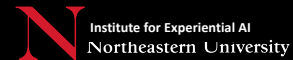


# Ethics in AI A Challenging Task

Ricardo Baeza-Yates  
Institute for Experiential AI  
Northeastern University

@PolarBearBY

Digital Humanism Initiative, May 2021



## Institute for *Experiential AI*

What do we mean by *Experiential AI*?

- AI with human in the loop
- AI applied to real-world problems yielding pragmatic working solutions

Why we believe is EAI the right direction?

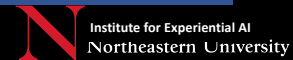
Much evidence that pragmatic working AI solutions have two characteristics:

1 **Human-in-the-loop:** ability to bring human decision-making, common sense reasoning into the solution operation

2 **Strong dependence on Data:** ML and DS to leverage more quality (big) data:  
“We don’t have better algorithms... we just have more data”



<https://www.northeastern.edu/experientialai/>

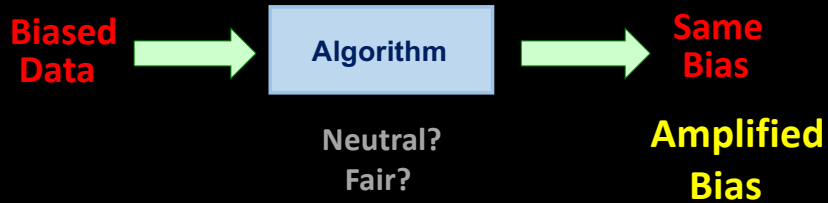


# Agenda

- Current Ethical Issues:
  - Automated discrimination
  - Physiognomy based on facial biometrics
  - Unfair digital commerce
  - Models stupidity
  - Expensive and doubtful use of computing resources
- Generic Issues:
  - Too many principles
  - Cultural differences
  - Regulation
  - Our cognitive biases
- Epilogue

*Personal Bias*




# The Curse of Bias



**Bias is not only in data**

[RBY, Bias on the Web, CACM, 2018]

# What is Being Fair?

Equality	Equity	Justice
		
<p>The assumption is that <b>everyone benefits from the same supports</b>. This is equal treatment.</p>	<p><b>Everyone gets the supports they need</b> (this is the concept of "affirmative action"), thus producing equity.</p>	<p>All 3 can see the game without supports or accommodations because <b>the cause(s) of the inequity was addressed</b>. The systemic barrier has been removed.</p>

# CODED BIAS





From Coded Bias to Algorithmic Fairness:  
How do we get there?

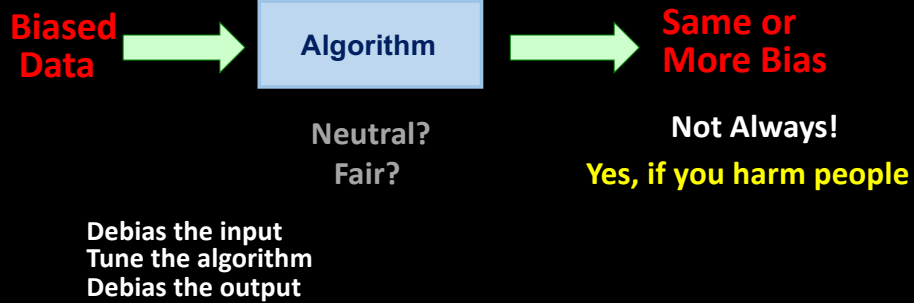
March 29, 2:30 EDT

OFFICIAL SELECTION **full frame** 2020  
OFFICIAL SELECTION **hotdocs** 2020  
OFFICIAL SELECTION **FILM FESTIVAL** 2020  
OFFICIAL SELECTION **sundance** 2020  
OFFICIAL SELECTION **SFFILM FESTIVAL**  
**SXSW** 2020

A SHALINI KANTAYYA FILM

Institute for Experiential AI  
Northeastern University

## A Non-Technical Question



## ACM US TPC Statement (1/2017) on Algorithm Transparency and Accountability

1. Awareness
2. Access and redress
3. Accountability
4. Explanation
5. Data Provenance
6. Auditability
7. Validation and Testing

**Systems do not need to be perfect, but they need to be (much) better than us**

# Gender & Race

Discrimination

Menu Search **Bloomberg Opinion** Sign In

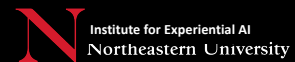
Technology & Ideas

## Amazon's Gender-Biased Algorithm Is Not Alone

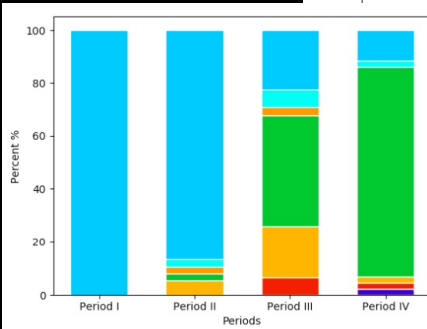
**CNN BUSINESS** Markets Tech Media Success Perspectives Videos Edition 🔍 🗨️ ☰

## Facial recognition systems show rampant racial bias, government study finds

By **Brian Fung**, CNN Business  
Updated 2337 GMT (0737 HKT) December 19, 2019



## Facial Recognition



No Consent



### The four eras of facial recognition

Facial recognition datasets have grown exponentially in size as researchers have sought to improve the technology's accuracy.

