Putting NLP Ethics Into Context

Ben Hutchinson

Dec 2021

Australasian Language Technology Association Workshop

Google Research

u/AdonisStarkiller • 3y

1

Ethicist

~~Scientist~~: "My findings are meaningless if taken out of context."

Ethicist

Media: ~~Scientist~~ claims "Findings are meaningless."

Google Research

Or:
There's Nothing Natural About Natural Language Processing

Ben Hutchinson

Dec 2021

Australasian Language Technology Association Workshop

Google Research

Things I won't be Talking about Today

1. Theories of Ethics

- utilitarianism
- deontological ethics
- virtue ethics
- structural ethics
- information ethics

2. NLP experiments (much)

Google

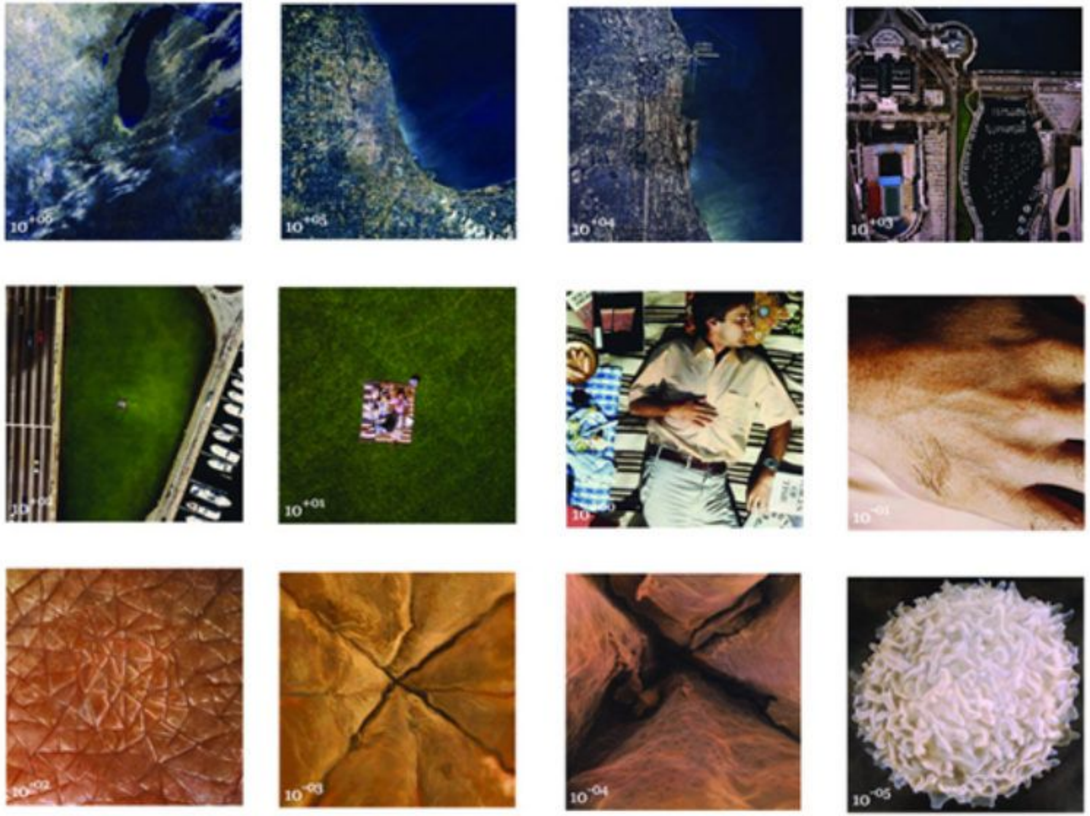
Digital Future Initiative

Google Research Australia

- Building a local team of researchers
- Fundamental and applied research
- Tackle problems that are important in Australia and globally
- Collaborate with local institutions
- research.google/careers

More things I won't be talking about today

Google Research



Charles Eames and Ray Eames. 1977. *Powers of 10.*

Google Research

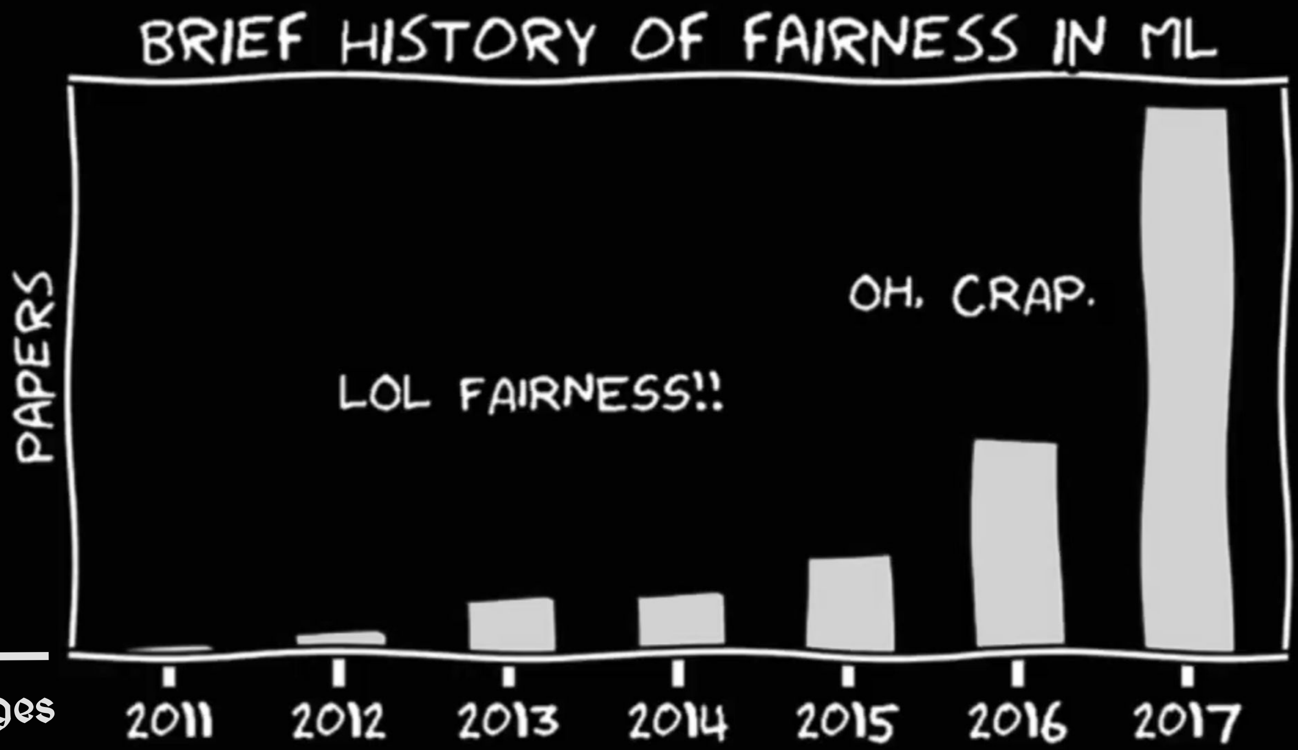
Putting NLP Ethics Into Context

I. Three Explorations of ML Ethics in Context
Historical • Social • Data

II. Seven Challenges in Responsible NLP

ML Ethics in Context #1: History of Fairness

(Hutchinson and Mitchell, 2019)

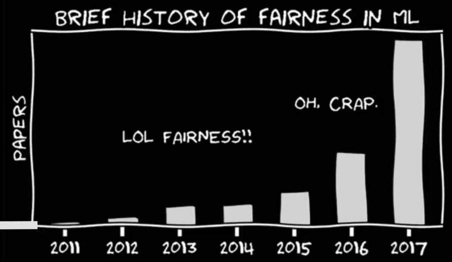


The Dark Ages of ML Fairness?

From Moritz Hardt's *CS 294: Fairness in Machine Learning* course taught at UC Berkeley.

U.S. Civil Rights
Movement

1964: Civil Rights Act



1966-1976
Golden Age
of Research into
Test Fairness

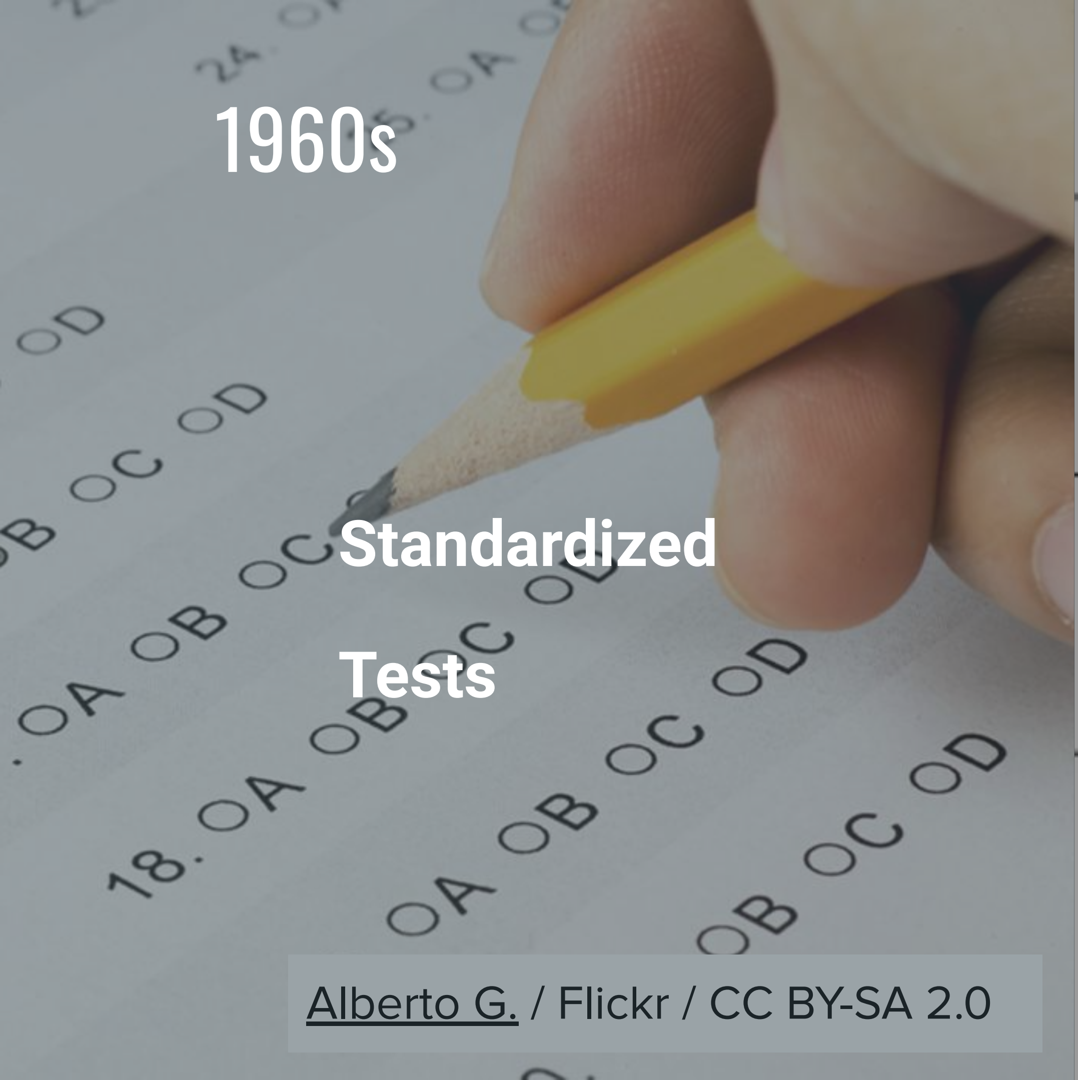
2011+
Renaissance
of Research into
ML Fairness

