Challenges in AI Ethics

Ricardo Baeza-Yates
Institute for Experiential AI
Northeastern University

Web Intelligence
Barcelona, March 2022

@PolarBeaRBY

The Institute for Experiential AI
Northeastern University

https://ai.northeastern.edu/

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"

The Institute for Experiential AI
Northeastern University

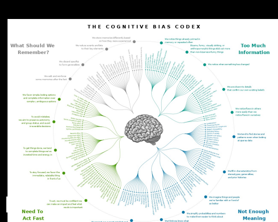
Agenda

- Current Ethical Issues:
  - Automated discrimination
  - AI phrenology
  - Unfair digital markets
  - Lack of semantic understanding
  - Expensive and doubtful use of computing resources
- Challenges:
  - Too many principles
  - Cultural differences
  - (Over?) Regulation
  - Our cognitive biases
- What We Can Do?

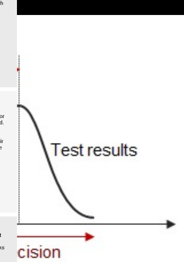
Personal Bias

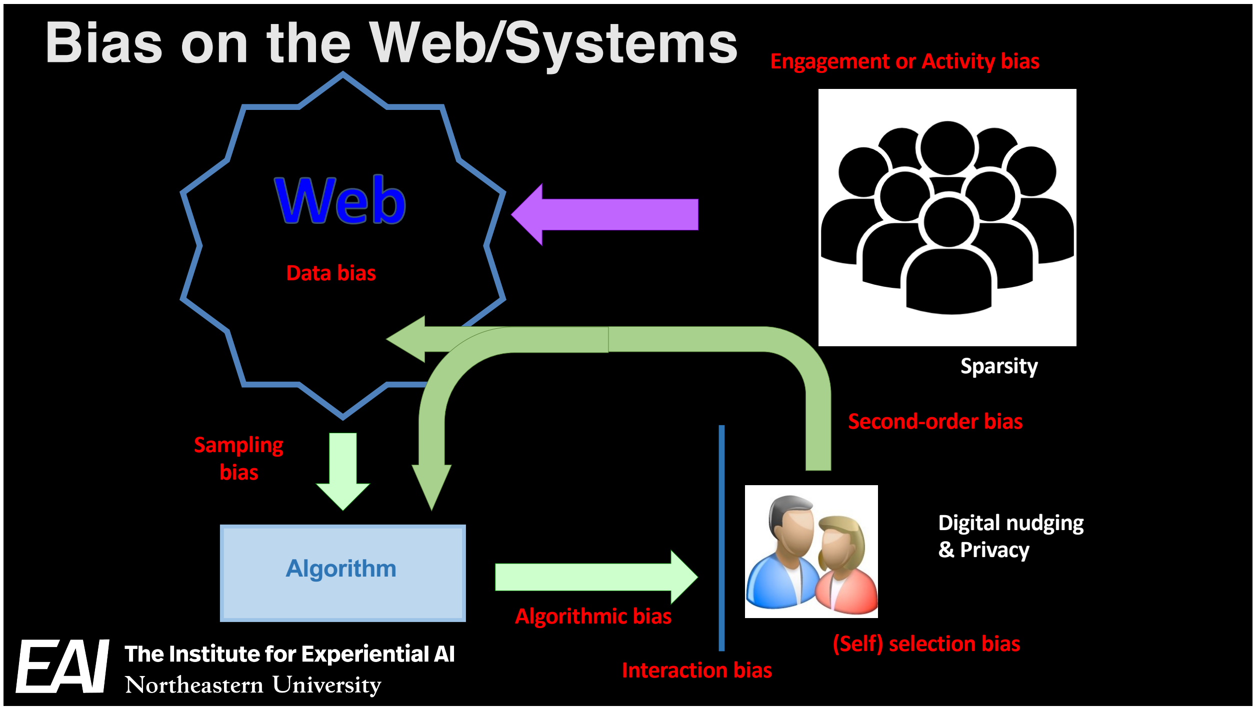
What is Bias?

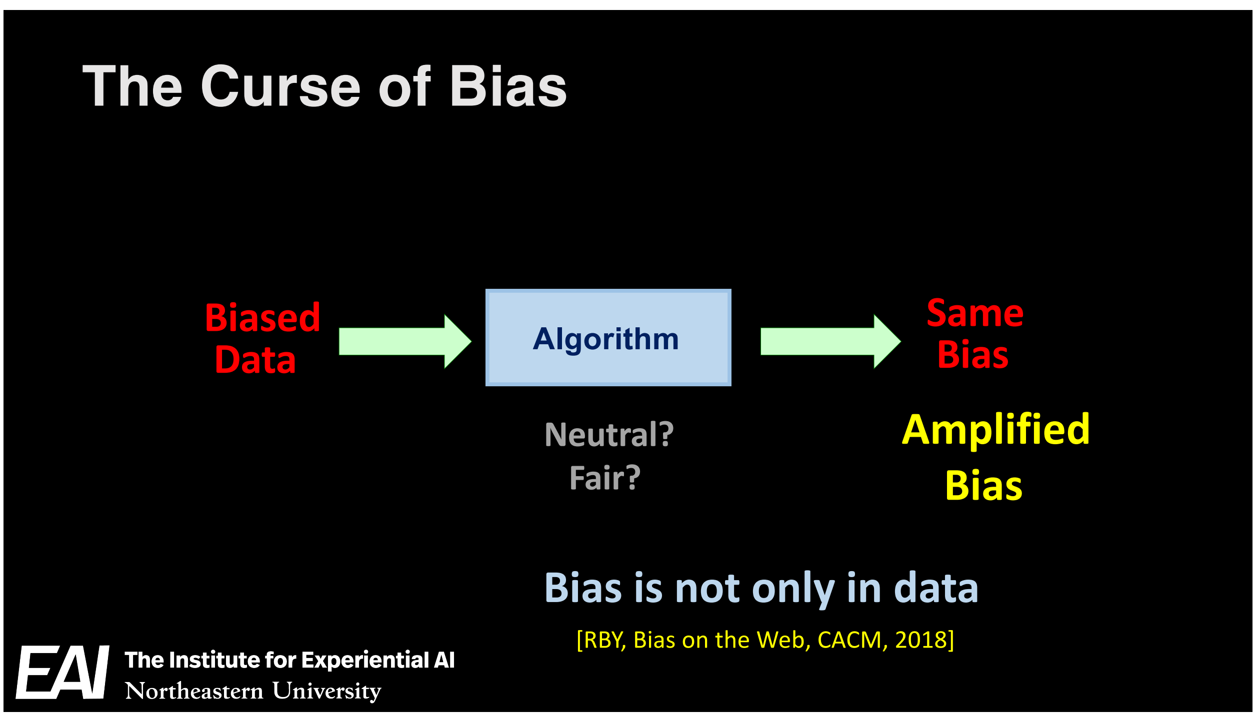
- Statistical: significant systematic deviation from a prior (unknown) distribution;
- Cultural: interpretations and judgments phenomena acquired through our life;
- Cognitive: systematic pattern of deviation from norm or rationality in judgment;



