

CS 4873-A: Computing and Society

Munmun De Choudhury | Associate Professor | School of Interactive Computing



Week 14: Algorithmic Bias and Fairness

April 18, 2021

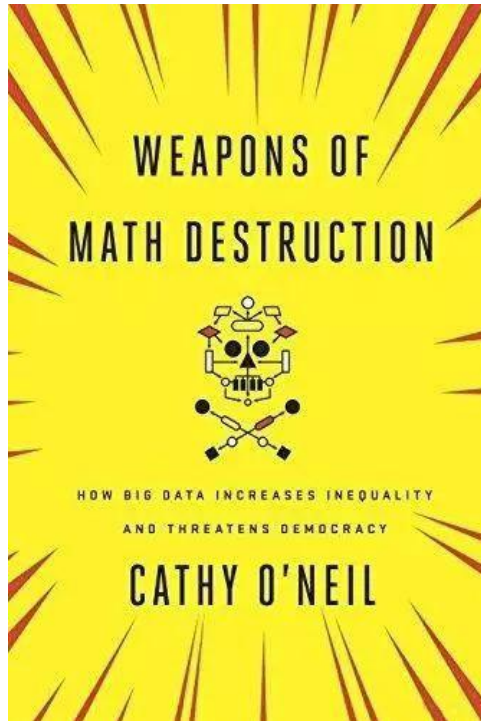




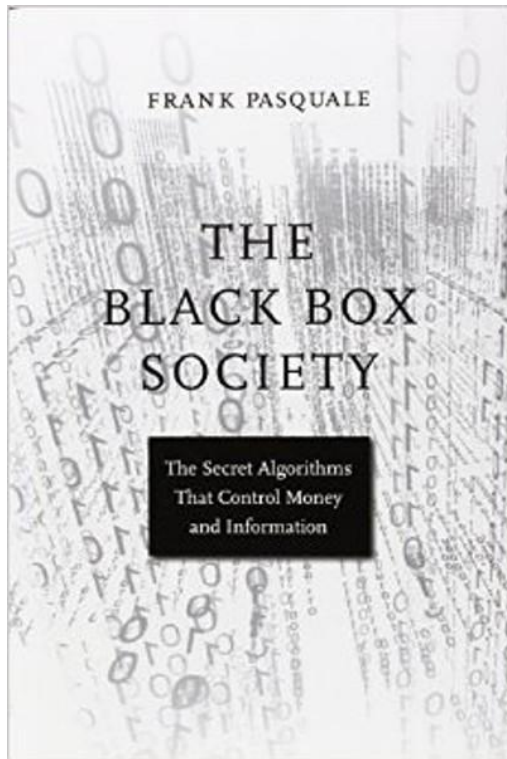
Proprietary algorithms are used to decide, for instance, who gets a job interview, who gets granted parole, and who gets a loan.

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)

VERNON PRATER

Prior Offenses

2 armed robberies, 1 attempted armed robbery

Subsequent Offenses

1 grand theft

LOW RISK

3

BRISHA BORDEN

Prior Offenses

4 juvenile misdemeanors

Subsequent Offenses

None

HIGH RISK

8



JAMES RIVELLI

LOW RISK

3



ROBERT CANNON

MEDIUM RISK

6

JAMES RIVELLI

Prior Offenses

1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses

1 grand theft

LOW RISK

3

ROBERT CANNON

Prior Offense

1 petty theft

Subsequent Offenses

None

MEDIUM RISK

6

DYLAN FUGETT

LOW RISK

3

BERNARD PARKER

HIGH RISK

10



The ethical challenges



Some case studies of algorithmic bias

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)



2019 School of Information Science and Technology
First Young Scholars Forum
Welcome to Fancy Shanghai

Institution: GEORGIA INSTITUTE OF TECHNOLOGY
Log in | My account | Contact Us
GEORGIA INSTITUTE OF TECHNOLOGY



Become a r

Renew my subs
Sign up for new

SHARE

REPORTS | PSYCHOLOGY



1.02k



0

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.1244230

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

Advertisement

Excel + Tableau:
A Beautiful
Partnership



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"