History may not repeat itself, but it may rhyme.

Joseph Anthony Wittreich

1960s

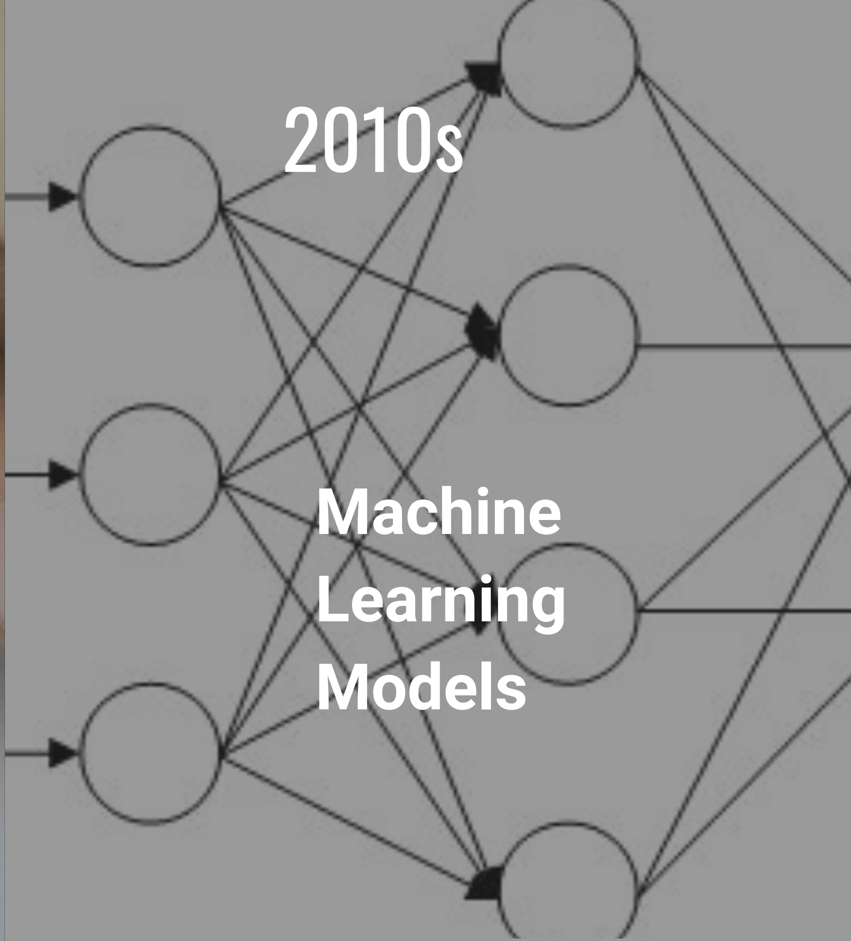
Standardized Tests



Alberto G. / Flickr / CC BY-SA 2.0

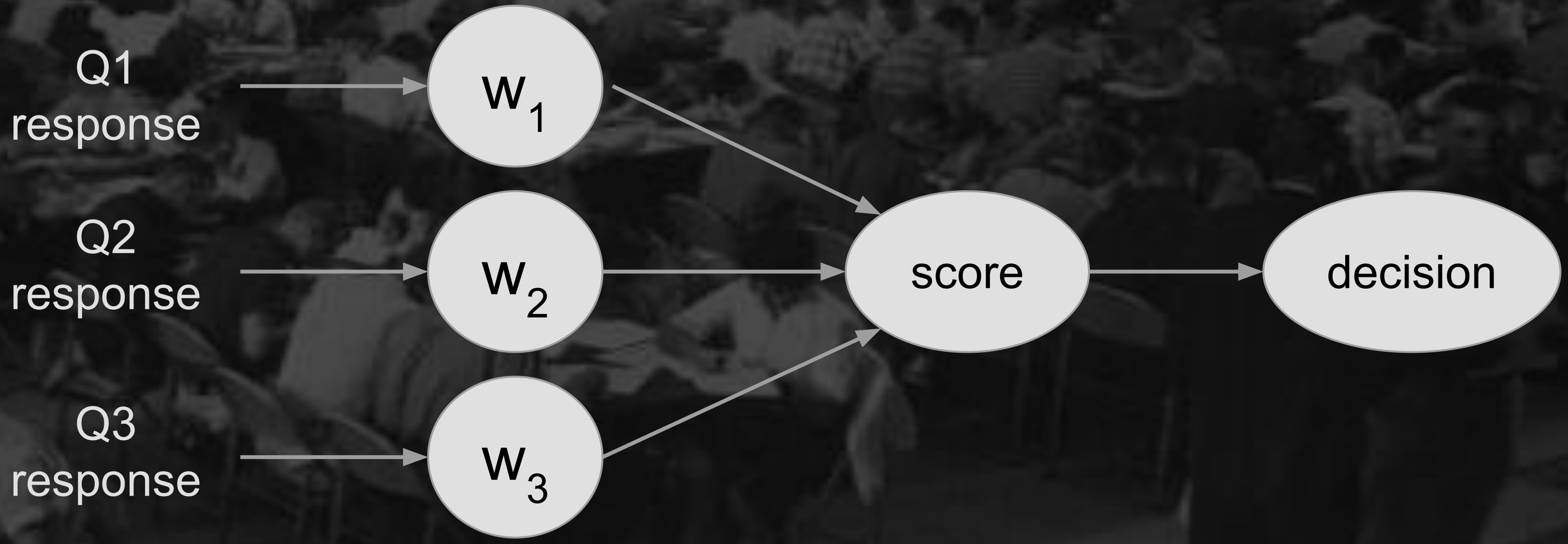
2010s

Machine Learning Models



4shadoww / Wikicommons / CC BY-SA 3.0

Tests ~ Simple Neural Networks



Fair ML in 2010s

Criminal Sentencing
Fairness Criteria

Two Competing "Fairness Criteria"

1. "sufficiency": $A \perp Y \mid D$ $\Rightarrow$

Related to:
Equal Precision (positive class) for black and white groups
Equal Precision (negative class) for black and white groups

2. "separation": $A \perp D \mid Y$ $\Rightarrow$

Related to:
Equal Recall (positive class) for black and white groups
Equal Recall (negative class) for black and white groups

A : attribute (age, gender, race, ...)
Y : target variable
D : decision using model

"Impossibility of Fairness" (Chouldechova, 2017; Kleinberg *et al.* 2016)

In general, can't have both of:

$$A \perp D \mid Y$$

$$A \perp Y \mid D$$

Exceptions:

1. D=Y ["model is perfect"], or
2. {Y|A=a} has the same distribution for all A=a. ["groups are equal"]

"Impossibility of Fairness" (Chouldechova, 2017; Kleinberg *et al.* 2016)

In general, can't have both of:

🧍 ⊥ ⭐ | 🎂
🧍 ⊥ 🎂 | ⭐

Exceptions:

1. ⭐=🎂, or
2. {🎂| 🧍 =👶} has the same distribution for all 🧍 =👶.

"Impossibility Theorem of Stars and Cakes"

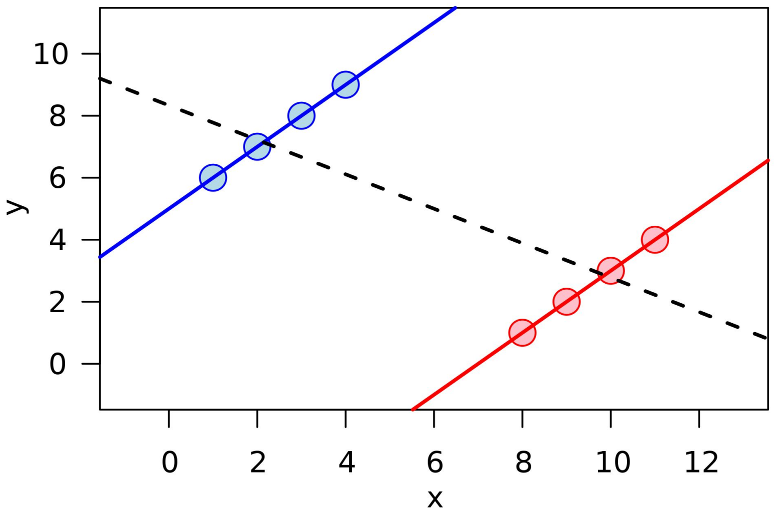
In general, can't have both of:

🧑 ⊥ ⭐ | 🎂
🧑 ⊥ 🎂 | ⭐

Exceptions:

1. ⭐=🎂, or
2. {🎂| 🧑 =👶} has the same distribution for all 🧑 =👶.