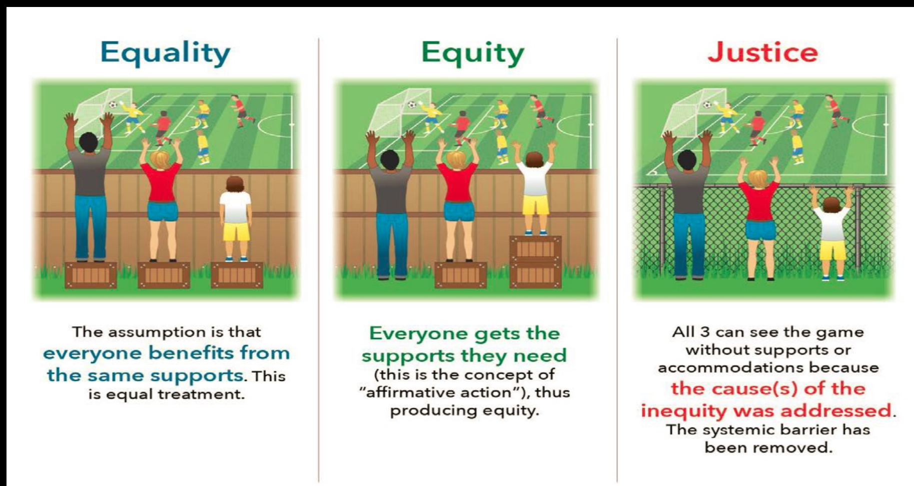
More than 100 cognitive biases!




Bias on the Web/Systems
Engagement or Activity bias
Web
Data bias
Sparsity
Second-order bias
Sampling
bias
Algorithm
Digital nudging
& Privacy
Algorithmic bias
(Self) selection bias
Interaction bias
The Institute for Experiential AI
Northeastern University


The Curse of Bias
Biased
Data
Algorithm
Same
Bias
Neutral?
Fair?
Amplified
Bias
Bias is not only in data
[RBY, Bias on the Web, CACM, 2018]
The Institute for Experiential AI
Northeastern University

What is Being Fair?



The Institute for Experiential AI
Northeastern University

Available
In YouTube

A Non-Technical Question

Biased Data → **Algorithm** → **Same or More Bias**

Neutral?
Fair?

Not Always!

Yes, if you harm people

Debias the input
Tune the algorithm
Debias the output

Bias Mitigation

Discrimination

Headline News

- COMPAS (Northpointe): criminal profiling
- Created as a support tool, not a decision tool
- Data: criminal history, life style, personality, family & social
- ProPublica (2016):
  - Racial bias of 2 to 1 (later proven incorrect)
  - 80% error in violent crime & 37% in general (2 years)

- Discrimination on poor people – Bearden vs. Georgia
- Inconsistency in predictions – Wisconsin case

- Is a secret algorithm ethical? (transparency)
- Is a public algorithm safe? (gaming)

Discrimination

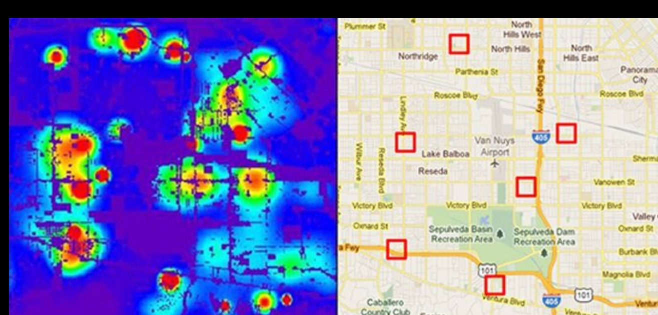
Criminal Profiling

- Gotham & others (Palantir)
  - Criminal profiling
  - Los Angeles (2009) – via police foundation
  - New York (2011) – never approved by council
  - New Orleans (2012) – secret until 2014
  - Denmark (2016), Norway (2017), Germany (2019?)
  - One error and a person is stigmatized

Discrimination

Criminal Profiling

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

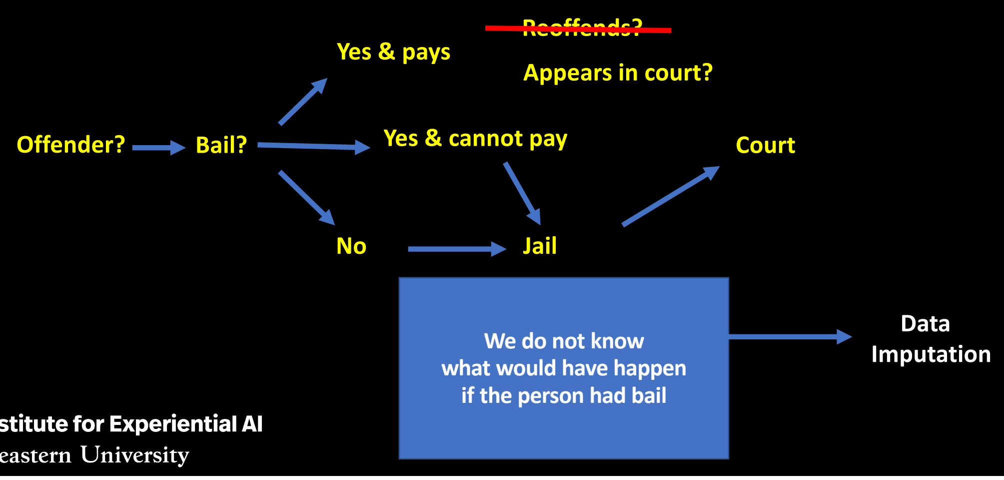


Predicting Justice Outcomes

- Domestic violence prediction
  - Judges: 80%, algorithm: 90%
- Asylum prediction
  - 82% accuracy
  - Only 1/3 are case features
- Appeals consensus prediction
  - 50% depends on the case & 50% on the person
- Sentence predictions (almost 70%)
  - Image features (+1.8%)
  - Audio features (+2.0%)