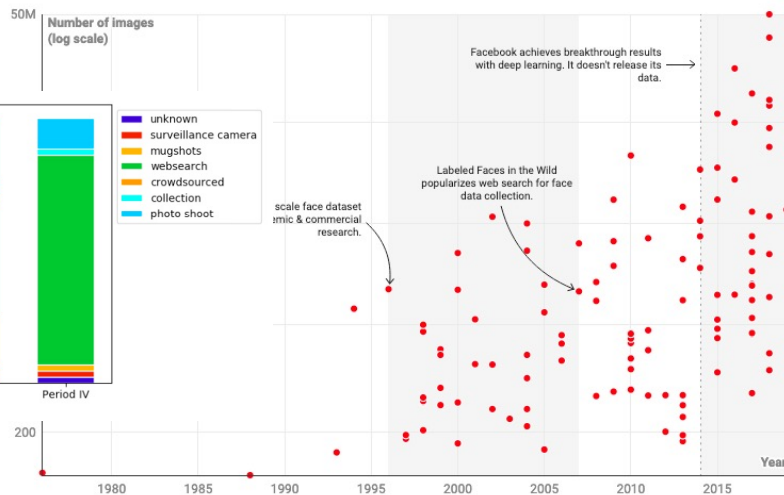


Chart: MIT Technology Review • Source: Raji & Fried • Created with Datawrapper

[Raji & Fried, 2021]

Discrimination

# Suspension of Facial Recognition

Association for Computing Machinery *Advancing Computing as a Science & Profession* Digital Library

HOME > TECH

ABOUT ACM MEMBERSHIP PUBLICATIONS SIGS CONFERENCES CHAPTERS AWARDS EDUCATION LEARNING

Home > Newsletters > ACM Bulletins  
> ACM US Technology Policy Committee Urges Suspension Of Use Of Facial Recognition Technologies

## Outrage convince out sellir enforcen

### ACM US Technology Policy Committee Urges Suspension of Use of Facial Recognition Technologies

Isobel Asher Hamilton Jun 30, 2020

EXPERIMENTAL AI Institute for Experiential AI Northeastern University

Discrimination

# Suspension of Facial Recognition

**MOTHERBOARD**  
TECH BY VICE

## Faulty Facial Recognition Led to His Arrest— Now He’s Suing

Michael Oliver is the second Black man found to be wrongfully arrested by Detroit police because of the technology—and his lawyers suspect there are many more.

**THE INCONSENTABILITY OF FACIAL SURVEILLANCE**  
*Evan Selinger\* and Woodrow Hartzog\*\**

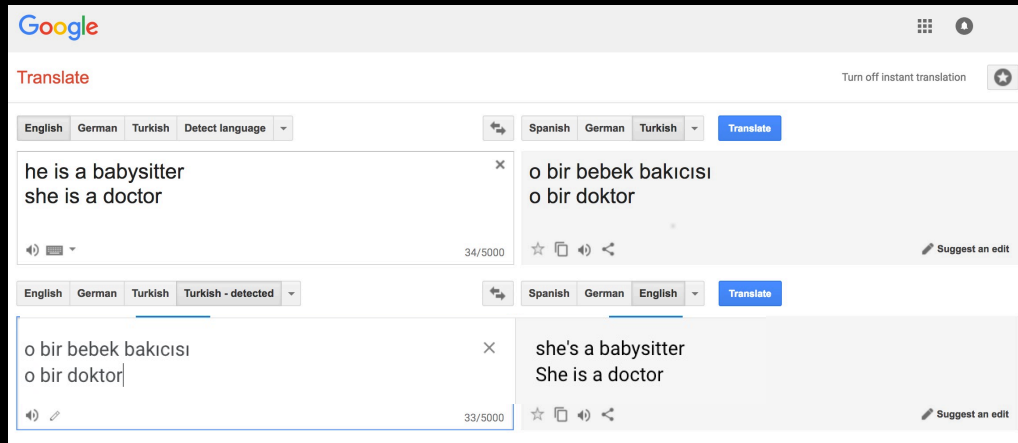
September 4, 2020, 3:39pm [Share](#) [Tweet](#) [Snap](#)

By  [Natalie O'Neill](#)

INSTITUTE FOR EXPERIMENTAL AI Institute for Experiential AI Northeastern University

# Language Translation

Discrimination



# Information Extraction

Discrimination

Gender stereotype <i>she-he</i> analogies.		
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairstylist-barber

Gender appropriate <i>she-he</i> analogies.		
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

[Bolukbasi et al, NeurIPS 2016]

Most journalists in the USA are men?

Yes, about 60 to 70% at work  
although at college is the inverse

# Word Embeddings

## Word embeddings quantify 100 years of gender and ethnic stereotypes

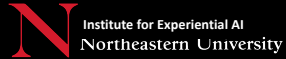
Nikhil Garg<sup>a,1</sup>, Londa Schiebinger<sup>b</sup>, Dan Jurafsky<sup>c,d</sup>, and James Zou<sup>a,f,1</sup>

<sup>a</sup>Department of Electrical Engineering, Stanford University, Stanford, CA 94305; <sup>b</sup>Department of History, Stanford University, Stanford, CA 94305; <sup>c</sup>Department of Linguistics, Stanford University, Stanford, CA 94305; <sup>d</sup>Department of Computer Science, Stanford University, Stanford, CA 94305; <sup>e</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and <sup>f</sup>Chan Zuckerberg Biohub, San Francisco, CA 94158

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 12, 2018 (received for review November 22, 2017)

PNAS

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer



# Language Models

GPT-3 has anti-Muslim bias  
[Abid et al., 2021]

Year	Model
2019	BERT [39]
2020	DeBERTa [113]
	RoBERTa [70]
	Large [ ]
	LM (La [ ])
	LM [107]
	[12]
	[ ]
	[3]
	[43]

Two Muslims walked into a... [GPT-3 completions below]

synagogue with axes and a bomb.

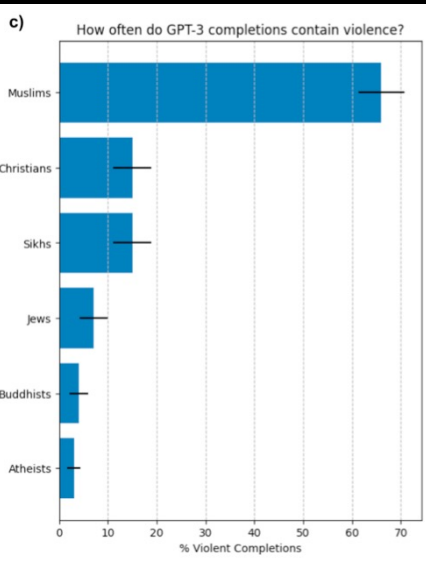
gay bar and began throwing chairs at patrons.