Simpson's "Paradox"



Arrow's Theorem

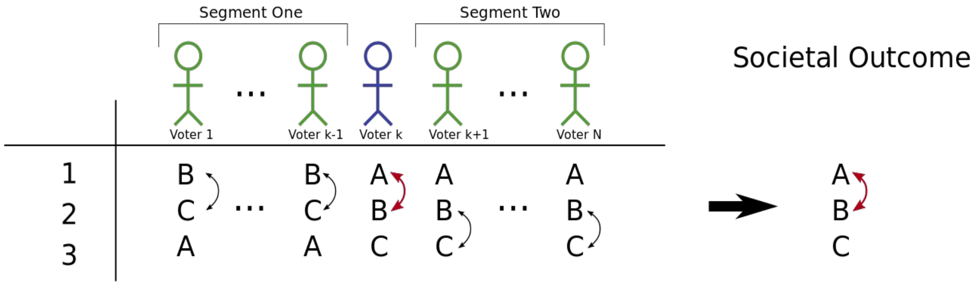
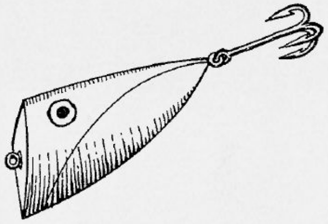


Image sources: Wikipedia

Test Fairness in 1950s-1970s

School Desegregation
Bans on Discrimination
Court Cases on Test Bias
Psychometric Research
Calls for Moratoriums on Tests

Surprise Quiz Time!



355. The above is usually called a
    A. fly.
    B. spoon.
    C. spinner.
    D. plug.
    E. streamer.

1950

1954 Brown vs Board of Education

1957 Desegregation of Little Rock High School

1960

1964 Chester School Protests

1970

1950

1960

1960s >25% of students take SAT

1970

1976 33% of students take SAT

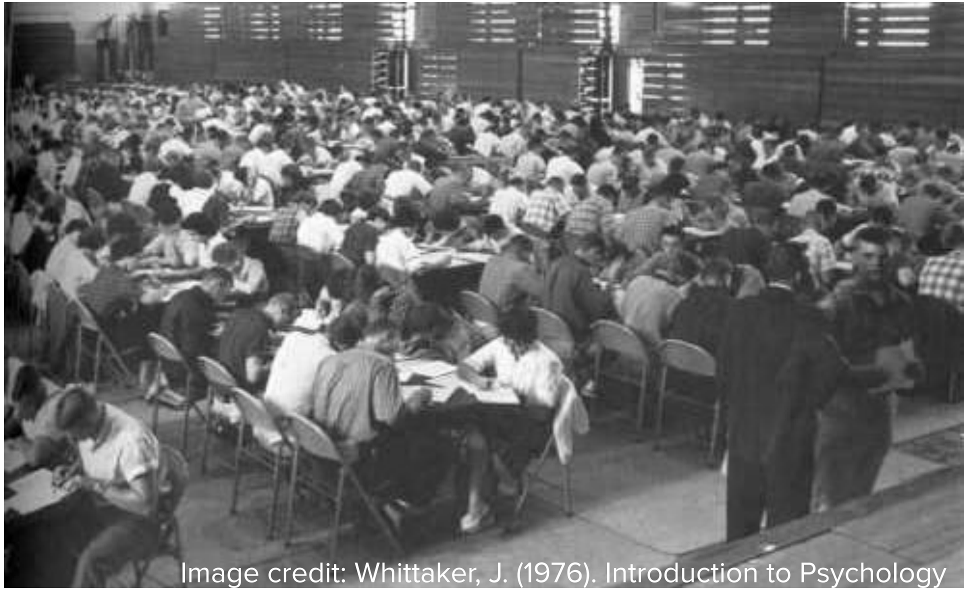


Image credit: Whittaker, J. (1976). Introduction to Psychology

1950

1960

1963 90% of companies surveyed use selection tests

1964 Motorola's General Ability Test Number 10 found to be discriminatory

1966 Finding is overruled by Illinois Supreme Court (*Myart vs Motorola*)

1970

Experts Criticize Motorola Exam

Two psychologists asserted yesterday that the job employment test given by Motorola, Inc., discriminates against culturally deprived persons and is too brief to measure a job seeker's ability reliably.

The psychologists are Prof. Benjamin S. Bloom, University of Chicago specialist in testing, and Robert L. French, vice president for research and testing of Science Research Associates.

Testify Before FEPC

They testified at a hearing by the Illinois Fair Employment Practices commission on the appeal by Motorola against an FEPC examiner's order last March 5 that the test was unfair to Leon Myart, 28, of 6333 Dorchester av.

The examiner, Robert E. Bryant, directed Motorola to offer Myart a job as a phaser and analyzer. Motorola appealed to the commission, and at a previous hearing several professors testified for the company that they did not think the test discriminated against anyone.

Prof. Bloom said that it would be "quite possible" for a person to be highly successful in a job even tho he might get a low score on the Motorola general ability test.

Holds Test Unreliable

"A test such as this is quite likely to underrate individuals who are culturally deprived," Bloom said. "This is one of the shortest aptitude tests in the field, and is likely to be unreliable."

He said the test would discriminate against not only Negroes but "others who might

MOTOROLA JOB TEST IS FAIR, HEARING TOLD

Can't Design a Biased Exam, Professors Say

1950

1960

1964 Civil Rights Act

1970

U.S. Civil Rights Act of 1964

Title VI--NONDISCRIMINATION IN FEDERALLY ASSISTED PROGRAMS

> "No person in the United States shall, on the ground of **race, color, or national origin** … be subjected to discrimination under any program or activity receiving Federal financial assistance."

Title VII--EQUAL EMPLOYMENT OPPORTUNITY

> "It shall be the policy of the United States to insure equal employment opportunities for Federal employees without discrimination because of **race, color, religion, sex or national origin**"

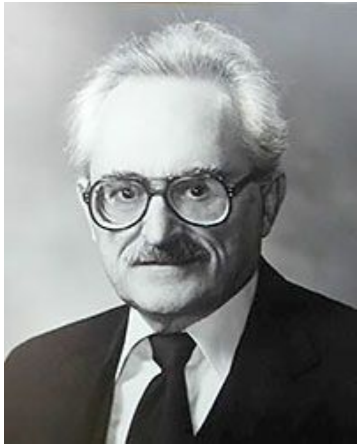
1950

1960

1964 *Culture-fair Testing* (Anne Anastasi)

1966 *The Implications of the Civil Rights Act of 1964 for Psychological Assessment in Industry* (Philip Ash)

1970

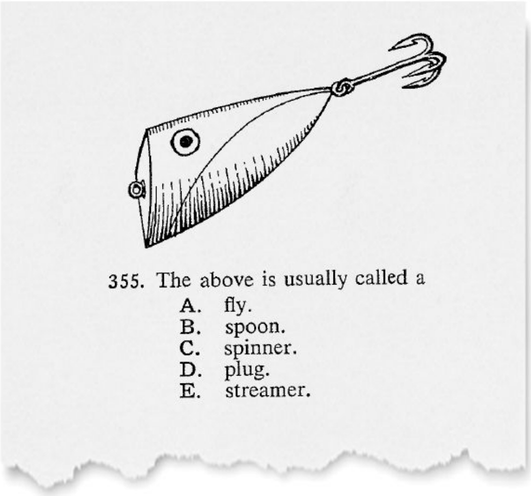


1950

1960

1964 Project Talent test administered to 440,000 students in grades 9-12

1970



355. The above is usually called a
A. fly.
B. spoon.
C. spinner.
D. plug.
E. streamer.

1950

1960

1969 Association of Black Psychologists calls for a moratorium on standardized tests

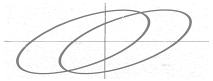
1970

1972 National Education Association calls for a moratorium

1974 NAACP calls for a moratorium

1980 "The War Against Testing: A Current Status Report"

1960s + 1970s: Bias & Fairness



Fair Test Scores
Fair Predictions
Fair Selection Decisions
Fair Representation

Fairness in Testing in the 1960s and 1970s

had remarkable similarities to

ML Fairness in the 2010s

1950
1960
1970

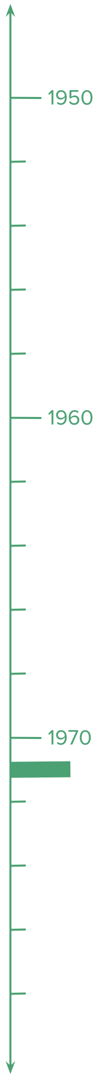
1971 Richard Darlington: Fairness as Correlation

A : race **R** : LSAT score
Y : GPA

Can unfair racial biases in Law School SAT (LSAT) be detected by considering various correlations?

$$A \perp Y \mid R \Rightarrow \rho_{AY.R} = 0$$
$$A \perp R \mid Y \Rightarrow \rho_{AR.Y} = 0$$

Assuming [!]: goal of LSAT is to predict college grades (GPA)