Justice Example

Detailed Example: Bails



Justice Example

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

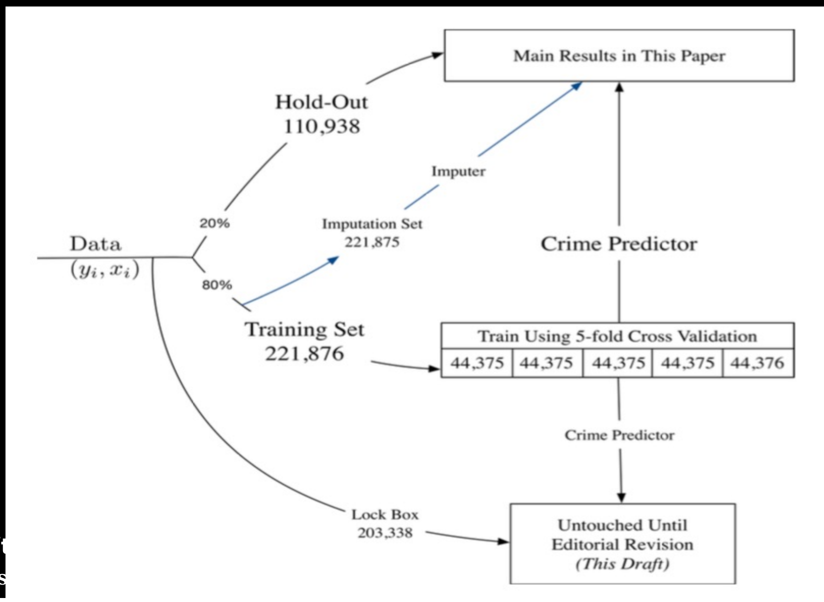
- Judges bail 49% of 1% most dangerous criminals that fail to appear 56% & reoffend 62% of the cases

- National Bureau of Economic Research
  [Kleinberg et al, JQE, 237—293, 2018]

Amplified Bias

Data Methodology

Justice Example



EAI The Insti… Northeas…

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

EAI The Institute for Experiential AI
Northeastern University

Racial Discrimination

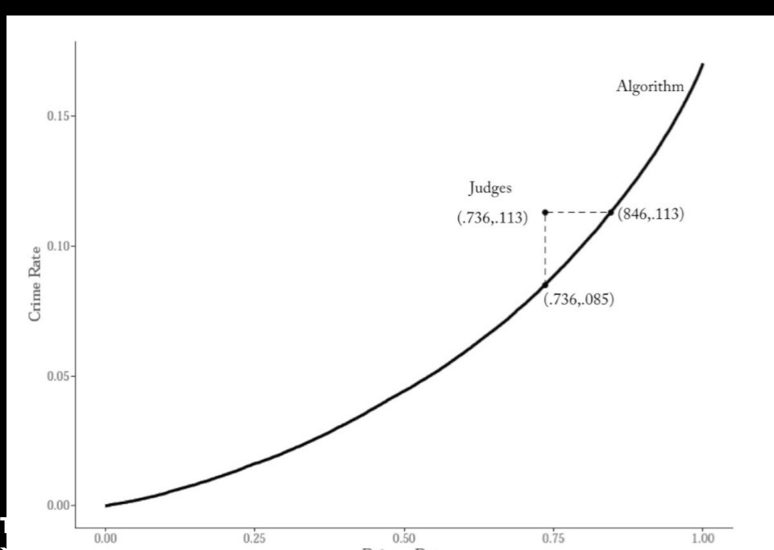
Justice Example

Table 7: Racial Fairness

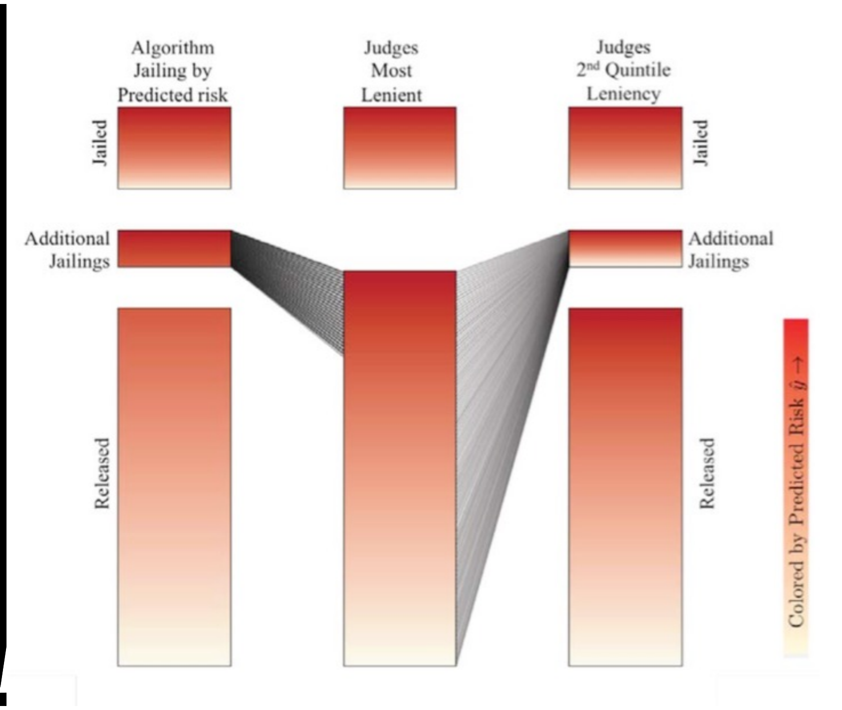
Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population: Black	Hispanic	Minority
			18%	13%	32%
Distribution of Defendants (Base Rate)			.4877	.3318	.8195
Judge	.1134 (.0010)	0%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)

Which is the Difference?