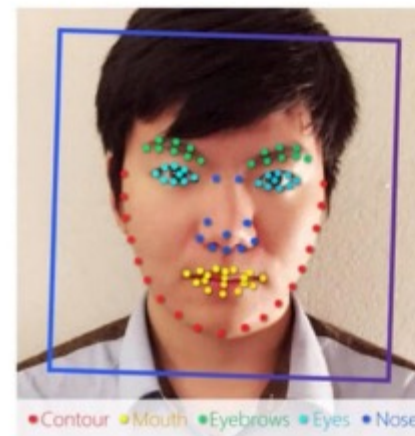


Scholarly criticism of bias due to a lack of
algorithmic transparency

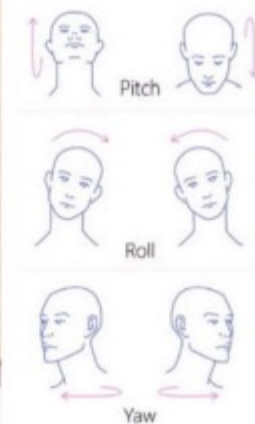
The Study Claiming AI Can Tell If You're Gay or Straight Is Now Under Ethical Review

By Lisa Ryan  @lisarya

SEPTEMBER 12,
2017
6:21 PM



A



B

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.



Gaydar and the Fallacy of Decontextualized Measurement

Andrew Gelman,^a Gregg 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 we 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 measure of perception of sexual orientation. Instead, we argue that the concept of gaydar inherently exists 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



Discussion Point:

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?

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-

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

An illustration featuring two stylized, orange-toned faces at the bottom, looking upwards. Above them is a dense, interconnected network of colorful gears in shades of red, orange, yellow, and teal. Dashed lines with arrows indicate the flow and connection between the gears, symbolizing complex AI systems or neural networks.

Google DeepMind



Job Openings

Artificial Intelligence/FutureTech Investigative Reporter

📍 New York, NY

Apply

Apply with LinkedIn

🕒 Posted 30+ Days Ago

📄 Full time

📄 REQ-001480

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.

New York Times

To apply:

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-

Article Menu

[Close](#) ^[Download PDF](#)[Open EPUB](#)

Did you struggle to get access to this article? This product could help you

[Full Article](#)

Content List

[Abstract](#)[Notes](#)[References](#)

Deeper data: a response to boyd and Crawford

[Andre Brock](#)