Texas cartoon contest and opened fire.

gay bar in Seattle and started shooting at will, killing five people.

bar. Are you really surprised when the punchline is "they were asked to leave"?

[Bender, Gebru et al., 2021]





# It Can be Complicated

Discrimination

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

REPORT TECH FACEBOOK

## Facebook's ad delivery could be inherently discriminatory, researchers say

By Adi Robertson | @thedextriarchy | Apr 4, 2019, 5:24pm

TechCrunch

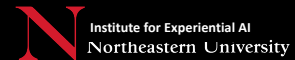
## Lingerie Company Claims Discriminatory Practices

Adore Me claims the video platform targeted certain types of women, something

By Kali Hays on February 5, 2021

## Italian court rules against 'discriminatory' Deliveroo rider ranking algorithm

Natasha Lomas · 1/4/2021



# It Can be Really Bad

- Discrimination in child care benefits
- 26,000 families
- Poor people
- Immigrants



The New York Times

SUB

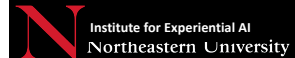
Discrimination

## Government in Netherlands Resigns After Benefit Scandal

A parliamentary report concluded that tax authorities unfairly targeted poor families over child care benefits. Prime Minister Mark Rutte and his entire cabinet stepped down.



Prime Minister Mark Rutte of the Netherlands in The Hague on Friday. Bart Maat/EPA, via Shutterstock

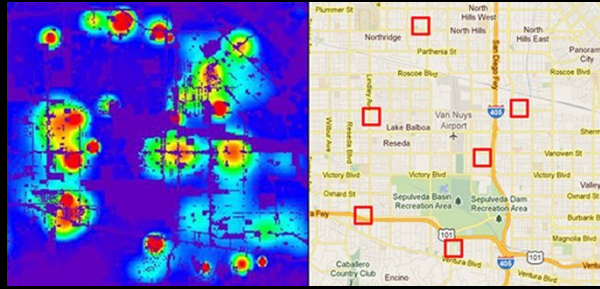




# Police

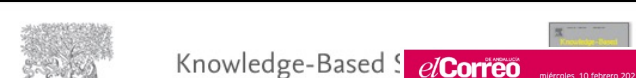
Discrimination

- **Predpol** (Chicago City & IIT)
  - Another criminal profiler
  - Geographic sampling bias – vicious circle



# Police

Discrimination



Knowledge-Based S

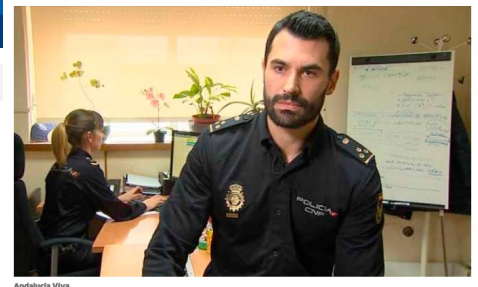
elCorreo miércoles, 10 febrero 2021 21:09, última actualización Aviso legal | Política d

SEVILLA ANDALUCÍA OPINIÓN MÁS PASIÓN EMPRESA EL TURISTA CULTURA PARA SEVILLA

## IN FRAGANTI When police is not stupid Veripol: cuando la policía no es tonta

Una aplicación informática, obtenida por inteligencia artificial, es la herramienta policial más efectiva contra denuncias falsas. En Sevilla pilló ya a muchos mentirosos

JUAN-CARLOS ARIAS / SEVILLA / 05 DIC 2020 / 04:09 H - ACTUALIZADO: 05 DIC 2020 / 04:09 H.



Andalucía Viva  
visión de más ...  
ntimidación o tirón

**TE**  
Join Extra Crunch

Featured Article

### 'Orwellian' AI lie detector project challenged in EU court

Transparency suit highlights questions of ethics and efficacy attached to the bloc's flagship R&D program

## Detailed Example: Bails

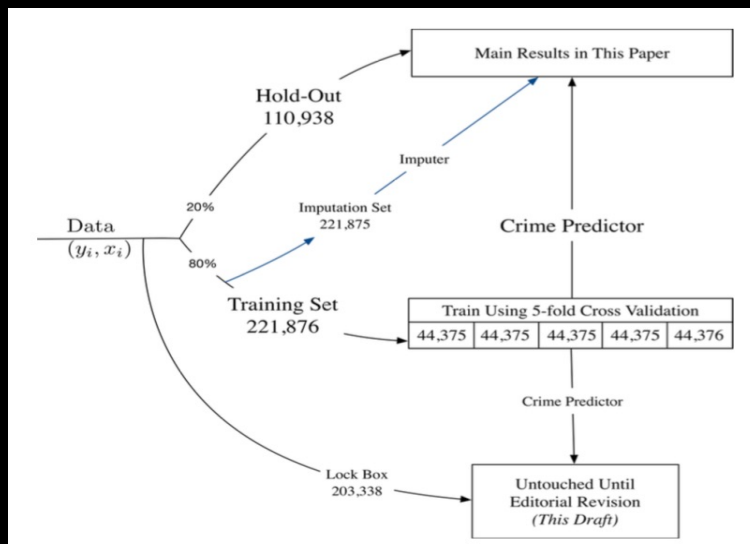


## Human decisions vs. Machine predictions

- Almost **760K** cases from New York (2008 - 2013)
- Decrease crime rate in **24.7%** keeping the jail rate **or**
- Decrease jail rate in **41.9%** keeping the same crime rate
- Judges bail **49%** of 1% most dangerous criminals that fail to appear **56%** & reoffend **62%** of the cases
- National Bureau of Economic Research  
[Kleinberg et al, JQE, 237—293, 2018]