1971 Richard Darlington

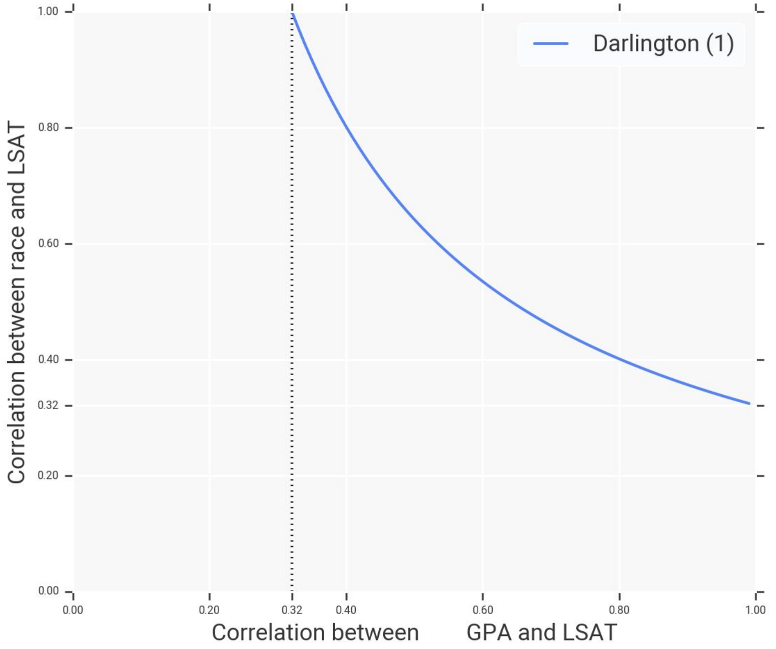
A : race
Y : GPA
R : LSAT score

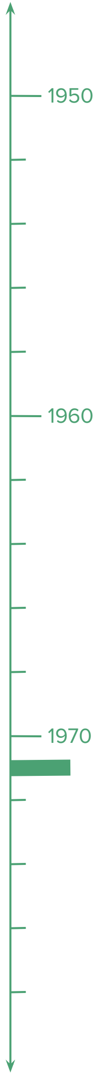
What is a fair relationship between race & test score?

1. $\rho_{AY.R} = 0$

Entailed by A⊥Y | R ("sufficiency") when A, Y, R are multivariate normal.

"Fair" values of cultural discrimination, according to different definitions of fairness, for ρ(race, gpa) = 0.321





1971 Richard Darlington

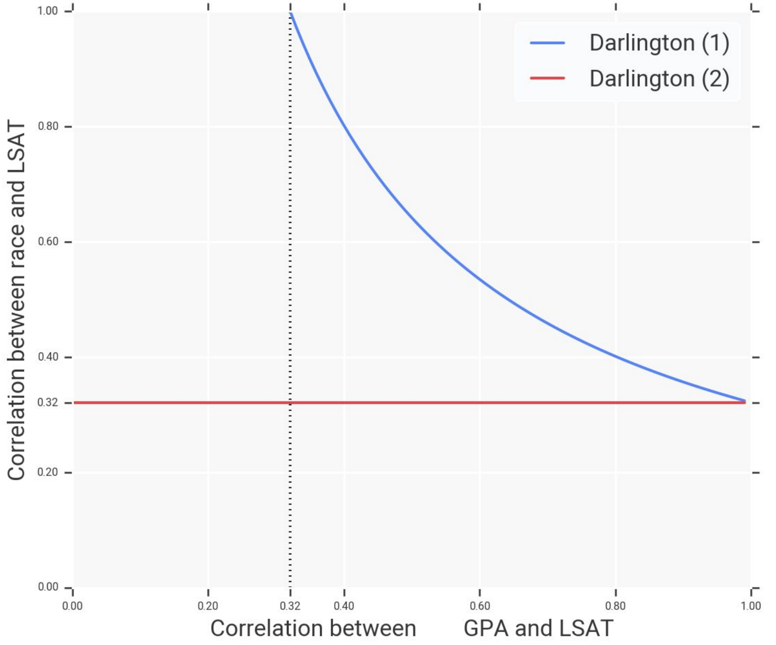
What is a fair relationship between race & test score?

2. $\rho_{AR} = \rho_{AY}$

Aim to select an equal proportion of people from each group as are qualified within that group.

A : race
R : LSAT score
Y : GPA

"Fair" values of cultural discrimination, according to different definitions of fairness, for ρ(race, gpa) = 0.321



1950

1960

1970

1971 Richard Darlington

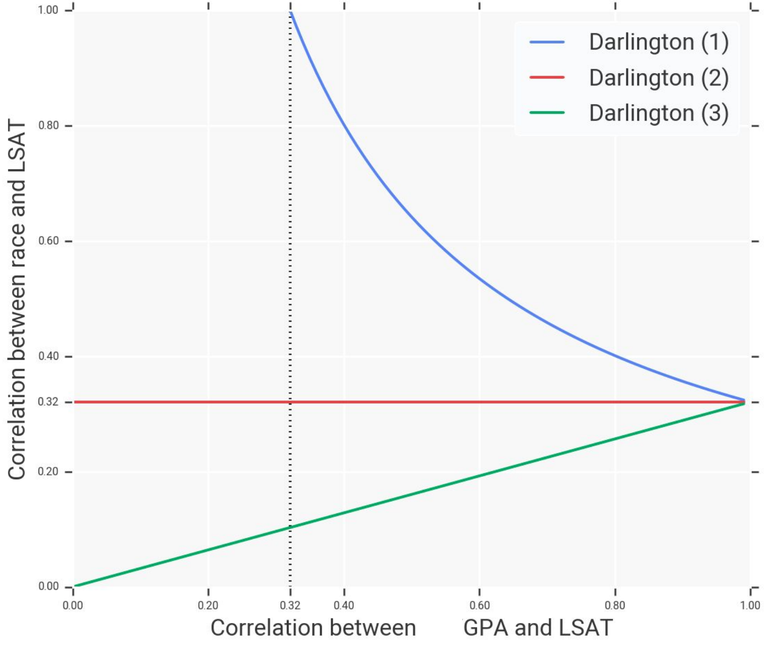
What is a fair relationship between race & test score?

3. $\rho_{AR.Y} = 0$

Entailed by $A \perp R \mid Y$ ("separation") when A, Y, R are multivariate normal.

A : race
R : LSAT score
Y : GPA

“Fair” values of cultural discrimination, according to different definitions of fairness, for ρ(race, gpa) = 0.321



1950

1960

1970

1971 Richard Darlington

A : race **R** : LSAT score

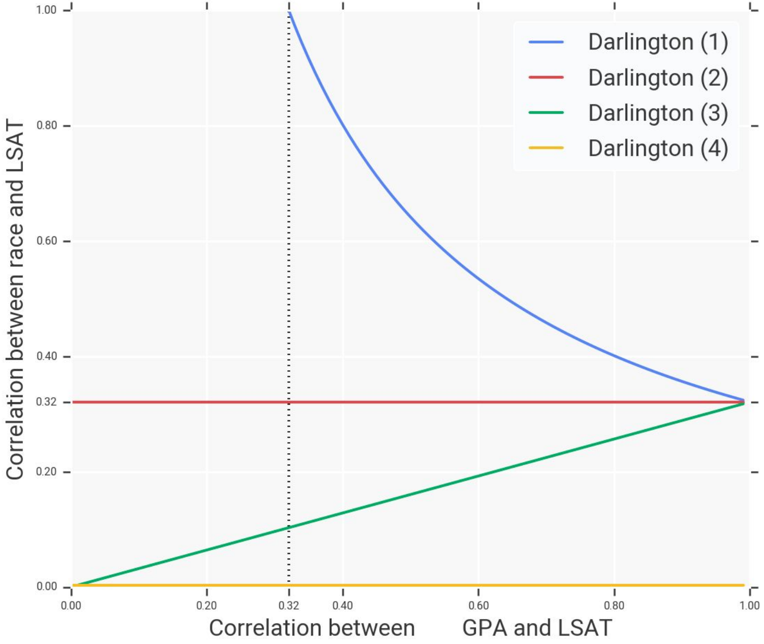
Y : GPA

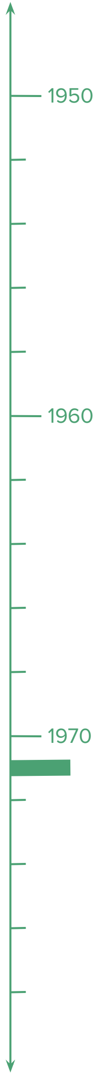
What is a fair relationship between race & test score?

4. $\rho_{AR} = 0$

Relaxation of $A \perp R$.

“Fair” values of cultural discrimination, according to different definitions of fairness, for ρ(race, gpa) = 0.321





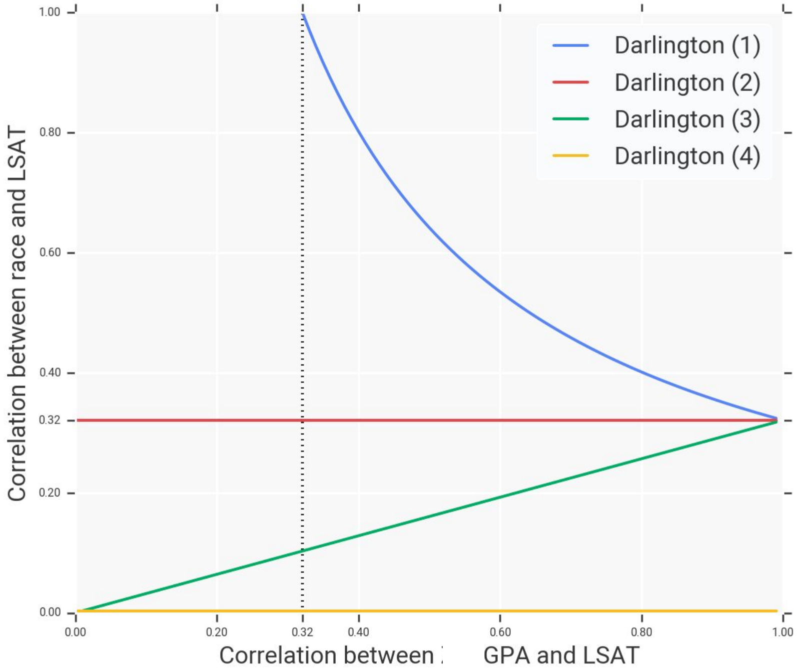
1971 Richard Darlington

Four definitions are **incompatible** unless one of

1. $\rho_{RY} = 1$ [i.e. "test is perfect"]
2. $\rho_{RY} = 0$ [i.e. "test is useless"]
3. $\rho_{AY} = 0$ [i.e. "groups are equal"]