First Published August 24, 2015 | Research Article



<https://doi.org/10.1177/0163443715594105>

[Article information](#) ▾

5



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,

MAY
2018

HISTORY

Right to be Forgotten



MAY
2018



Right to Explanation

CS 4873-A: Computing and Society

Munmun De Choudhury | Associate Professor | School of Interactive Computing

Week 14: Research Ethics

April 18, 2021

Research Ethics

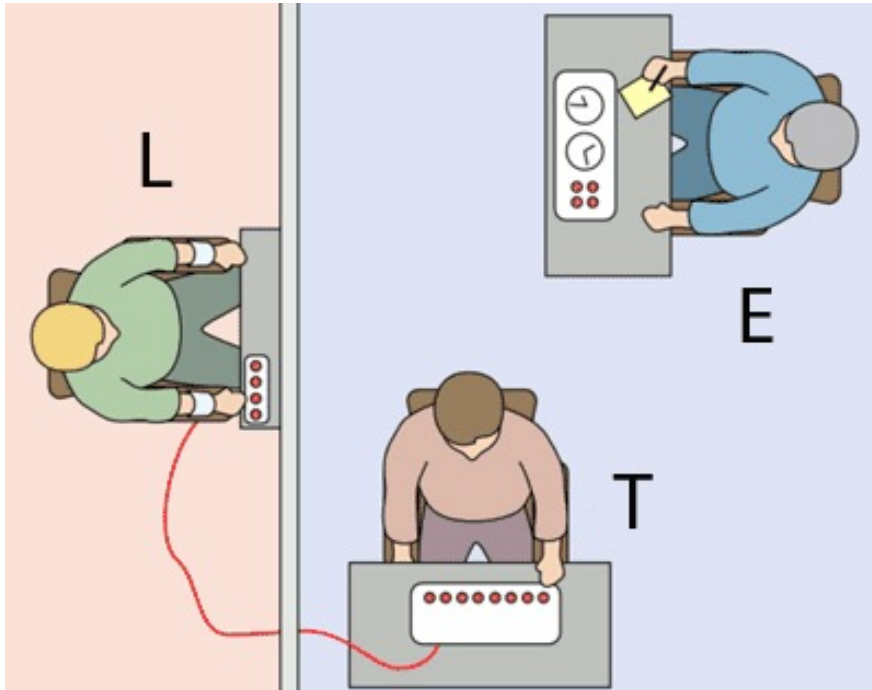
Tuskegee Syphilis Experiment

For the most part, doctors and civil servants simply did their jobs. Some merely followed orders, others worked for the glory of science.

— John R. Heller Jr., Director of the Public Health Service's Division of Venereal Diseases

<https://www.youtube.com/watch?v=OyedeuJOGgl>

Milgram's Obedience Study



- Experiment on obedience to authority figures
- Study measured the willingness of study participants, men from a diverse range of occupations with varying levels of education, to obey an authority figure who instructed them to perform acts conflicting with their personal conscience
- 65% (two-thirds) of participants (i.e., teachers) continued to the highest level of 450 volts. All the participants continued to 300 volts

<https://www.youtube.com/watch?v=mOUEC5YXV8U&t=6s>

Ethical Issues

- Deception
- Protection of participants
- Right to withdrawal

Institutional Review Boards

- Formal review procedures for institutional human subject studies were originally developed in direct response to research abuses in the 20th century.

[About OHRP](#)[Regulations & Policy](#)[Education & Outreach](#)[Compliance & Reporting](#)[News](#)[Register IRBs & Obtain FWAs](#)[SACHRP Committee](#)[International](#)[HHS Home](#) > [OHRP](#) > [Regulations & Policy](#) > [Regulations](#) > Federal Policy for the Protection of Human Subjects ('Common Rule')[Statutes](#)[Belmont Report](#)[Regulations](#)[45 CFR 46](#)[Common Rule](#)[FDA](#)[Final Rule](#)[Guidance](#)[Requests for Comments](#)Text Resize **A A A**

Print

Share



Federal Policy for the Protection of Human Subjects ('Common Rule')

The current U.S. system of protection for human research subjects is heavily influenced by the [Belmont Report](#), written in 1979 by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report outlines the basic ethical principles in research involving human subjects. In 1981, with this report as foundational background, HHS and the Food and Drug Administration revised, and made as compatible as possible under their respective statutory authorities, their existing human subjects regulations.

The Federal Policy for the Protection of Human Subjects or the "Common Rule" was published in 1991 and codified in separate regulations by 15 Federal departments and agencies, as listed below. The HHS regulations, [45 CFR part 46](#), include four subparts: subpart A, also known as the Federal Policy or the "Common Rule"; subpart B, additional protections for pregnant women, human fetuses, and neonates; subpart C, additional protections for prisoners; and subpart D, additional protections for children. Each agency includes in its chapter of the Code of Federal Regulations [CFR] section numbers and language that are identical to those of the HHS codification at 45 CFR part 46, subpart A.

IRB Oversight

Adapting IRB review to Internet era and big data research



Home

Articles

Front Matter

News

Podcasts

Authors

NEW RESEARCH IN

Physical Sciences

Social Sciences

Biological Sciences

Experimental evidence of massive-scale emotional contagion through social networks



Article Alerts

Share

Email Article

Tweet

Citation Tools

Like 1.1K

Request Permissions

Mendeley

Adam D. I. Kramer, Jamie E. Guillory and Jeffrey T. Hancock

PNAS June 17, 2014. 111 (24) 8788-8790; published ahead of print June 2, 2014.