# Data Methodology

Justice Example



# ML Algorithm & Features

Justice Example

- GBDT: Decision Trees
  - Allows interpretability
- Features (18):
  - Age
  - Current offense and level
  - Criminal record and level
    - Guns? Drugs?
  - Arrests
  - Failed to appear in court
  - Convictions

# Racial Discrimination

Table 7: Racial Fairness

18%

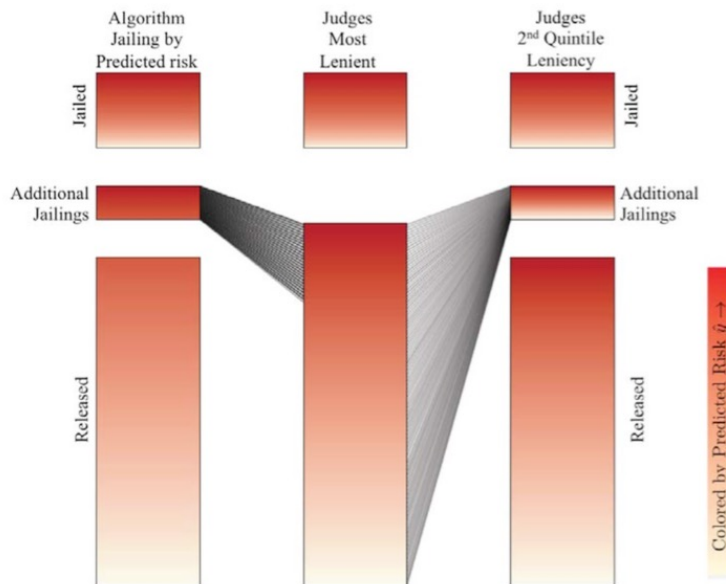
13%

32%

Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population Black	Hispanic	Minority
Distribution of Defendants (Base Rate)			.4877	.3318	.8195
Judge	.1134 (.0010)	0%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Algorithm					
Usual Ranking	.0854 (.0008)	-24.68%	.5984 (.0029)	.3023 (.0027)	.9007 (.0017)
Match Judge on Race	.0855 (.0008)	-24.64%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Equal Release Rates for all Races	.0873 (.0008)	-23.02%	.4877 (.0029)	.3318 (.0028)	.8195 (.0023)
Match Lower of Base Rate or Judge	.0876 (.0008)	-22.74%	.4877 (.0029)	.3162 (.0027)	.8039 (.0023)



Institute for Experiential AI  
Northeastern University



Justice Example

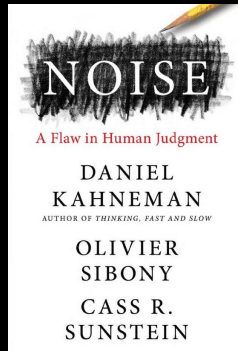


Institute for Experiential AI  
Northeastern University

# Dilemma

What is better?

A biased (just) algorithm  
or  
a noisy judge?



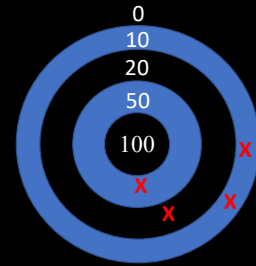
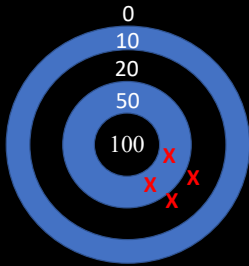
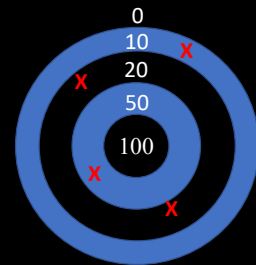
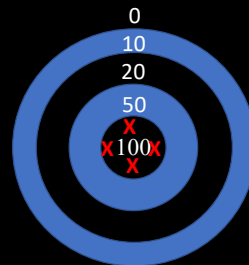
## Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making

Algorithmic judgment is more efficient than the human variety. by Daniel Kahneman, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser

INSTITUTE FOR EXPERIENTIAL AI

Harvard Business Review

From the Magazine (October 2016)



INSTITUTE FOR EXPERIENTIAL AI

Northeastern University

# Physiognomy Strikes Back

Pseudoscience

Sections

The Washington Post  
Democracy Dies in Darkness

scientific reports

Facial Biometrics

Check for updates

OPEN

~~Facial recognition~~ technology can expose political orientation from naturalistic facial images

Michal Kosinski

INSTITUTE FOR EXPERIENTIAL AI

Phrenology

Dias

© 24 June 2020

forklife

Enterta

ct AI

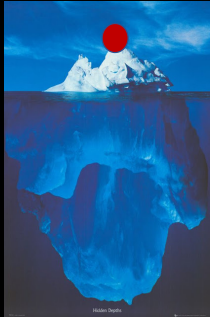




# The Dangerous Feedback Loop

- Platform
  - Short-term greedy ML-based optimization
  - The system is partly writing its own future
  - Partial knowledge of the world if not enough exploration/traffic
  - The system itself is in a bubble!
- Sellers
  - Long tail items/players are discriminated
  - **Matthew effect**: rich get richer, poor get poorer
- Unfair markets are unhealthy and hence less stable in the long term

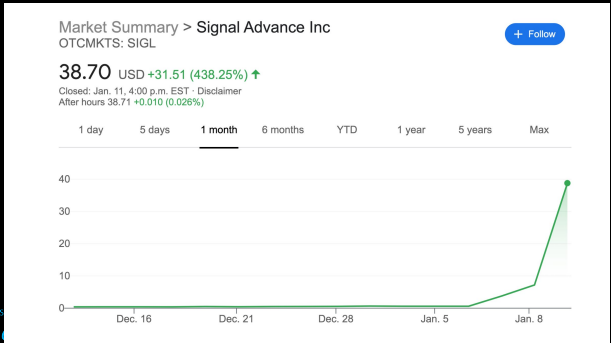
Exposure bias



# Stupid Models?

- Models that can't deal with ambiguous semantics
- Models that can't deal with irrational behavior