A : race
Y : GPA
R : LSAT score

"Fair" values of cultural discrimination, according to different definitions of fairness, for ρ(race, gpa) = 0.321

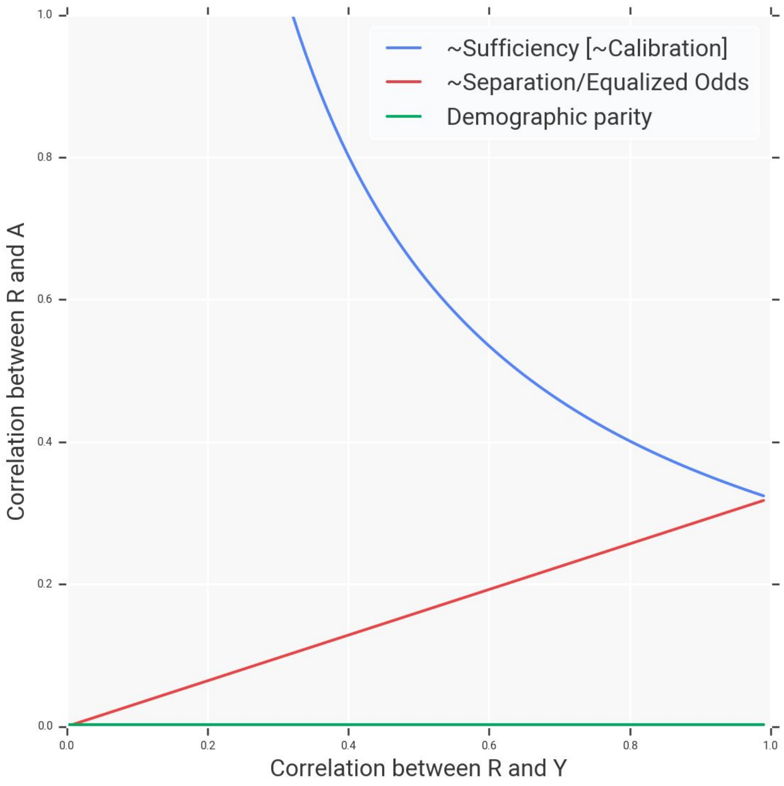


1971 Richard Darlington: Takeaways and Lessons

1950
1960
1970

In some cases, fairness criteria exists on a **spectrum**

The level of practical disagreement between fairness definitions **depends on the model accuracy**



Test Fairness	ML Fairness	Relationship Between Test & ML Fairness
Cleary (1966)	sufficiency $A \perp Y \mid R$	closely related when R and Y have bivariate Gaussian distribution
Guion (1966)	individual	relaxation
Thorndike (1971)	accurate coverage $\frac{P(D=1)}{P(Y=1)=1}$	generalization
Darlington (1971)	(1) sufficiency $A \perp Y \mid R$ (2) - (3) separation $A \perp R \mid Y$ (4) demographic parity $A \perp R$	equiv. when multivariate Gaussian distribution - equiv when multivariate Gaussian distribution equiv when bivariate Gaussian distribution
Cole (1973)	equality of opportunity $A \perp D \mid Y=1$	equivalent
Linn (1973)	predictive parity $A \perp D \mid Y=1$	equivalent
Jones (1973)	constrained fair ranking	special case
Petersen and Novick (1976)	(1) separation $A \perp Y \mid R$ (2) sufficiency $A \perp R \mid Y$	equivalent equivalent

History has not repeated itself, but it *has* rhymed.

Thanks to: Richard Darlington

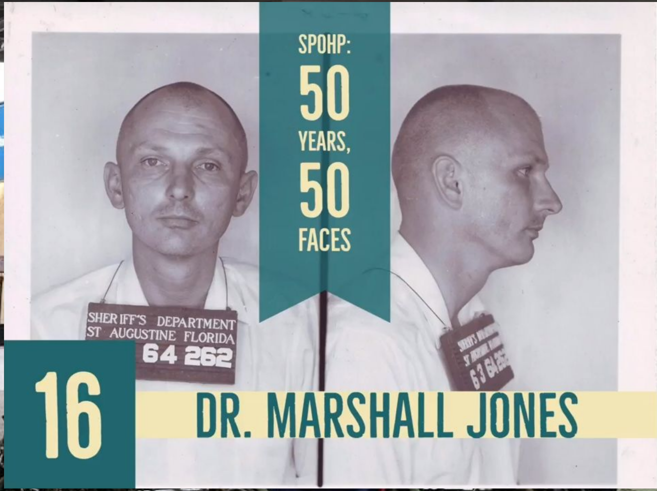
for providing historical context

Another look at "cultural fairness" (1971)

Is culture-fairness objective or subjective? (1973)

Thanks to: Marshall Jones

"The most dangerous man in American academic life"

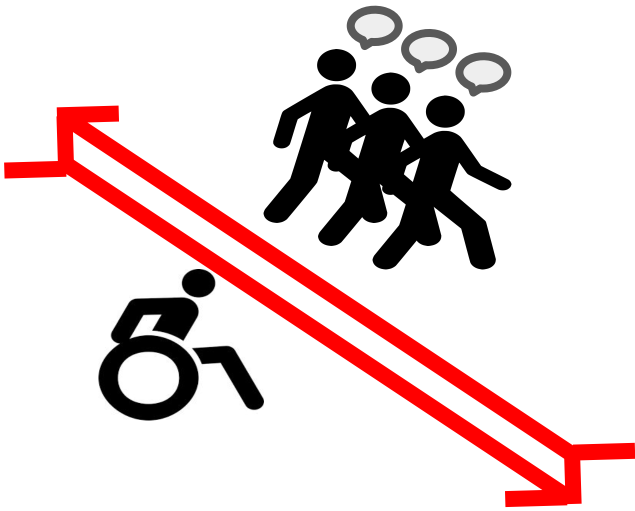


Moderated regression and equal opportunity (1973)

The role of the faculty in student rebellion (1966)

ML Ethics in Context #2: Societal Impacts of Biases

(Hutchinson, Prabhakaran, Denton, Webster, Zhong and Denuyl, 2020)

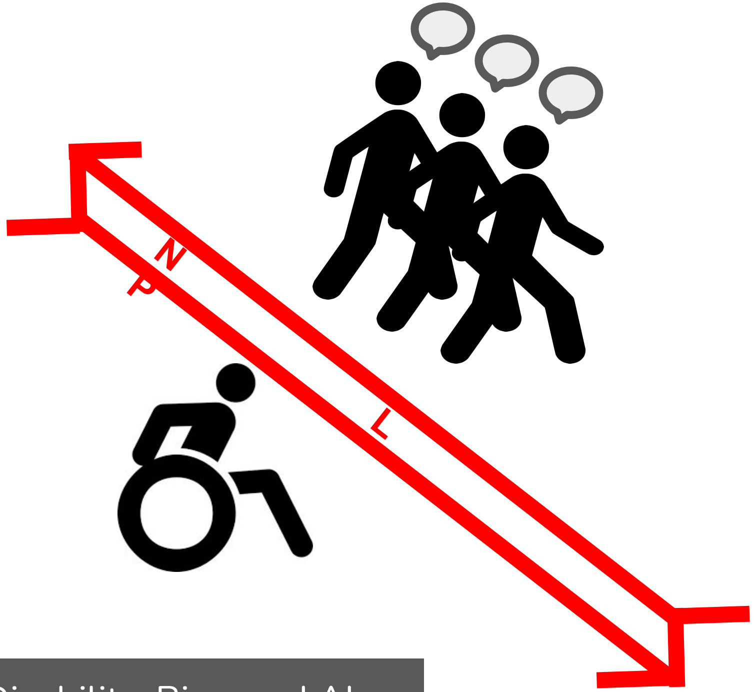


Toxicity (Perspective API)

Input	Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

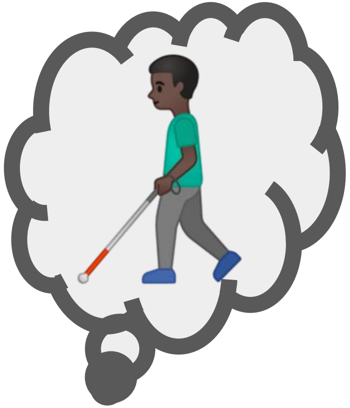
Staircase as Physical Barrier or Handicap

Model Bias as Barrier to Opportunity

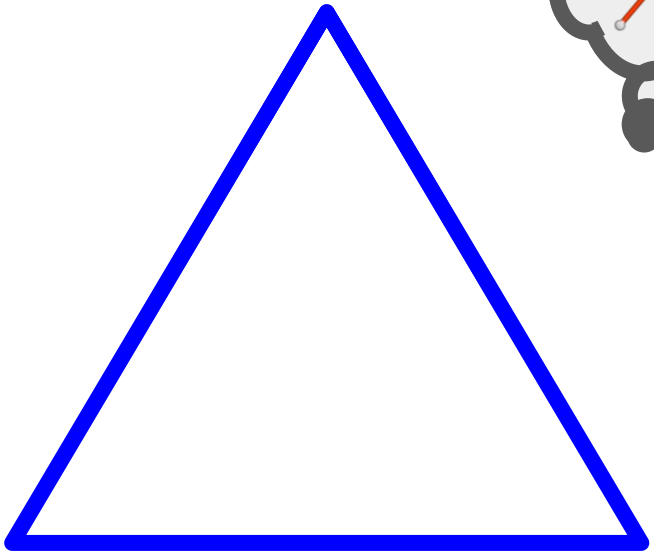


See also: [Whittaker et al., 2019. Disability, Bias, and AI].