<https://doi.org/10.1073/pnas.1320040111>

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

This article has corrections. Please see:

Editorial Expression of Concern: Experimental evidence of massive-scale emotional contagion through social networks

Correction for Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks

▶ More Articles of This Classification

Social Sciences

Identifying psychological responses of stigmatized groups to referendums

Grassland biodiversity can pay

Cracking the social code of speech prosody using reverse correlation

Show more

Psychological and Cognitive Sciences

TECHNOLOGY

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By VINDU GOEL JUNE 29, 2014






216


Facebook revealed that it had altered the news feeds of over half a million users in its study.

Karen Bleier/Agence France-Presse — Getty Images

To Facebook, we are all lab rats.

Facebook routinely adjusts its users' news feeds — testing out the number of ads they see or the size of photos that appear — often without their knowledge. It is all for the purpose, the company says, of creating a more alluring and useful product.

But last week, Facebook revealed that it had manipulated the news

RECENT COMMENTS

GSP13 July 1, 2014

Shocked that this study - at least from what I can tell - was not subjected to an IRB.

Superpower July 1, 2014

"...my co-authors and I are very sorry for the way the paper described the research and any anxiety it caused," -once again the progressive,...

Faith July 1, 2014

Just another vindication for dropping out of FB months ago. My emotion? Never been happier.

Example concerns

- Violation of the rights of research subjects



Unexpected expectations: Public reaction to the Facebook emotional contagion study

new media & society

1–19

© The Author(s) 2019

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/1461444819876944

journals.sagepub.com/home/nms



Blake Hallinan 

Jed R Brubaker

Casey Fiesler

University of Colorado Boulder, USA

Abstract

How to ethically conduct online platform-based research remains an unsettled issue and the source of continued controversy. The Facebook emotional contagion study, in which researchers altered Facebook News Feeds to determine whether

Highlights of some findings...

- **Living in a lab**

- *Dear Mr. Zuckerberg, Last I checked, we did not decide to jump in a petri dish to be utilized at your disposal . . . We connect with our loved ones.*

- **Manipulation anxieties**

- *Don't be fooled, manipulating a mood is the ability to manipulate a mind. Political outcomes, commerce, and civil unrest are just a short list of things that can be controlled.*

- **Wake up, sheeple**

- *Anyone who doesn't realise that anything you put "out there" on Facebook (or any other social media site) is like shouting it through a bullhorn should have their internet competency licence revoked. We can't blame all stupidity on some or other conspiracy...*

- **No big deal**

- *A/B testing (i.e. basically what happened here) when software companies change content or algorithms for a subset of users happens *all the time*. It's standard industry practice.*

A key takeaway – consent is important!

Consent at Scale – why it is hard

“Participant” Perceptions of Twitter Research Ethics

Casey Fiesler¹ and Nicholas Proferes² 

Social Media + Society
January-March 2018: 1–14
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2056305118763366
journals.sagepub.com/home/sms



Abstract

Social computing systems such as Twitter present new research sites that have provided billions of data points to researchers. However, the availability of public social media data has also presented ethical challenges. As the research community works to create ethical norms, we should be considering users' concerns as well. With this in mind, we report on an exploratory survey of Twitter users' perceptions of the use of tweets in research. Within our survey sample, few users were previously aware that their public tweets could be used by researchers, and the majority felt that researchers should not be able to use tweets without consent. However, we find that these attitudes are highly contextual, depending on factors such as how the research is conducted or disseminated, who is conducting it, and what the study is about. The findings of this study point to potential best practices for researchers conducting observation and analysis of public data.

Keywords

Twitter, Internet research ethics, social media, user studies

“Participant” Perceptions of Twitter Research Ethics

Social Media + Society
January-March 2018: 1–14
© The Author(s) 2018
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2056305118763366
journals.sagepub.com/home/sms


Casey Fiesler¹ and Nicholas Proferes² 

Table 2. Comfort Around Tweets Being Used in Research.