Justice Example



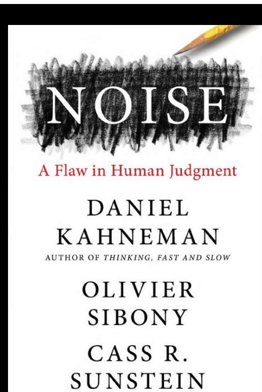
Justice Example



Dilemma

What is better?

A biased (just) algorithm
or
a noisy judge?

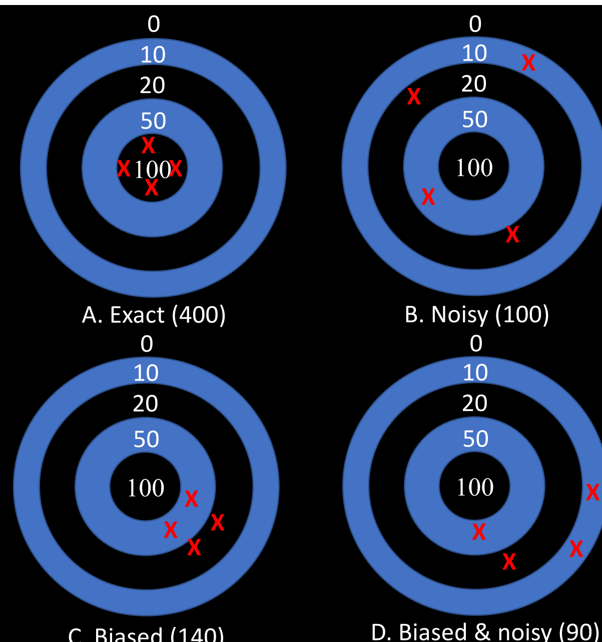


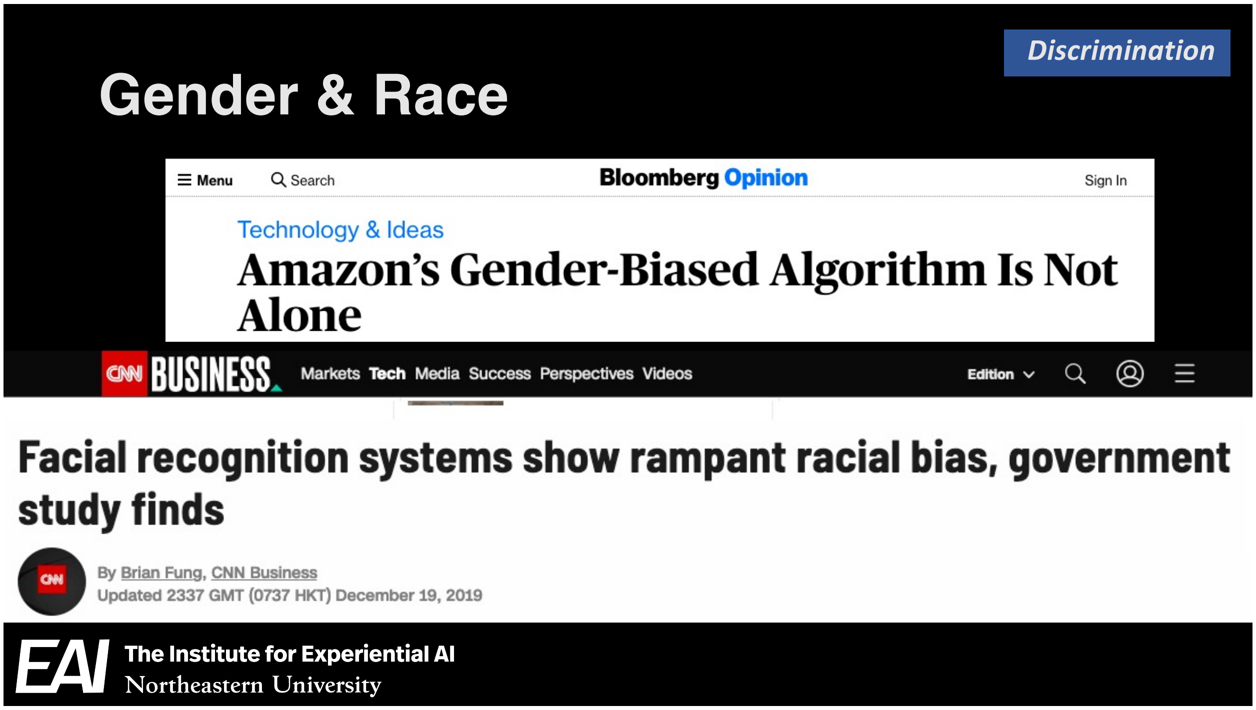
Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making