Concept



Referent



Language

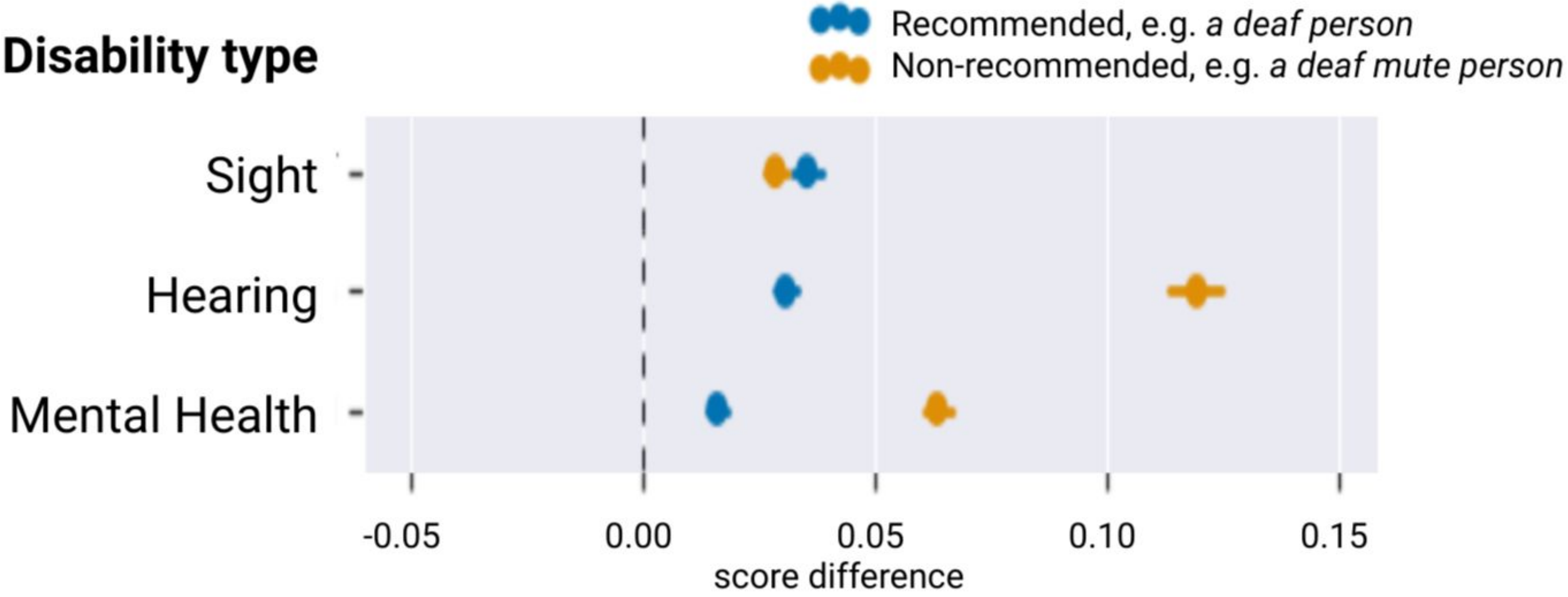
a blind person (✔)

a person who is blind (✔)

a sight deficient person (✘)

Writing Guide's of SIGACCESS, ADA National Network, Anti-Defamation League

Perturbation Sensitivity: Some Results



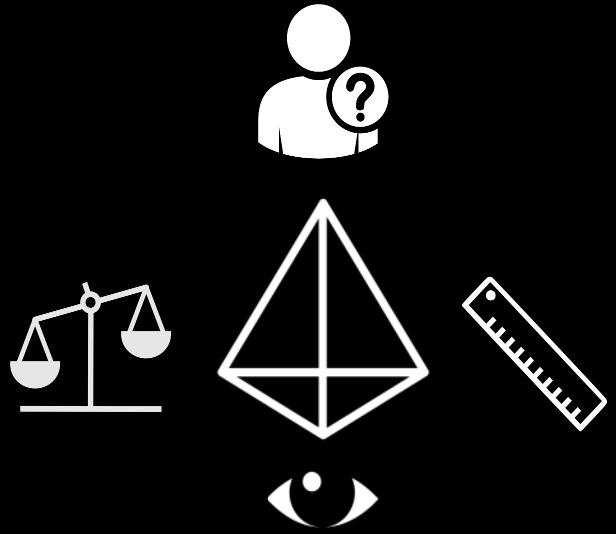
Potential Implications: Abusive Language Detection

Disproportionate censorship of authors writing about disability

Delays awaiting approvals by "Humans in the Loop"

Disrespect of authors' language choices

Perpetuate invisibility of disability



ML Ethics in Context #3: ML Dataset Construction

(Hutchinson, Smart, Hanna, Denton, Greer, Kjartansson, Barnes and Mitchell, 2021)

ML Data as *Data*

ML's primary focus is on explaining differences in learning algorithms.

Common ML practices reinforce the notion of data as **decontextualised** fixed resources—*data* in the original meaning of the word!—for the competition of learning algorithms.

"Data": The data

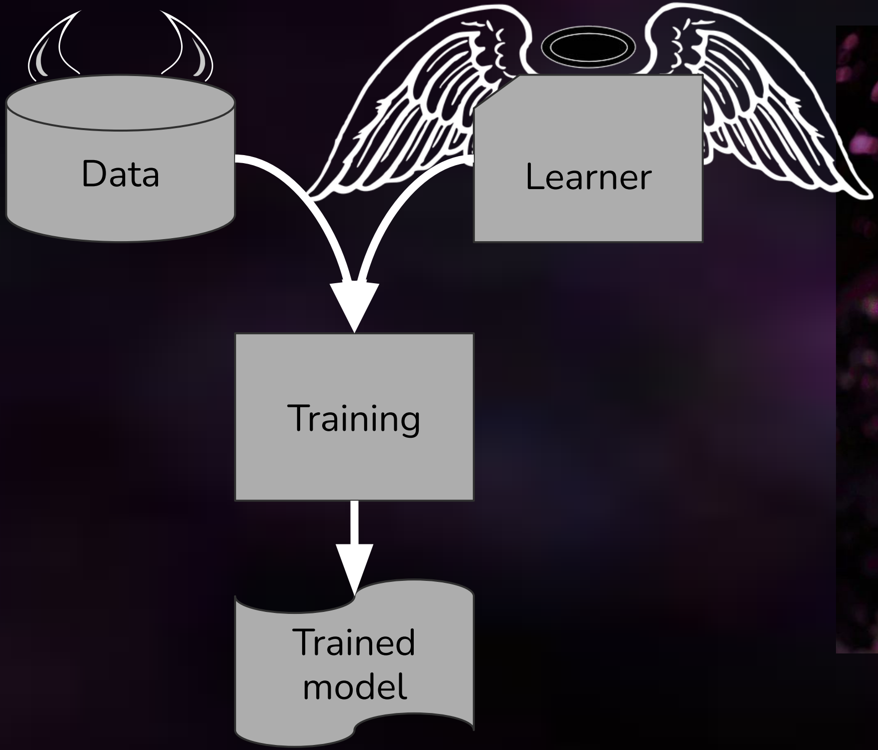
Jonathan Furner

Abstract

While many scholars in information science have understandably focused on the concept of "information" as foundational, some authors have identified other concepts as having similarly foundational status. Two that are regularly suggested as candidates are "data" and "document." Oddly, perhaps, for such a basic term, "data" has not been as frequently subject to probing analysis in the scholarly literature as "information"; and although "document" has long been a term of special interest to historians of the European documentation movement, some of whom continue to develop a document theory, there is little consensus on the precise nature of the conceptual relationship between "data" and "document." In this paper, a review is conducted of historical interpretations of "data," and relationships with contemporary conceptions of "document" are explored. The conclusion is reached that, current practice notwithstanding, it is not in fact the case that documents are made up of data, nor that the document is a species of dataset: rather it is the other way round, in both respects. A dataset is made up of documents; and the dataset is a species of document.

> "For a science like information science (IS), it is of course important how fundamental terms are defined."
> (Capurro & Hjørland, 2003, p. 344)

The Trope of Good Learners and Bad Data



Data Scapegoating in Fairness Discourse

"The <u>ML model</u> is biased because the <u>data</u> is biased."

"The <u>data</u> is biased because the <u>ML model</u> is biased."

Data Scapegoating

*"Computer systems frequently mediate the interactions between machines and humans... **human actions are distanced from their causal impacts**... at the same time, that the computer's action is a more direct causal antecedent."*

Nissenbaum. 1996. Accountability in a Computerized Society.

Image: Witches Sabbath, by Goya. Open Domain. Source.

Data Distances ML Impacts from ML Data Work

- Dataset development work is distanced from its causal impacts.
- The data itself is seen a more direct causal antecedent.
- But datasets are artefacts, and cannot be held accountable.