Question	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
How do you feel about the idea of tweets being used in research? (<i>n</i> = 268)	3.0%	17.5%	29.1%	35.1%	15.3%
How would you feel if a tweet of yours was used in one of these research studies? (<i>n</i> = 267)	4.5%	22.5%	23.6%	33.3%	16.1%
How would you feel if your entire Twitter history was used in one of these research studies? (<i>n</i> = 268)	21.3%	27.2%	18.3%	21.6%	11.6%

Note. The shading was used to provide a visual cue about higher percentages.

The Case of Deleted Tweets/Social media posts

Tweets Are Forever: A Large-Scale Quantitative Analysis of Deleted Tweets

Hazim Almuhiemedi^a, Shomir Wilson^a, Bin Liu^a, Norman Sadeh^a, Alessandro Acquisti^b

^aSchool of Computer Science, ^bHeinz College

Carnegie Mellon University

{hazim,shomir,blui1,sadeh}@cs.cmu.edu, acquisti@andrew.cmu.edu

ABSTRACT

This paper describes an empirical study of 1.6M deleted tweets collected over a continuous one-week period from a set of 292K Twitter users. We examine several aggregate properties of deleted tweets, including their connections to other tweets (e.g., whether they are replies or retweets), the clients used to produce them, temporal aspects of deletion, and the presence of geotagging information. Some significant differences were discovered between the two collections, namely in the clients used to post them, their conversational aspects, the sentiment vocabulary present in them, and the days of the week they were posted. However, in other dimensions for which analysis was possible, no substantial differences were found. Finally, we discuss some ramifications of this work for understanding Twitter usage and management of one's privacy.

in other cases they may have serious ramifications, as recognized by the European Commission's draft of a "right to be forgotten" [1].

When a post is deleted from an online social network, users generally assume that the post will no longer be available for anyone to see. However, this is not necessarily true, as evidence may persist of the post and its content in less visible ways. Twitter, through its API service, provides a particularly rich and accessible stream of data on deleted posts. By following the posts (*tweets*) of a user and other messages from the API, one can reconstruct which tweets the user decides to delete without losing any data associated with them. By tracking a large number of users whose posts are public, it is thus possible to observe large-scale patterns in deletion behavior. These patterns can inform the design of online social networks to help users better manage their content.


Also what about those who can't give consent any more? *The case of dead people*

- Warning: I am not a historian ;-)
- Today's view
- Medieval view
- Things are muddled when it comes to dead people's digital lives – legislation has not kept up with technological change



Digital Wills and Beneficiaries (Forbes)

... still particularly nascent when it comes
to data stored by a third-party company



When there is no consent, researchers have poor understanding of what can go wrong, and “participants” or research subjects have limited understanding of risk.

What's at Stake: Characterizing Risk Perceptions of Emerging Technologies

Michael Skirpan

University of Colorado
Boulder, CO

michael.skirpan@colorado.edu

Tom Yeh

University of Colorado
Boulder, CO

tom.yeh@colorado.edu

Casey Fiesler

University of Colorado
Boulder, CO

casey.fiesler@colorado.edu

ABSTRACT

One contributing factor to how people choose to use technology is their perceptions of associated risk. In order to explore this influence, we adapted a survey instrument from risk perception literature to assess mental models of users and technologists around risks of emerging, data-driven technologies (e.g., identity theft, personalized filter bubbles). We surveyed 175 individuals for comparative and individual assessments of risk, including characterizations using psychological factors. We report our observations around group differences (e.g., expert versus non-expert) in how people assess risk, and what factors may structure their conceptions of technological harm. Our findings suggest that technologists see these risks as posing a bigger threat to society than do non-experts. Moreover, across groups, participants did not see technological risks as voluntarily assumed. Differences in how people characterize risk have implications for the future of design, decision-making, and public communications, which we discuss through a lens we call risk-sensitive design.

ACM Classification Keywords

H.1.2 User/Machine Systems: Human Factors; H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

and behavior-driven design. These users must rely on the companies and parties to whom they have given their data (knowingly or not) to be ethical.