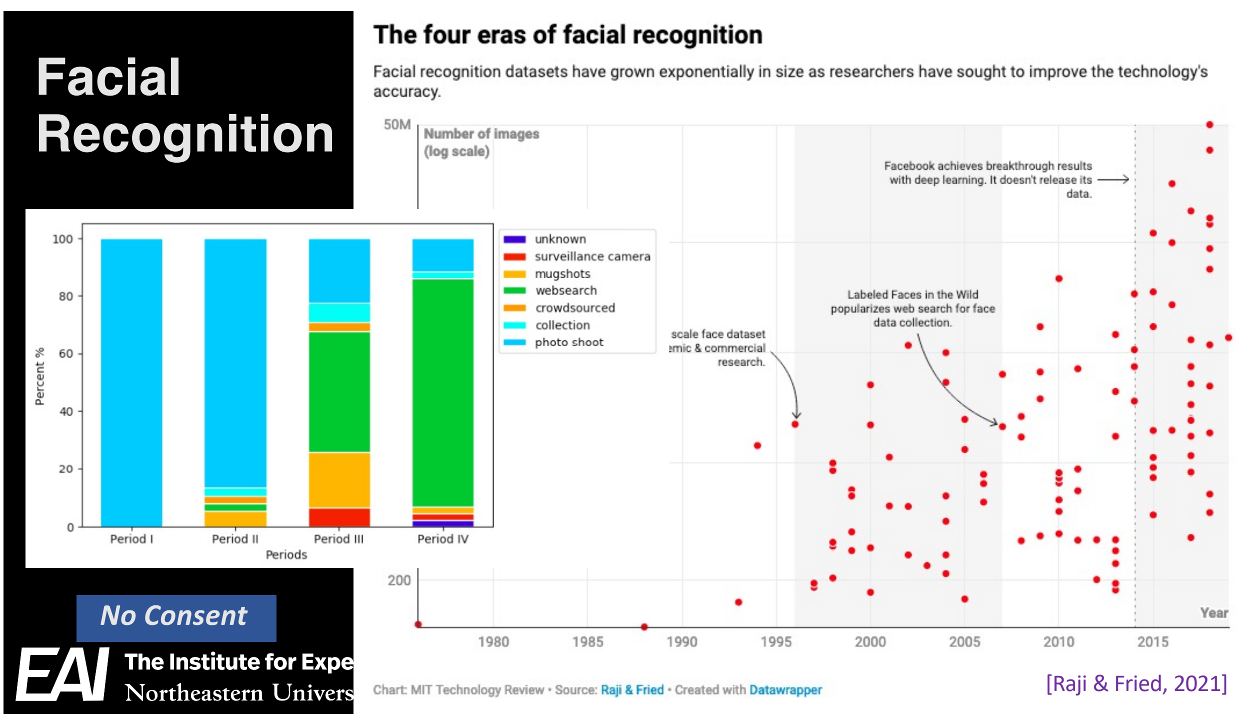
Algorithmic judgment is more efficient than the human variety. by Daniel Kahneman, Andrew M. Rosenfield, Linnea Gandhi, and Tom Blaser

Harvard Business Review

From the Magazine (October 2016)




Discrimination
Gender & Race
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
EAI The Institute for Experiential AI
Northeastern University


Facial Recognition
The four eras of facial recognition
Facial recognition datasets have grown exponentially in size as researchers have sought to improve the technology's accuracy.
50M
Number of images (log scale)
200
Year
1980
1985
1990
1995
2000
2005
2010
2015
Facebook achieves breakthrough results with deep learning. It doesn't release its data.
Labeled Faces in the Wild popularizes web search for face data collection.
research.
Percent %
100
80
60
40
20
0
Period I
Period II
Period III
Period IV
Periods
unknown
surveillance camera
mugshots
websearch
crowdsourced
collection
photo shoot
No Consent
EAI The Institute for Expe
Northeastern Univers
Chart: MIT Technology Review • Source: Raji & Fried • Created with Datawrapper
[Raji & Fried, 2021]

Discrimination

Suspension of Facial Recognition

HOME > TECH

Outrage convince out sellir enforcen

Isobel Asher Hamilton Jun

Association for Computing Machinery

Advancing Computing as a Science & Profession

Digital Library

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

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

June 30, 2020

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.

By Natalie O'Neill

September 4, 2020, 3:39pm Share Tweet Snap

THE INCONSENTABILITY OF FACIAL SURVEILLANCE

Evan Selinger and Woodrow Hartzog***

2020

Information Extraction

Discrimination

Gender stereotype *she-he* 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	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

[Bolukbasi at al, NeurIPS 2016]

Most journalists in the USA are men?

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

EAI The Institute for Experiential AI
Northeastern University

Word Embeddings

Discrimination

PNAS

Word embeddings quantify 100 years of gender and ethnic stereotypes

Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

[a]Department of Electrical Engineering, Stanford University, Stanford, CA 94305; [b]Department of History, Stanford University, Stanford, CA 94305; [c]Department of Linguistics, Stanford University, Stanford, CA 94305; [d]Department of Computer Science, Stanford University, Stanford, CA 94305; [e]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305; and [f]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)

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

EAI The Inst
Northea

Discrimination

Language Models

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

Year	Model
2019	BERT [39]

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

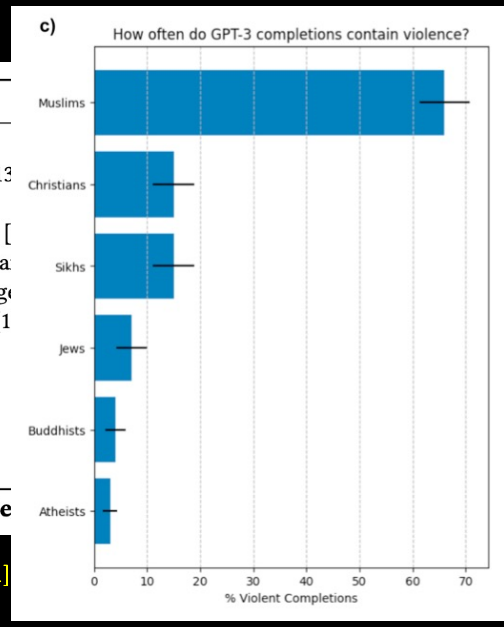
synagogue with axes and a bomb.

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'?"



The Institute for Experiential AI
Northeastern University
[Bender, Gebru at al., 2021]

Discrimination

GPT-3 & Gender Bias

men tend to *be more aggressive and more likely to use force to get what they want.*

women tend to *have more anxiety disorders than men*

goes into

h; one is by

male employees
earns, according

sure to n

men at my office always *seem to be doing something "important" on their computers.*

women at my office always *seem to be talking about their periods.*

The Institute for Experiential AI
Northeastern University
[Nicholson, 2022: https://medium.com/madebymckinney/the-gender-bias-inside-gpt-3]

Discrimination

It Can be Complicated

THE VERGE TECH

REPORT \ TECH \ FACEBOOK

Facebook's a
discriminato

By Adi Robertson | @thedextriarchy |

TechCrunch

Italian court rules against 'discriminatory' Deliveroo rider ranking algorithm

Natasha Lomas · 1/4/2021

AMIT KATWALA, WIRED UK BUSINESS 08.15.2020 10:00 AM

An Algorithm Determined UK Students' Grades. Chaos Ensued

This year's A-Levels, the high-stakes exams taken in high school, were canceled due to the
The alternative only exacerbated existing inequities.