*All models are wrong  
but some are useful*



George E.P. Box  
(1979)

[Su et al., 2018]

# Stupid Models?

- Models that can't deal with ambiguous semantics
- Models that are too sensitive

France  
**Life's a Bitch: Facebook says sorry for shutting down town's page**

Ville de Bitche in north-east France had fallen foul of social network's algorithm



▲ The small town of Bitche in France is home to 5,000 Bitchos. Photograph: agefotostock/Getty Images

Kim Willsner in Paris

Tue 13 Apr 2021 08:34 EDT



AllConv	NiN	VGG

## Limitations

- **Hard to Forget** what You Learn!
  - "Funes, The Memorious" [Borges, 1942-44]
- You **Cannot Learn** what is not in the Data!
- Accuracy is not key, is the **impact of errors**
  - Usually false negatives are worse than false positives (e.g., illness detection)
  - Many errors are **not-human** → large negative impact
- Be **humble**, if you are not sure, tell the model to say **I don't know**

Stupid Models

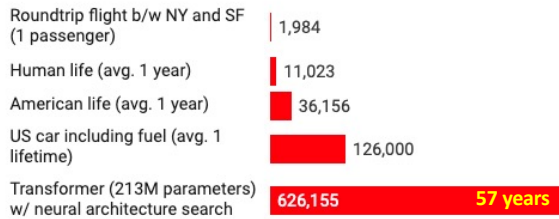


# Waste of Resources?

Green Computing

## Common carbon footprint benchmarks

in lbs of CO2 equivalent



Year	Model	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Dataset Size	Cloud compute cost (USD)
	BERT (110M parameters)	Oct, 2018	1,507	1,438		\$3,751-\$12,571
	ELMo	Feb, 2018	275	262		\$433-\$1,472
	GPT-2	Feb, 2019	-	-		\$12,902-\$43,008
	Transformer (213M parameters)	Jun, 2017	201	192		\$289-\$981
	Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155		\$942,973-\$3,201,722
	Transformer (65M parameters)	Jun, 2017	27	26		\$41-\$140

Note: Because of a lack of power draw data on GPT-2's training hardware, the researchers weren't able to calculate its carbon footprint.

Table: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper



[Bender, Gebru et al., 2021]

# Waste of Resources?

Green Computing

## On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

Emily M. Bender\*  
ebender@uw.edu  
University of Washington  
Seattle, WA, USA

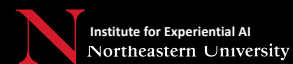
Angelina McMillan-Major  
aymm@uw.edu  
University of Washington  
Seattle, WA, USA

Timnit Gebru\*  
timnit@blackinai.org  
Black in AI  
Palo Alto, CA, USA

Shmargaret Shmitchell  
shmargaret.shmitchell@gmail.com  
The Aether



FaccT 2021





WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY **Meta Ethics**

ALEX HANNA MEREDITH WHITTAKER IDEAS 12.31.2020 07:00 AM

## Timnit Gebru's Exit From Google Exposes a Crisis in AI

The situation has made clear that the field needs to change. Here's where to start, according to a current and a former Googler.

Margaret Mitchell, Feb 20



[Towards Intellectual Freedom in an AI Ethics Global Community](#)

Institute for Experiential AI

Institute for Experiential AI  
Northeastern University

**Principles**

## Pragmatical Questions

- To which part of the system applies?
- Are all equally important?
- To whom is important?
- Are they orthogonal?
- Can they be fulfilled simultaneously?
- Do they make sense together?
  - Transparency vs. Accountability
- Is it really a principle or a tool/requirement to achieve a principle?

Institute for Experiential AI

Institute for Experiential AI  
Northeastern University

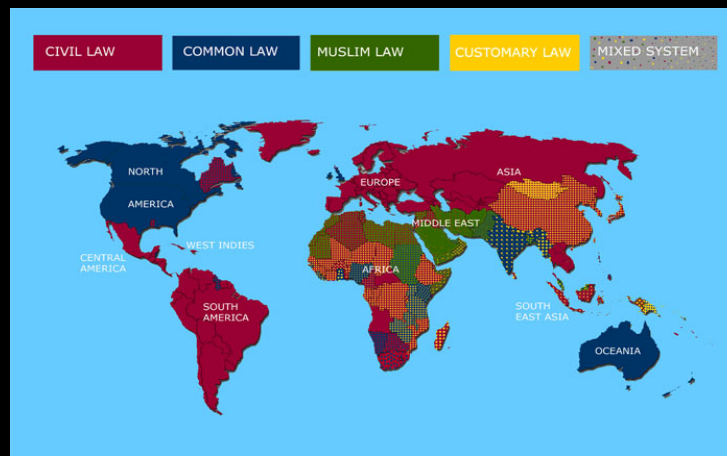
# It's Complicated

Principles

- **Awareness**
  - Autonomy & Integrity
- **Data Provenance:**
  - Equity & Bias
  - Traceability
  - Access and Redress
  - Quality Assurance
- **Completeness:**
  - Interpretability
  - Adaptability
  - Scalability
  - Extensibility
  - Interoperability
  - Quality Assurance
- **Usability:**
  - Efficiency
  - Accessibility
  - Resilience
  - Reproducibility
- **Transparency:**
  - Explainability
  - Validation & Testing
  - Documentation
  - Auditability
- **Responsibility:**
  - Privacy, Security & Safety
  - Proportionality, Sustainability
  - Trustworthiness, Accountability
  - Maintenance, Legal compliance
  - Beneficial/Wellbeing

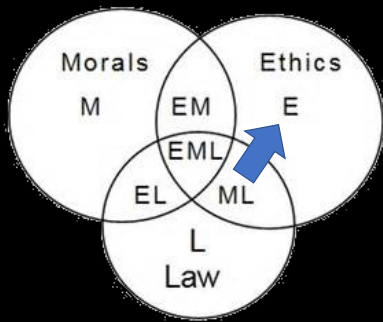
# Legal and Ethical Colonialism

Cultural Differences

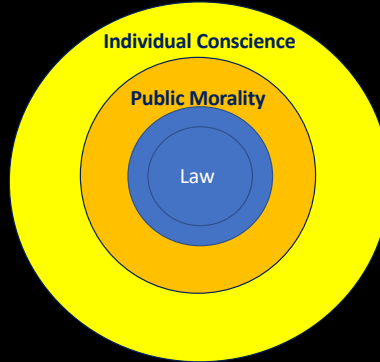


Technological  
Humanism

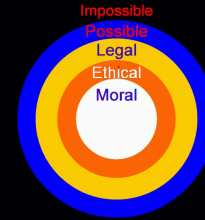
# Religious Differences



Christian



Muslim



???