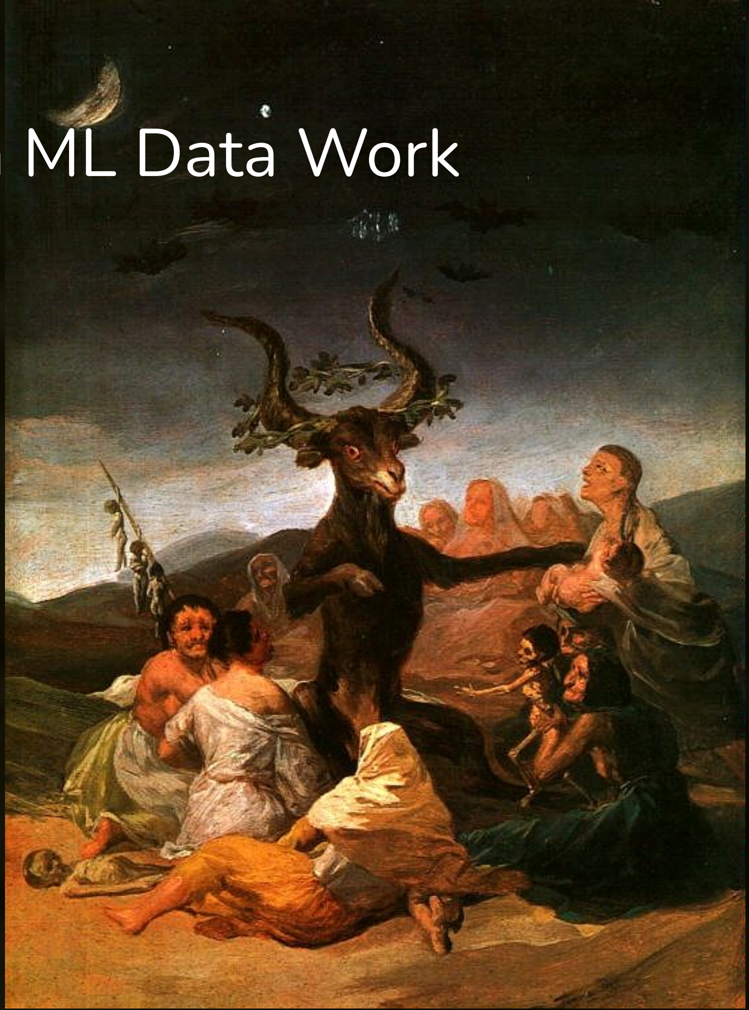


Image: Witches Sabbath, by Goya. Open Domain. Source.

Data Workers have Lower Status

*"that [i.e. data] work is done by workers with **lower status in the workplace**."*

Møller. 2020 Who does the work of data?

*"the **lionized work** of building novel models"*
Sambisavan, *et al.* 2020. Ibid.

*"AI **superstars**"*

*"deep learning **savant**"*
Ari. 2018. The rise and rise of AI in Africa.

Image: The Fall of the Rebel Angels, by Pieter Bruegel the Elder. CC. Source.

Is NLP Data Work Lower Status?

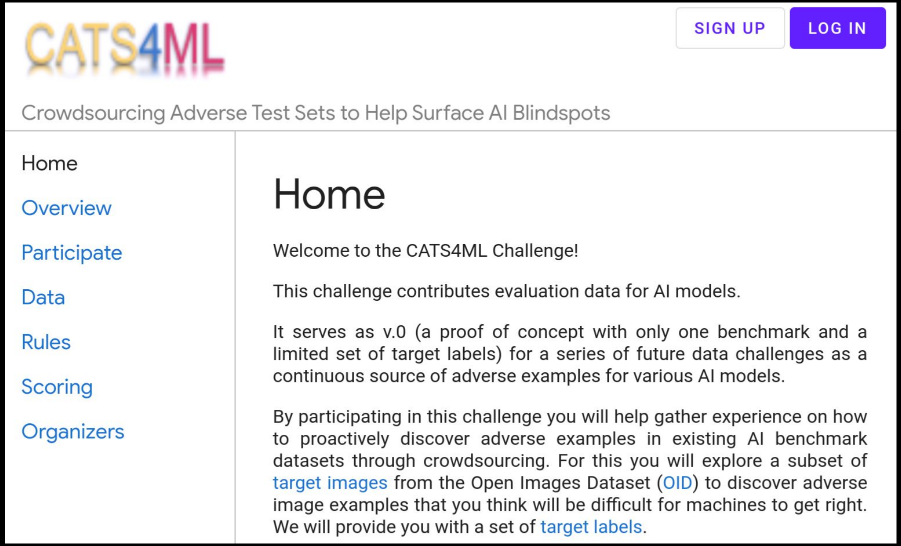
LREC

EMNLP

Image: The Fall of the Rebel Angels, by Pieter Bruegel the Elder. CC. [Source](#).

Recognize Value of AI Dataset Expertise

- Echo calls by Jo and Gebru (2020) for work on the theory and practice of AI Dataset Development
- More recognition of skilled data work, including conferences and prizes



Fair Pay Analogs of Fair ML

(Peng, Naecker, Hutchinson, Smart & Noorosi, 2020)

Pay Fairness Criterion #1

$$\text{Group} \perp \text{Pay} \mid \widehat{\text{Work}}$$

- implies if two groups do the same work they should be paid the same
- violated if two groups do the same work but one is paid more

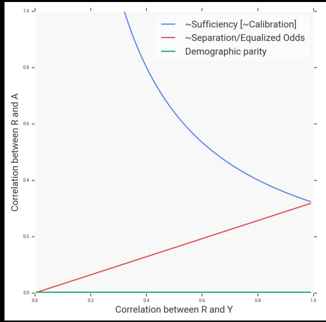
Pay Fairness Criterion #2

$$\text{Group} \perp \widehat{\text{Work}} \mid \text{Pay}$$

- implies if two groups are paid the same they should have done the same work
- violated if two groups are paid the same but one does more work

Impossibility of Fair Pay

(Peng, Naecker, Hutchinson, Smart & Noorosi, 2020)



In general, can't have both of:

🧍 ⊥ ⭐ | 🎂

🧍 ⊥ 🎂 | ⭐

In general, can't have both of:

$$\text{Group} \perp \text{Pay} \mid \text{Work}$$

$$\text{Group} \perp \text{Work} \mid \text{Pay}$$

Exceptions:

1. $\widehat{\text{Work}} = \text{Work}$
2. All groups have the same distribution of Pay

Dataset Development is Political

Requires acknowledging:

- impacts
    - what is enabled?
    - what is encouraged?
- roles, stakes and expertise of others

Data Science as Political Action:
Grounding Data Science in a Politics of Justice

Ben Green
bgreen@g.harvard.edu
Berkman Klein Center for Internet & Society at Harvard University
Harvard John A. Paulson School of Engineering and Applied Sciences

Critique and Contribute:
A Practice-Based Framework for Improving Critical Data Studies and Data Science

Gina Neff,[1,*] Anissa Tanweer,[2] Brittany Fiore-Gartland,[3] and Laura Osburn[4]

Recap of Part I

History often rhymes

Social perspectives matter

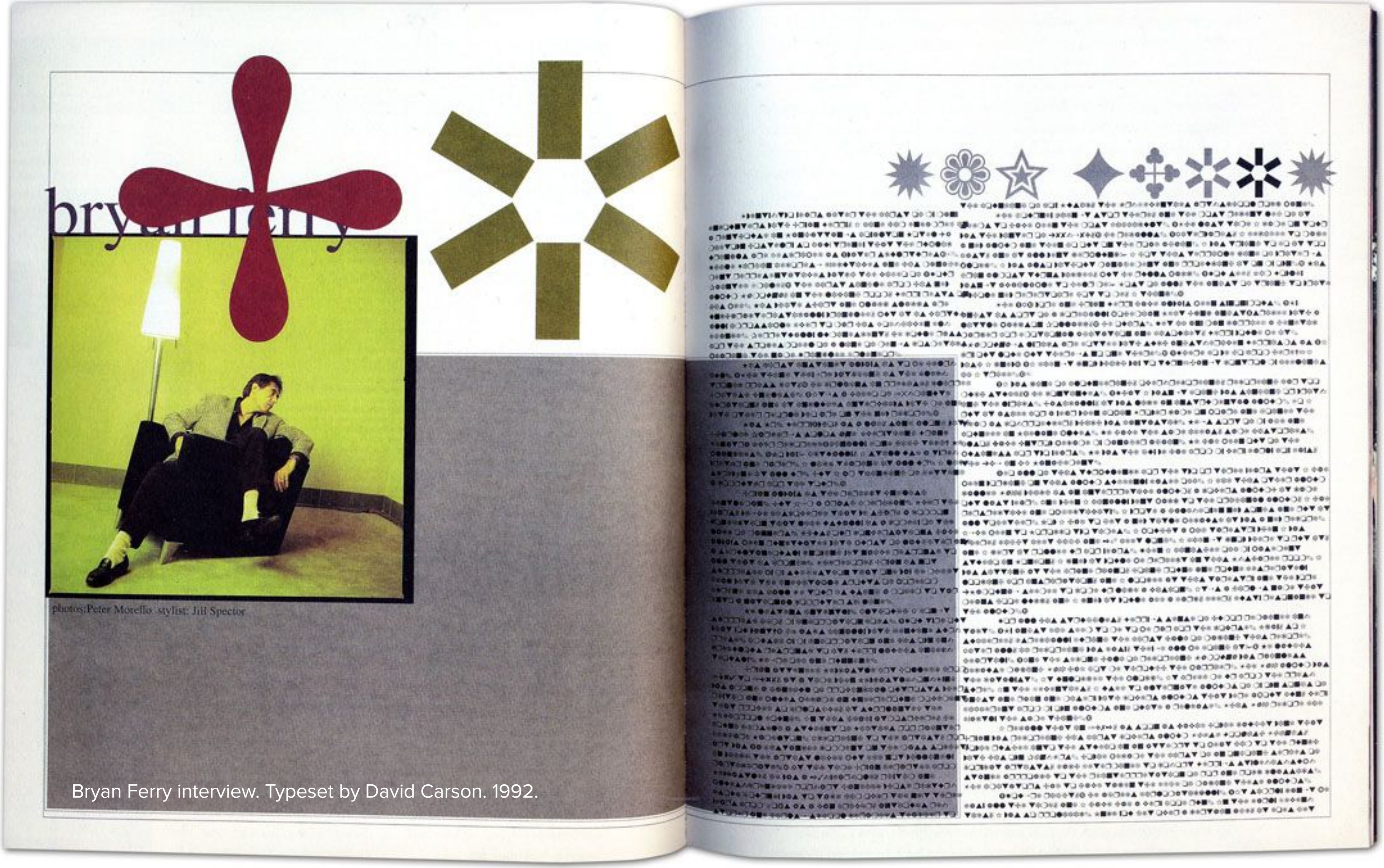
Dataset development is political

Part II:
Seven Challenges in Responsible NLP

Google Research

There's nothing natural about natural language.