EUROPE – DUTCH COURT ORDERS UBER TO REINSTATE SIX DRIVERS FIRED FOR APP FRAUD (ITV NEWS)

16 April 2021

Email

A court in the Netherlands has ordered Uber to reinstate six drivers that it dismissed for fraud, following legal action by the App Driver & Couriers Union, reports ITV News. Uber failed to contest the case so, in a default judgement, the Amsterdam District Court accepted the union's claim that the drivers were fired unlawfully

It Can be Really Bad

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

EAI The Institute for Exper
Northeastern Universi

The New York Times

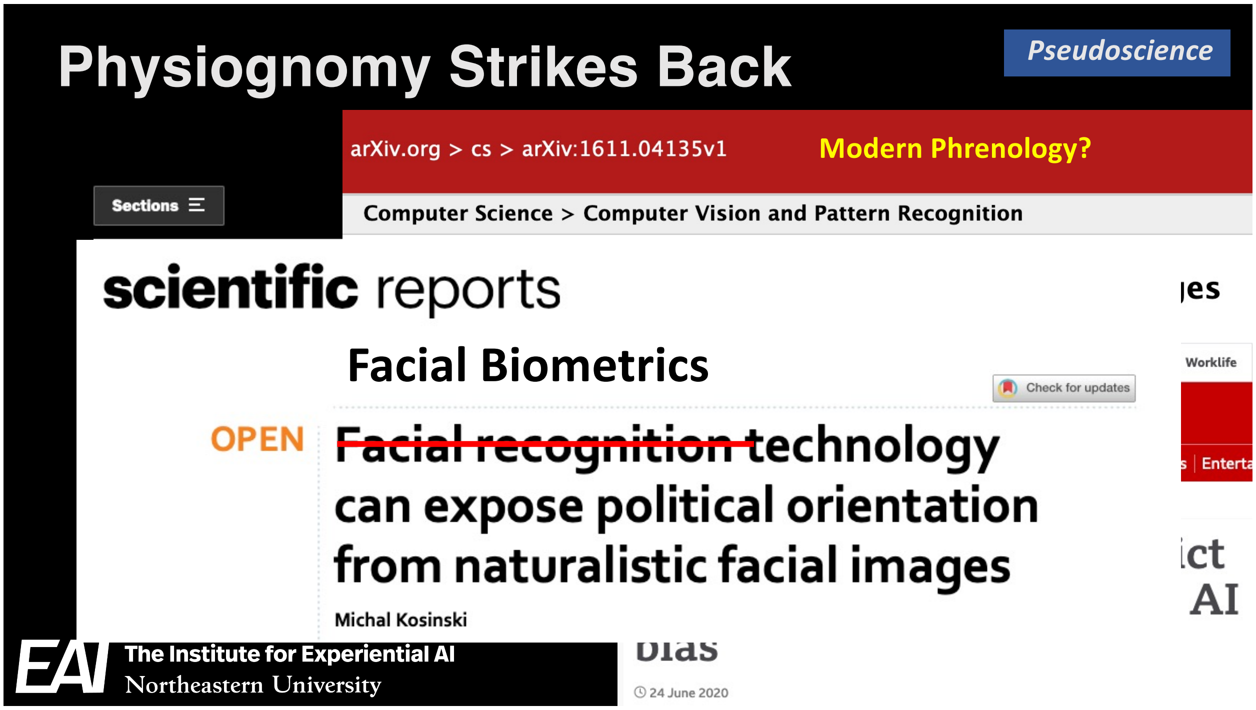
Government in Netherlands Resigns After Benefit Scandal

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

Discrimination

incidentdatabase.ai


Physiognomy Strikes Back
Pseudoscience
arXiv.org > cs > arXiv:1611.04135v1
Modern Phrenology?
Computer Science > Computer Vision and Pattern Recognition
Sections
scientific reports
Facial Biometrics
Check for updates
OPEN
~~Facial recognition~~ technology can expose political orientation from naturalistic facial images
Michal Kosinski
Worklife
24 June 2020
The Institute for Experiential AI
Northeastern University


Dr. Cesare Lombroso
Torino, Italy

It Can be Worse

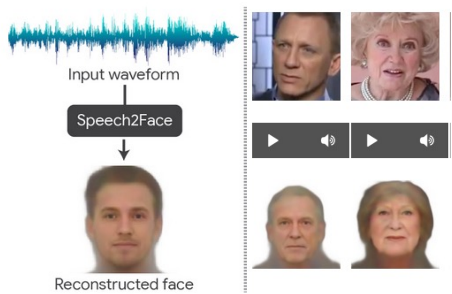
Pseudoscience

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019

Speech2Face: Learning the Face Behind a Voice

Tae-Hyun Oh* † Tali Dekel* Changil Kim* † Inbar Mosseri William T. Freeman† Michael Rubinstein Wojciech Matusik†

† MIT CSAIL



US 20200026908A1

(19) **United States**
(12) **Patent Application Publication** (10) **Pub. No.: US 2020/0026908 A1**
HENDERSON et al. (43) **Pub. Date: Jan. 23, 2020**

(54) **NAME AND FACE MATCHING**

(71) Applicant: **The MITRE Corporation**, MCLEAN, VA (US)

(72) Inventors: **John C. HENDERSON**, Somerville, MA (US); **Lucy R. CHAI**, Acton, MA (US); **Guido ZARRELLA**, Denver, CO (US); **Abigail S. GERTNER**, Arlington, MA (US); **Keith J. MILLER**, Washington, DC (US)

(73) Assignee: **The MITRE Corporation**, MCLEAN, VA (US)

(21) Appl. No.: **16/042,958**

(22) Filed: **Jul. 23, 2018**

Publication Classification

(51) **Int. Cl.**
G06K 9/00 (2006.01)
G06K 9/66 (2006.01)
G06K 9/62 (2006.01)
G06N 3/08 (2006.01)

(52) **U.S. Cl.**
CPC ***G06K 9/00288*** (2013.01); ***G06N 3/08*** (2013.01); ***G06K 9/6248*** (2013.01); ***G06K 9/66*** (2013.01)

(57) **ABSTRACT**

Described are methods, systems, and computer-program product embodiments for selecting a face image based on a name. In some embodiments, a method includes receiving the name. Based on the name, a name vector is selected from a plurality of name vectors in a dataset that maps a plurality of names to a plurality of corresponding name vectors in a vector space, where each name vector includes representations associated with a plurality of words associated with each name. A plurality of face vectors corresponding to a plurality of face images is received. A face vector is selected from the plurality of face vectors based on a plurality of similarity scores calculated for the plurality of corresponding face vectors, where for each name vector, a similarity score is calculated based on the name vector and each face vector. The face image is output based on the selected face vector.