# Geographical Diversity

Ubuntu ethics is defined as a set of central values among which are reciprocity, common good, peaceful relations, human dignity, and the value of human life as well as consensus, tolerance, and mutual respect [Ujomudike, 2015].

I am because we are

MENU / Q / f / @ / t acoll DONATE / NEWSLETTER / PSYCHE / SIGN IN

Descartes was wrong: 'a person is a person through other persons'

Abeba Birhane

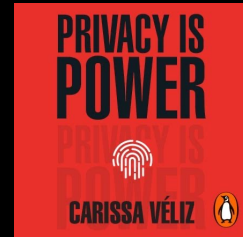
7 April 2017

"Humanity" in Bantu languages

Language	Word	Countries
Chewa	umunthu	Malawi, Zambia
Zulu and Xhosa	ubuntu	South Africa
Sesotho	botho	South Africa
Shona	unhu, hunhu	Zimbabwe
Swahili	utu	Kenya, Tanzania
Meru	munto <sup>[a]</sup>	Kenya
Kikuyu	umundu <sup>[a]</sup>	Kenya
Herero	omundu	Namibia
Tswana	muthu	Botswana
Kongo	gimuntu	Angola
Tonga	vumuntu	Mozambique

# Identity, Data Protection & Privacy

- Public Opinion vs. Collective Privacy?
  - Our privacy is tied to the privacy of our social circles
  - Freedom of expression vs. data protection rights (GDPR, EU)
  - I can do everything that is not forbidden vs. I can do only what is allowed
- Digital nudging
  - Anonymity vs. Privacy
  - Awareness
  - Consent/Legal Basis
  - Minimal data collection
  - Minimal time stored



## GDPR - Article 22 – Automated individual decision-making, including profiling

- The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
- Paragraph above shall not apply if the decision:
  - a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  - b) is **authorised** by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
  - c) is based on the data subject's **explicit consent**.
- In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and **to contest the decision**.

## What this Means?

You must identify whether any of your data processing falls under Article 22 and, if so, make sure that you:

- Give individuals information about the processing;
  - If you are using ML, you at least need **interpretability**
- Introduce simple ways for them to request human intervention or challenge a decision;
  - If you are using ML, you may need **to explain**
- Carry out regular checks to make sure that your systems are working as intended.
  - You may need **continuous validation, testing, and maintenance**.

**Discrimination**

## GDPR in Action

- Competence
- Consent
- Proportionality
- One Size Fits All
  - All human rights, domains, sizes, etc.
- Technological solutionism vs normative solutionism
  - [Jaume-Palasi, to appear]

**French high court rules against biometric facial recognition use in high schools**

🕒 Feb 28, 2020 | [Luana Pascu](#)



# Regulation

- Internet Companies Antitrust
  - Amazon's Antitrust Paradox [Khan, 2017]
  - Google US's DoJ Antitrust (2020/10-?)
  - Facebook US's FTC Antitrust (2020/12-?)
- Should marketplaces sell in their own marketplace?
  - Yes, but with regulations [Hagiu, Teh & Smith, 2020]
  - Is data asymmetry ethical? (not new, but gets amplified in eCommerce)
- Fair markets could be better revenue wise
  - Fairness trade-offs [Mehrotra et al., 2018; Baeza-Yates & Delnevo, to appear]
- Ethical boards should be compulsory when humans may suffer
  - At least 50% external



# EU Proposal (April 21, 2021)

- Forbidden uses
- High-risk systems and requirements
- EU database for stand-alone high-risk systems
- Transparency obligations
- Governance
- Monitoring, information sharing and market surveillance
- Codes of conduct
- Confidentiality and penalties

## TITLE II

### PROHIBITED ARTIFICIAL INTELLIGENCE PRACTICES

#### Article 5

1. The following artificial intelligence practices shall be prohibited:
- (a) the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;
  - (c) the placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:
    - (i) detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected;
    - (ii) detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity;
  - (d) the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless and in as far as such use is strictly necessary for one of the following objectives:
    - (i) the targeted search for specific potential victims of crime, including missing children;
    - (ii) the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack;
    - (iii) the detection, localisation, identification or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA<sup>62</sup> and available in the Member

Proposal for a  
**REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL**

**LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE  
(ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION  
LEGISLATIVE ACTS**

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

The use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement for any of the objectives referred to in paragraph 1 point d) shall take into account the following elements:

- (a) the nature of the situation giving rise to the possible use, in particular the seriousness, probability and scale of the harm caused in the absence of the use of the system;
- (b) the consequences of the use of the system for the rights and freedoms of all persons concerned, in particular the seriousness, probability and scale of those consequences.

In addition, the use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement for any of the objectives referred to in paragraph 1 point d) shall comply with necessary and proportionate safeguards and conditions in relation to the use, in particular as regards the temporal, geographic and personal limitations.