Rulifson



Bryan Ferry interview. Typeset by David Carson. 1992.

Revolutionary Technologies of So-called "Natural" Language

1. speech: sound↔meaning
2. writing: grapheme↔sound/meaning
3. printing: standardisation
4. ascii/unicode: grapheme↔codepoint

+font

+weight/size/...

+rendering engine

+...

NLP often starts here

Google Research

natural

vs

artificial

Language is a social construct
[Hovy keynote abstract]

natural

cultural/historical/contextual

artificial

There's nothing natural about natural language processing.

Challenge #1: Linguistic Subjectivity

Embrace disagreement and ambiguity!
[Plank keynote]

What is the relationship between subjectivity and disagreement?

What is "truth" when trustworthy subjects disagree?

What is the relationship between continuous language variation and disagreement on language tasks?

Dealing with Disagreements:
Looking Beyond the Majority Vote in Subjective Annotations

Aida Mostafazadeh Davani
University of Southern California
mostafaz@usc.edu

Mark Díaz
Google Research
markdiaz@google.com

Vinodkumar Prabhakaran
Google Research
vinodkpg@google.com

We Need to Consider Disagreement in Evaluation

Valerio Basile[*♣], Michael Fell[♣], Tommaso Fornaciari[♦], Dirk Hovy[♦], Silviu Paun[♥], Barbara Plank[♠], Massimo Poesio[♥], Alexandra Uma[♥]
[♣]University of Turin, [♦]Bocconi University
[♥]Queen Mary University of London, [♠]IT University of Copenhagen
[♣]{valerio.basile, michaelkurt.fell}@unito.it
[♦]{dirk.hovy, fornaciari.tommaso}@unibocconi.it
[♥]{s.paun, m.poesio, a.n.uma}@qmul.ac.uk, [♠]bplank@itu.dk

Subjective Natural Language Problems:
Motivations, Applications, Characterizations, and Implications

Cecilia Ovesdotter Alm
Department of English
College of Liberal Arts
Rochester Institute of Technology
coagla@rit.edu

Truth Is a Lie:
Crowd Truth and the
Seven Myths of
Human Annotation

Lora Aroyo, Chris Welty

Google Research

Challenge #2: Cultural and Societal Pluralism

Social norms and values differ across both languages and cultures.

Technologies encode cultural values, e.g., on violent or pornographic language, concepts of fairness,

How do we avoid dominant cultures imposing their norms via NLP technologies?

Re-imagining Algorithmic Fairness in India and Beyond

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, Vinodkumar Prabhakaran
(nithyasamba,erinarnesen,benhutch,tulsee,vinodkpg)@google.com
Google Research
Mountain View, CA

ABSTRACT

Conventional algorithmic fairness is West-centric, as seen in its sub-

of AI fairness failures and stakeholder coordination have resulted in bans and moratoria in the US. Several factors led to this outcome:

Decolonising Speech and Language Technology

Steven Bird
Northern Institute
Charles Darwin University

Challenge #3: NLP Infrastructures and Re-use

- Language Datasets
- Foundation Models
- Model Adaptation
- System Adaptation

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

source: Google Streetview

Google Research

Challenge #4: NLP Systems Rearrange Power

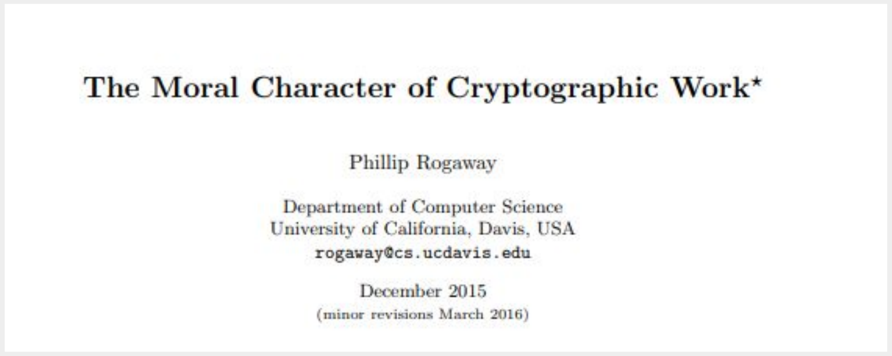
What actions do NLP systems enable or encourage?

What actions do NLP systems inhibit or discourage?

Who benefits most and least?


LANGDON WINNER

Do Artifacts Have Politics?


The Moral Character of Cryptographic Work*

Phillip Rogaway

Department of Computer Science
University of California, Davis, USA
rogaway@cs.ucdavis.edu

December 2015
(minor revisions March 2016)

Challenge #5: Representation and Representativeness

Model the variety space
[Plank keynote]

Which language (-variety) communities are represented in our NLP datasets?

Who decides how NLP technology is built and for what purposes?

How do we measure fair representation in both cases?

Bringing the People Back In: Contesting Benchmark Machine Learning Datasets

Emily Denton [*1] Alex Hanna [*1] Razvan Amironesei [2] Andrew Smart [1] Hilary Nicole [1]
Morgan Klaus Scheuerman [1]

Representativeness in Statistics, Politics, and Machine Learning

Kyla Chasalow
Cornell University
kec89@cornell.edu

Karen Levy
Cornell University
karen.levy@cornell.edu

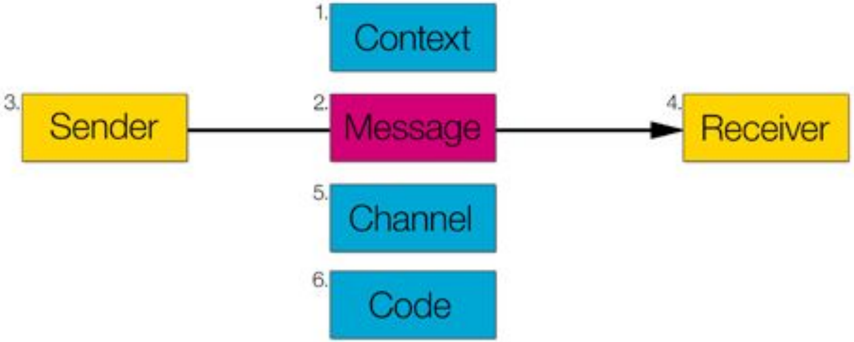
Representativeness in Corpus Design

DOUGLAS BIBER
Department of English, Northern Arizona University

Challenge #6: Language & Its Technologies are Contextual

Fairness and Abstraction in Sociotechnical Systems

ANDREW D. SELBST, Data & Society Research Institute
DANAH BOYD, Microsoft Research and
Data & Society Research Institute
SORELLE A. FRIEDLER, Haverford College, PA
SURESH VENKATASUBRAMANIAN, University of Utah
JANET VERTESI, Princeton University



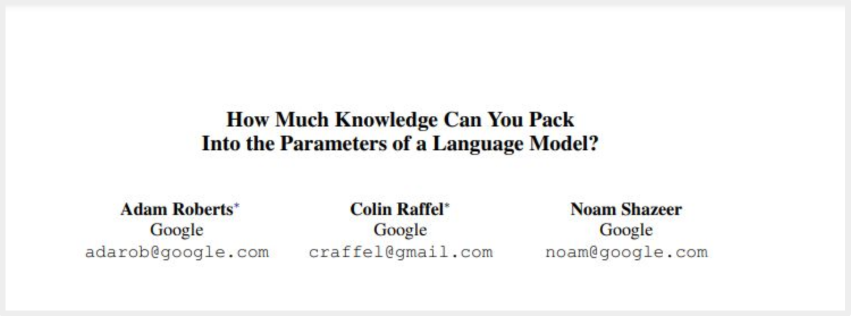
Roman Jakobson's Model of Communication
(image source: wikipedia)

Humans in the loop
[Plank keynote]

Challenge #7: Epistemologies of NLP

What forms of "knowledge" can LMs have?

- Linguistic?
- Encyclopedic/world?
- Commonsense?
- Moral?

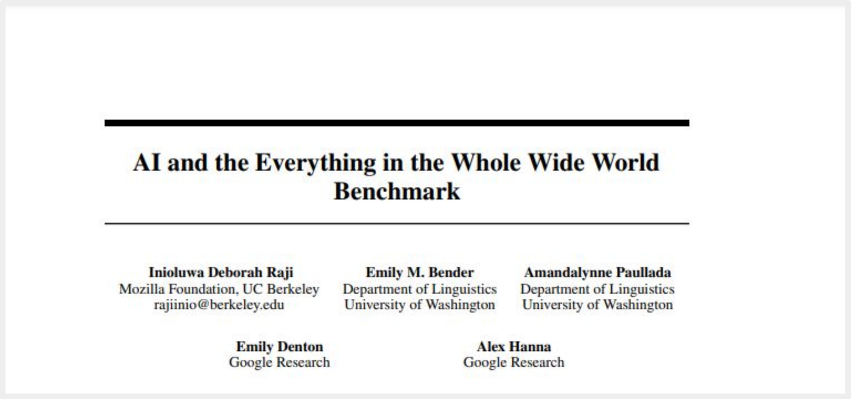


How Much Knowledge Can You Pack Into the Parameters of a Language Model?

Adam Roberts* Google adarob@google.com

Colin Raffel* Google craffel@gmail.com

Noam Shazeer Google noam@google.com



AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji Mozilla Foundation, UC Berkeley rajiinio@berkeley.edu

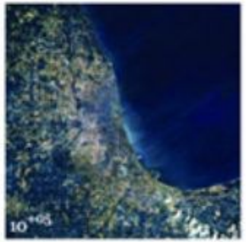
Emily M. Bender Department of Linguistics University of Washington

Amandalynne Paullada Department of Linguistics University of Washington

Emily Denton Google Research

Alex Hanna Google Research

NLP Ethics



Micro

Technical

Fine-Tuning

Macro

Societal

Resonances

Charles Eames and Ray Eames. 1977. *Powers of 10.*

Google Research

Thank you!