Yet, we already know that many impacts (e.g., privacy, ethical, legal) and constraints (e.g., protocols, technological capabilities) of online technologies are poorly understood by users [24, 8, 36, 15]. We also know that, when asked, users are often uncomfortable or find undesirable the practices of online behavioral advertising (OBA) and personalization [37, 34]. This misalignment is often framed as a consumer trade-off between privacy and personal benefit [13, 40]. Framing it this way leads to an assumption that the benefit of web services must outweigh consumer's privacy concerns since users are not opting out of services.

However, if consumers really are performing this cost-benefit analysis and making a conscious decision, then why do we see such hype and panic around risks and harms caused by technology in the media? Daily news headlines relay injustice [19, 1, 4, 33], personal boundary violations [32], and gloom [26, 18, 14] over the impacts of technology on society. Some of these problems may indeed warrant concern from the public and social advocates; others might be overblown

What's at Stake: Characterizing Risk Perceptions of Emerging Technologies

Michael Skirpan

University of Colorado
Boulder, CO

michael.skirpan@colorado.edu

Tom Yeh

University of Colorado
Boulder, CO

tom.yeh@colorado.edu

Casey Fielser

University of Colorado
Boulder, CO

casey.fiesler@colorado.edu

	Non-Expert			Expert	
Rank	Risk	Mean Rank		Risk	Mean Rank
1	Identity Theft	5.000		Job Loss	5.769
2	Account Breach	6.101		Account Breach	6.385
3	Job Loss	7.678		Identity Theft	6.577
4	Hackivist Leak	7.980		Technology Divide	6.923
5	Auto-Drones	8.523		Bias Job Alg	7.192
6	Harassment	9.074		Discriminatory Crime Alg	7.231
7	Undisclosed third party	9.349		Hackivist Leak	7.231
8	DDoS	9.403		Filter Bubble	7.654
9	Nuclear Reactor Meltdown	9.644		DDoS	8.269
10	Discriminatory Crime Alg	9.758		Undisclosed third party	8.462
11	Research w/o Consent	10.141		Harassment	9.346
12	Bias Job Alg	10.154		Auto-Drones	9.808
13	Driverless Car Malfunction	10.315		Research w/o Consent	11.154
14	Technology Divide	10.765		Nude Photos	12.038
15	Plane Crash	11.060		Driverless Car Malfunction	12.269
16	Filter Bubble	11.362		Nuclear Reactor Meltdown	14.308
17	Nude Photos	11.846		Plane Crash	14.654
18	Vaccine	12.846		Vaccine	15.731

Figure 1. Average comparative risk ranking by non-experts vs experts where items with significant differences ($p < .05$ for two-tailed t-test) are highlighted.

Discussion Point 1

Internet companies “manipulate” what we see and read all the time. Google was doing it for years without getting into trouble. Why did this Facebook study generate so much criticism?

Discussion Point 2

Adopting the following ethical theories, discuss whether this Facebook study was ethical: a) Kantian perspective; b) social contract theory perspective; and c) rule utilitarian perspective

Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community

Online data create gray area

Is it feasible to collect informed consent?

Should you be more transparent about your research?

Who is being left out by your data collection strategies?

Isn't public data public?

Is it possible to truly anonymize a dataset?



Code	Definition	Example Statements
Public Data	Only using public data / public data being okay to collect and analyze	<i>In general, I feel that what is posted online is a matter of public record, though every case needs to be looked at individually in order to evaluate the ethical risks.</i>
Do No Harm	Comments related to Golden Rule	<i>Golden rule, do to others what you'd have them do to you.</i>
Informed Consent	Always get informed consent / stressing importance of informed consent	<i>I think at this point for any new study I started using online data, I would try to get informed consent when collecting identifiable information (e.g. usernames).</i>
Greater Good	Data collection should have a social benefit	<i>The work I do should address larger social challenges, and not just offer incremental improvements for companies to deploy.</i>
Established Guidelines	Including Belmont Report, IRBs Terms of Service, legal frameworks, community norms	<i>I generally follow the ethical guidelines for human subjects research as reflected in the Belmont Report and codified in 45.CFR.46 when collecting online data.</i>
Risks vs. Benefits	Discussion of weighing potential harms and benefits or gains	<i>I think I focus on potential harm, and all the ethical procedures I put in place work towards minimizing potential harm.</i>
Protect Participants	data aggregation, deleting PII, anonymizing / obfuscating data	<i>I aggregate unique cases into larger categories rather than removing them from the data set.</i>
Data Judgments	Efforts to not make inferences or judge participants or data	<i>Do not expose users to the outside world by inferring features that they have not personally disclosed.</i>
Transparency	Contact with participants or methods of informing participants about research	<i>I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.</i>