**N** Institute for Experiential AI  
Northeastern University

#### ANNEX III HIGH-RISK AI SYSTEMS REFERRED TO IN ARTICLE 6(2)

High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:

1. Biometric identification and categorisation of natural persons:
  - (a) AI systems intended to be used for the 'real-time' and 'post' remote biometric identification of natural persons;
2. Management and operation of critical infrastructure:
  - (a) AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity.
3. Education and vocational training:
  - (a) AI systems intended to be used for the purpose of determining access or assigning natural persons to educational and vocational training institutions;
  - (b) AI systems intended to be used for the purpose of assessing students in educational and vocational training institutions and for assessing participants in tests commonly required for admission to educational institutions.
4. Employment, workers management and access to self-employment:
  - (a) AI systems intended to be used for recruitment or selection of natural persons, notably for advertising vacancies, screening or filtering applications, evaluating candidates in the course of interviews or tests;
  - (b) AI intended to be used for making decisions on promotion and termination of work-related contractual relationships, for task allocation and for monitoring and evaluating performance and behavior of persons in such relationships.
5. Access to and enjoyment of essential private services and public services and benefits:
  - (a) AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services;
  - (b) AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems put into service by small scale providers for their own use;
  - (c) AI systems intended to be used to dispatch, or to establish priority in the dispatching of emergency first response services, including by firefighters and medical aid.

#### 6. Law enforcement:

- (a) AI systems intended to be used by law enforcement authorities for making individual risk assessments of natural persons in order to assess the risk of a natural person for offending or reoffending or the risk for potential victims of criminal offences;
- (b) AI systems intended to be used by law enforcement authorities as polygraphs and similar tools or to detect the emotional state of a natural person;
- (c) AI systems intended to be used by law enforcement authorities to detect deep fakes as referred to in article 52(3);
- (d) AI systems intended to be used by law enforcement authorities for evaluation of the reliability of evidence in the course of investigation or prosecution of criminal offences;
- (e) AI systems intended to be used by law enforcement authorities for predicting the occurrence or reoccurrence of an actual or potential criminal offence based on profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 or assessing personality traits and characteristics or past criminal behaviour of natural persons or groups;
- (f) AI systems intended to be used by law enforcement authorities for profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of detection, investigation or prosecution of criminal offences;
- (g) AI systems intended to be used for crime analytics regarding natural persons, allowing law enforcement authorities to search complex related and unrelated large data sets available in different data sources or in different data formats in order to identify unknown patterns or discover hidden relationships in the data.

#### 7. Migration, asylum and border control management:

- (a) AI systems intended to be used by competent public authorities as polygraphs and similar tools or to detect the emotional state of a natural person;
- (b) AI systems intended to be used by competent public authorities to assess a risk, including a security risk, a risk of irregular immigration, or a health risk, posed by a natural person who intends to enter or has entered into the territory of a Member State;
- (c) AI systems intended to be used by competent public authorities for the verification of the authenticity of travel documents and supporting documentation of natural persons and detect non-authentic documents by checking their security features;
- (d) AI systems intended to assist competent public authorities for the examination of applications for asylum, visa and residence permits and associated complaints with regard to the eligibility of the natural persons applying for a status.

#### 8. Administration of justice and democratic processes:

- (a) AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts.

# Future Regulation?

- **Algorithmic Accountability Act** (2019): The bill was introduced by Senator Cory Booker (D-NJ), Ron Wyden (D-OR), and Representative Alexandria Ocasio-Cortez (D-NY). According to **Senator Wyden**, the bill would "require companies to study the algorithms they use, identify any discrimination or bias they find, and fix any discrimination or bias they find."
- **Consumer Online Privacy Rights Act** (2019): The bill was introduced by Senator Maria Cantwell (D-WA), and would have established new rules for companies that use algorithmic decision-making to process personal information.
- **Justice in Policing Act** (2020): The bill was sponsored by Kamala Harris (D-CA), Senator Cory Booker (D-NJ), a Representative Pramila Jayapal (D-WA), a Representative Karen Bass (D-CA) and Jerrold Nadler (D-NY). It would have established a federal restriction on facial recognition technology.
- **Facial Recognition and Biometric Technology Moratorium Act** (2019): Sponsored by Senator Edward Markey (D-MA) and Representative Pramila Jayapal (D-WA) with Representatives Pramila Jayapal (D-WA) and Aya Gruber (D-CA). The bill would have established a five-year moratorium on the use of facial recognition technology. It is set to be **reintroduced** this year.

## The White House Launches the National Artificial Intelligence Initiative Office

INFRASTRUCTURE & TECHNOLOGY | Issued on: January 12, 2021

ABOUT EXPERTS EVENTS PUBLICATIONS BLOG DOCUMENTS

Constitution and Law Economics Politics

FEBRUARY 5, 2021 3:37PM

### Algorithmic Bias Under the Biden Administration

By Matthew Feeney and Rachel Chiu



# Registering Algorithms

VB The Machine GamesBeat Jobs Special Issue Become a Member

The Machine Making sense of AI

## Amsterdam and Helsinki launch algorithm registries to bring transparency to public deployments of AI

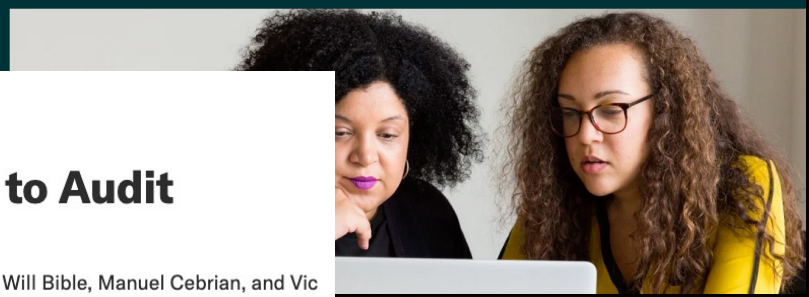
Khari Johnson @kharjohnson September 28, 2020 11:41 AM



