concepts and capabilities for line-of-business users, as well as extensive documentation, usage guidance, and industry-specific tutorials to enable data scientists and practitioners to incorporate the most appropriate tool for their problem into their work products. The architecture of the package has been engineered to conform to a standard paradigm used in data science, thereby further improving usability for practitioners. Such architectural design and abstractions enable researchers and developers to extend the toolkit with their new algorithms and improvements, and to use it for

MAY
2018

Right to be Forgotten


MAY
2018
Right to Explanation

From Treatment to Healing: Envisioning a Decolonial Digital Mental Health

Sachin R. Pendse
Georgia Institute of Technology
Atlanta, GA, USA
sachin.r.pendse@gatech.edu

Daniel Nkemelu
Georgia Institute of Technology
Atlanta, GA, USA
dnkemelu@gatech.edu

Nicola J. Bidwell
IUM, Windhoek, Namibia
Aalborg University, Copenhagen, Denmark
nic.bidwell@gmail.com

Sushrut Jadhav
University College London
London, UK
s.jadhav@ucl.ac.uk

Soumitra Pathare
Centre for Mental Health Law and Policy
Pune, India
spathare@cmhlp.org

Munmun De Choudhury*
Georgia Institute of Technology
Atlanta, GA, USA
munmund@gatech.edu

Neha Kumar*
Georgia Institute of Technology
Atlanta, GA, USA
neha.kumar@gatech.edu

ABSTRACT

The field of digital mental health is making strides in the application of technology to broaden access to care. We critically examine how these technology-mediated forms of care might amplify historical injustices, and erase minoritized experiences and expressions of mental distress and illness. We draw on decolonial thought and critiques of identity-based algorithmic bias to analyze the underlying power relations impacting digital mental health technologies today, and envision new pathways towards a decolonial digital mental health. We argue that a decolonial digital mental health is one that centers lived experience over rigid classification, is conscious of

ACM Reference Format:
Sachin R. Pendse, Daniel Nkemelu, Nicola J. Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing: Envisioning a Decolonial Digital Mental Health. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3491102.3501982

Content Warning: This work includes descriptions of mental illness, involuntary hospitalization, and suicide. This work also includes descriptions of colonialism, racism, slavery, and police brutality in the context of mental health. Additionally, Aboriginal