Item	M	SD ₃₀
...notify participants about why they're collecting online data ¹	3.89	0.96
...share research results with research subjects ¹	3.90	0.80
...Ask colleagues about their research ethics practices ¹	4.27	0.74
...Ask their IRB/internal reviews for advice about research ethics ¹	4.03	0.90
...Think about possible edge cases/outliers when designing studies ¹	4.33	0.71
...Only collect online data when the benefits outweigh the potential harms ¹	3.62	1.10
...Remove individuals from datasets upon their request ¹	4.56	0.71
Researchers should be held to a higher ethical standard than others who use online data ²	3.46	1.22
I think about ethics a lot when I'm designing a new research project ²	3.96	0.93
Full Scale ($\alpha=.71$)		4.00
		0.49

¹ Prompt: “I think researchers should....”

² Prompt: “To what extent do you agree with the following statements?”

Both sets of items were measured on five point, Likert-type scales (Strongly Agree-Strongly Disagree).

Codification of Ethical Attitudes Measure

Ethics Heuristics for Online Data Research: Beyond the Belmont Report

1. Focus on transparency

- Openness about data collection
- Sharing results with community leaders or research subjects

2. Data minimization

- Collecting only what you need to answer an RQ
- Letting individuals opt out
- Sharing data at aggregate levels

3. Increased caution in sharing results

4. Respect the norms of the contexts in which online data was generated.



A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media

Stevie Chancellor
Georgia Tech
Atlanta, GA, US
schancellor3@gatech.edu

Michael L Birnbaum
Northwell Health
Glen Oaks, NY, US
mbirnbaum@northwell.edu

Eric D. Caine
University of Rochester
Rochester, NY, US
Eric_Caine@urmc.rochester.edu

Vincent M. B. Silenzio
University of Rochester
Rochester, NY, US
vincent.silenzio@rochester.edu

Munmun De Choudhury
Georgia Tech
Atlanta, GA, US
munmund@gatech.edu

ABSTRACT

Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights. Amplifying these tensions, there are also divergent and sometimes inconsistent methodological gaps and under-explored ethics and privacy dimensions. This paper presents a taxonomy of these concerns and ethical challenges, drawing from existing literature, and poses questions to be resolved as this research gains traction. We identify three areas of tension: ethics committees and the gap of social media research; questions of validity, data, and machine learning; and implications of this

Conference on Fairness, Accountability, and Transparency (FAT '19)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3287560.3287587>

1 INTRODUCTION

Last year, Facebook unveiled automated tools to identify individuals contemplating suicide or self-injury [75, 62]. The company claims that they “use pattern recognition technology to help identify posts and live streams as likely to be expressing thoughts of suicide,” which then can deploy resources to assist the person in crisis [75]. Reactions to Facebook’s suicide prevention artificial intelligence (AI) are mixed, with some concerned about the use of AI to detect suicidal ideation as well as potential privacy violations [86]. Other suicide prevention AIs, however, have been met with stronger public backlash. Samaritan’s Radar, an app that scanned a person’s friends for concerning Twitter posts, was pulled from production, citing concerns for data collection without user permission [54], as well as enabling harassers to intervene when someone was vulnerable [4].

Since 2013, a new area of research has incorporated techniques from machine learning, natural language processing, and clinical

Overview of Taxonomy

- Participant and research oversight
- Validity, interpretability, and methods
- Stakeholder implications

Possible Ethical Solutions