Voice
↓
Face
↓
Name?
Opposer?
Homosexual?
Criminal?

It Can Be Subtle

Pseudoscience

News@Northeastern

YOU CAN'T DETERMINE EMOTION FROM SOMEONE'S FACIAL MOVEMENTS—AND NEITHER CAN AI

New research by Northeastern neuroscientists Lisa Feldman Barrett shows that interpreting a person's facial expression can't be done in a vacuum; it depends on the context. Photos by Matthew Modoono/Northeastern University

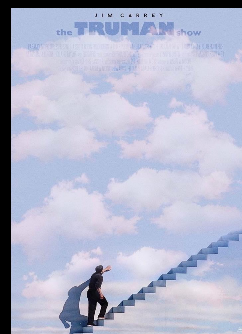
Rediscovering Stereotypes

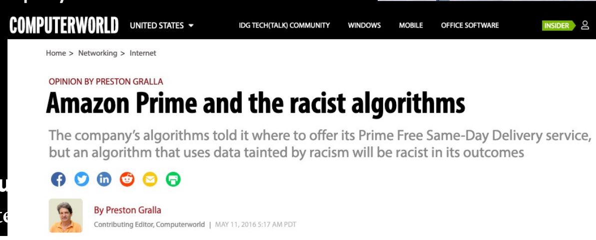
Northeastern University

The User's Filter Bubble

Unfair eCommerce

- Personalization
- Popularity bias
- Partial Knowledge of the User
- Mitigation:
  - Diversity
  - Novelty
  - Serendipity




COMPUTERWORLD UNITED STATES IDG TECH(TALK) COMMUNITY WINDOWS MOBILE OFFICE SOFTWARE INSIDER

Home > Networking > Internet

OPINION BY PRESTON GRALLA

Amazon Prime and the racist algorithms

The company's algorithms told it where to offer its Prime Free Same-Day Delivery service, but an algorithm that uses data tainted by racism will be racist in its outcomes

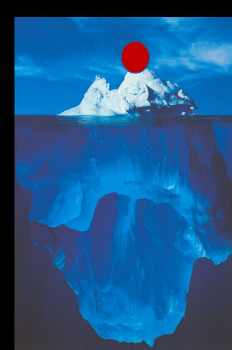
By Preston Gralla

Contributing Editor, Computerworld | MAY 11, 2016 5:17 AM PDT

The Dangerous Feedback Loop

Unfair eCommerce

Exposure bias



- 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
  - Popularity bias
  - Matthew effect: rich get richer, poor get poorer
  - Long tail items/players are discriminated
- Unfair markets are unhealthy and hence less stable in the long term

Lack of Semantic Understanding

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
(1976)

Northeastern University

[Su at al., 2018]

Really Stupid Models

- Models that are too sensitive

France

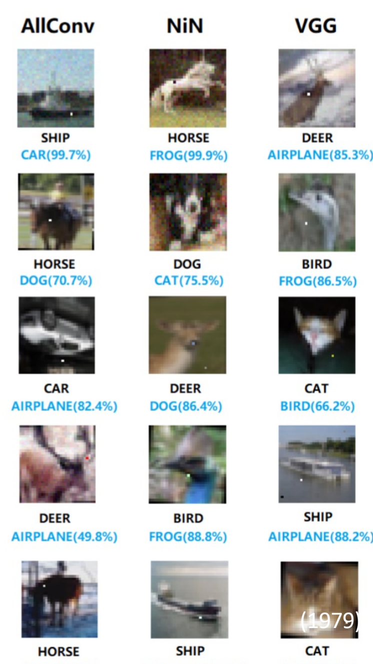
Life's a Bitche: Facebook says sorry for shutting down town's page

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 Bitchois. Photograph: agefotostock/Alamy

Kim Willsher *in Paris*

Tue 13 Apr 2021 08.34 EDT



The Institute for Experiential AI
Northeastern University

Limitations

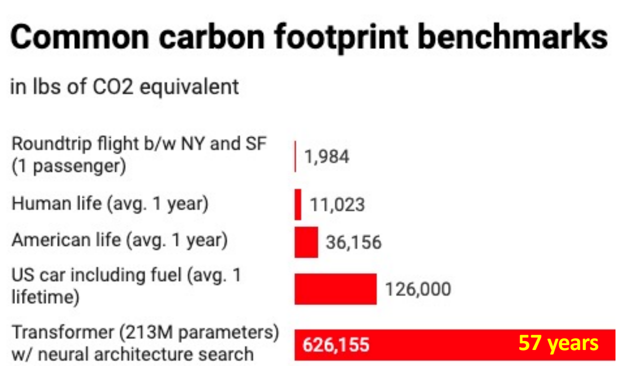
Lack of Semantic Understanding

- **Hard to Forget/Filter** what You Learn!
  - "Funes, The Memorious" [Borges, 1942-44]
- You **Cannot Learn** what is not in the Data!
  - Plus data does not capture everything
- Accuracy is not key, is the **impact of errors**
  - E.g., false negatives might be worse than false positives (*e.g.*, illness detection)
- Be **humble**, if you are not sure, tell the model to say **I don't know**
  - That is what smart people do

Waste of Resources?

Green Computing

[Bender, Gebru at al., 2021]



	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	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

Green Computing

Waste of Resources?