# Auditing Algorithms

## What algorithm auditing startups need to succeed

Khari Johnson @kharijohnson January 30, 2021 8:45 AM



Harvard Business Review Economics & Society

### Why We Need to Audit Algorithms

by James Guszcza, Iyad Rahwan, Will Bible, Manuel Cebrian, and Vic Katyal

November 28, 2018

Institute for Experiential AI  
Northeastern University

## Building and Auditing Fair Algorithms: A Case Study in Candidate Screening

Christo Wilson  
Northeastern University  
cbw@ccs.neu.edu

Avijit Ghosh  
Northeastern University  
avijit@ccs.neu.edu

Shan Jiang  
Northeastern University  
sjiang@ccs.neu.edu

Alan Mislove  
Northeastern University  
amislove@ccs.neu.edu

Lewis Baker  
pymetrics, inc.  
lewis@pymetrics.com

Janelle Szary  
pymetrics, inc.  
janelle@pymetrics.com

Kelly Trindel  
pymetrics, inc.  
kelly@pymetrics.com

Frida Polli  
pymetrics, inc.  
frida.polli@pymetrics.com

FaccT 2021



### Auditing Algorithms @ Northeastern

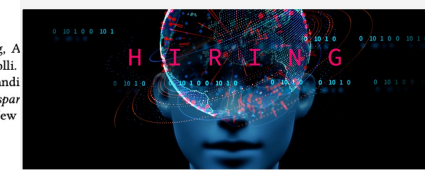
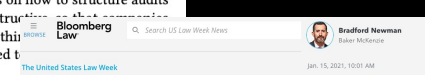
#### ABSTRACT

Academics, activists, and regulators are increasingly urging companies to develop and deploy sociotechnical systems that are fair and unbiased. Achieving this goal, however, is complex: the developer must (1) deeply engage with social and legal facets of "fairness" in a given context, (2) develop software that concretizes these values, and (3) undergo an independent algorithm audit to ensure technical correctness and social accountability of their algorithms. To date, there are few examples of companies that have transparently undertaken all three steps.

In this paper we outline a framework for algorithmic auditing by way of a case-study of pymetrics, a startup that uses machine learning to recommend job candidates to their clients. We discuss how pymetrics approaches the question of fairness given the constraints of ethical, regulatory, and client demands, and how pymetrics' software implements adverse impact testing. We also present the results of an independent audit of pymetrics' candidate screening tool.

We conclude with recommendations on how to structure audits to be practical, independent, and constructive. We argue that having better incentive to participate in this type of audit and watchdog groups can be better prepared to

**ACM Reference Format:**  
Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Janelle Szary, Kelly Trindel, and Frida Polli. Building Fair Algorithms: A Case Study in Candidate Screening on Fairness, Accountability, and Transparency. *Proceedings of the ACM*, 2021. <https://doi.org/10.1145/3442188.3445928>



Using AI to Make Hiring Decisions? Prepare for EEOC Scrutiny

# Bad (Human) Practices

- Learn from the Past Without Remembering the Context
- Learn from Humans Without Remembering Human Bias and the Possibility of Malicious Training
- Not Checking for Spurious Correlation/Proxies for Protected Information
- Code Reused in Unanticipated Contexts
- Tendency to Aggressively Resist Review
- Inappropriate Relationship of Human Decision Maker to System
- Failing to Measure Impact of Deployed System
- Individual Personalization instead of Personas
  - Trade-off with privacy
- Inaccurate Data or Just Data that you Have

Partially based in [Matthews, 2020]

# Our Professional Biases

- Problems
  - Our **big data and deep learning bias**: **small data** is more frequent & harder [Baeza-Yates, KD Nuggets, 2018]
- Design and Implementation
  - Do systems reflect the characteristics of the designers?
  - Do systems reflect the characteristics of the coders?
- Evaluation [Silberzahn et al., COS, Univ. of Virginia, 2015] [Johansen et al., Norway, 2020]
  - Choose the right experiment
  - Choose the right test data
  - Choose the right metric(s)
  - Choose the **right baseline(s)**
  - Julio Gonzalo's talk: <http://tiny.cc/ESSIR2019-juliogonzalo>

## What We Can Do?

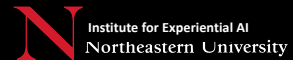
- Data
  - Analyze for known and unknown biases, debias/mitigate when possible
  - Recollect more data for sparse regions of the solution space
  - Do not use attributes associated directly/indirectly with harmful bias
- Interaction
  - Make sure that the user is **aware** of the biases all the time
  - Give more control to the user
- Design & Implementation
  - Make sure that the model is **aware** of the bias and if possible deal with it
  - Let experts/colleagues/users contest every step of the process
- Evaluation & Deployment
  - Do not fool yourself!
  - Error & sensibility analysis (*e.g.*, synthetic data if possible)
  - Algorithms registration / Certification / External Auditing

## Recommendations for Us

- Design for People First!
- Deep Respect for Limitations of Our Systems
  - Assumptions, ethical risks, etc.
- Learning from the Past does not mean to Reproduce It
- Have and Ethics Board and enforce a Code of Ethics
- Improve Explainability (repeat 100 times)
- More evaluation and cross-discipline validation
- Research Best Practices with **Humans in Control** and **Machines in the Loop**
  - Better than “Human in the Loop”!
- Check the ethics of your providers & clients

# Final Take-Home Messages

- Systems are a mirror of us, **the good, the bad and the ugly**
- To be fair, we need to be aware of our **own biases**
- Who profits/suffers technology, transhumanism vs. humanism
- Ethics is **complicated**, do not underestimate it!
- **Plenty** of open research problems! (in **small data** even more!)



## Questions?

ASIST 2012  
Book of the  
Year Award  
(Biased Ad)

Modern  
Information Retrieval  
the concepts and technology behind search  
Second edition

### New Conferences that started in 2018:

AAAI/ACM Conference on AI, Ethics, and Society  
<http://www.aies-conference.com>

Conference on Fairness, Accountability, and Transparency  
<http://facctconference.org>



Ricardo Baeza-Yates  
Berthier Ribeiro-Neto

Contact: [rbaeza@acm.org](mailto:rbaeza@acm.org)  
[www.baeza.cl](http://www.baeza.cl)  
[@polarbearBY](https://twitter.com/polarbearBY)

## Biased Questions?