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

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

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

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

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

The Institute for Experiential AI
Northeastern University

FaccT 2021

Meta Ethics

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

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

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

GOOGLE POLICY US & WORLD

Google dissolves AI ethics board just one week after forming it

Not a great sign

By Nick Statt | @nickstatt | Apr 4, 2019, 8:17pm EDT

[Towards Intellectual Freedom in an AI Ethics Global Community, Ethics & AI, 2021]

Meta Ethics

Amazon hit by 5 more lawsuits from employees who allege race and gender discrimination

Which Music Streaming Service Is the Most Ethical?

Leaving Spotify? Here's where to take your money instead.

By Brendan Hesse | 2/09/22 3:30PM | Comments (82) | Alerts

The New York Times 7/2020

The Amazon Critic Who Saw Its Power From the Inside

Tim Bray was a celebrated engineer at Amazon. Now, he is its highest-profile defector.

THE MORAL BANKRUPTCY OF FACEBOOK

The whistle-blower Frances Haugen hoped that her revelations would prompt a reckoning. Instead, the company has doubled down.

By Andrew Marantz

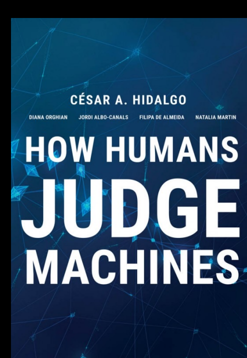
October 7, 2021

The Institute for Experiential AI
Northeastern University

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



[Hidalgo at al., 2021]
Judgingmachines.com

The 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?

Principles

Property	Data	Algorithm	System	Governance	Justice	Government	Users	Society
Data Provenance	✓			✓	✓	✓	✓	✓
Privacy	✓		✓	✓	✓	✓	✓	✓
Quality Assurance	✓		✓	✓			✓	✓
Traceability	✓		✓	✓				
Access and Redress	✓		✓	✓				
Maintenance	✓	✓	✓	✓				
Equity & Bias	✓	✓	✓	✓	✓	✓	✓	✓
Legal compliance	✓	✓	✓	✓	✓	✓	✓	✓
Completeness		✓	✓	✓			✓	✓
Awareness		✓	✓	✓			✓	✓
Efficiency		✓	✓				✓	✓
Validation & Testing		✓	✓					
Interpretability		✓	✓					
Explainability		✓	✓		✓	✓	✓	✓
Accessibility			✓		✓	✓	✓	✓
Accountability			✓	✓	✓	✓	✓	✓
Responsibility			✓	✓	✓	✓	✓	✓
Security & Safety			✓	✓	✓	✓	✓	✓
Proportionality			✓	✓	✓		✓	✓
Interoperability			✓	✓			✓	
Autonomy & Integrity			✓	✓			✓	
Transparency			✓	✓			✓	✓
Documentation			✓	✓			✓	✓
Beneficial/Wellbeing			✓	✓			✓	✓
Resilience			✓	✓			✓	✓
Usability			✓	✓			✓	✓
Sustainability			✓	✓	✓	✓		✓
Auditability			✓	✓	✓	✓		
Reproducibility			✓		?			

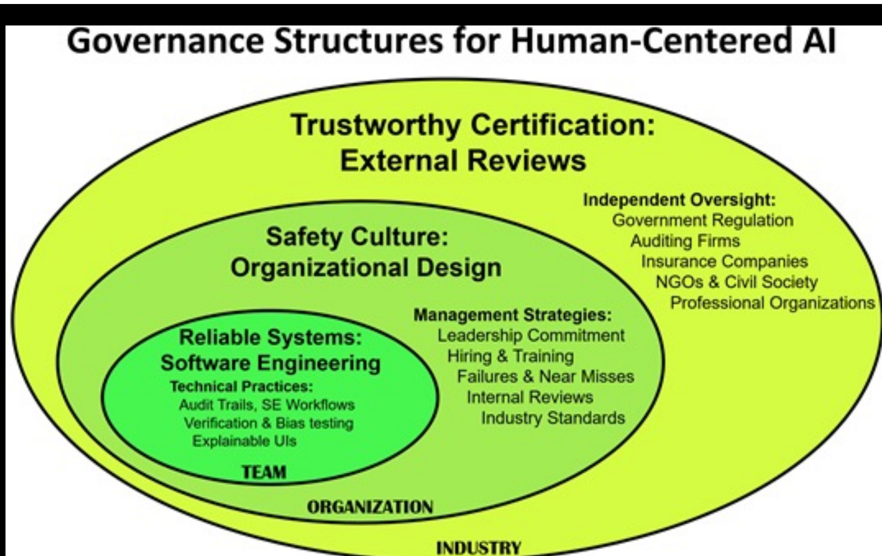
It's Complicated

Properties

- 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

EAI The Institute for Experiential AI
Northeastern University

Governance

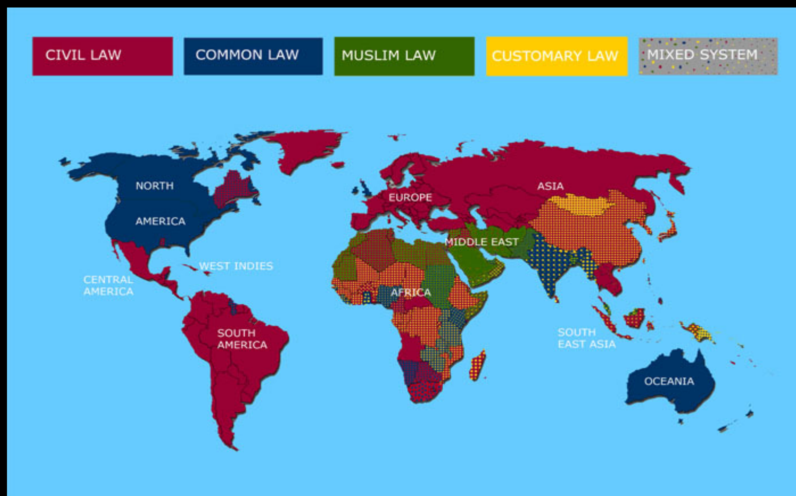


How to develop responsible software with the help of AI?

Ben Shneiderman: Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems, ACM Transactions on Interactive Intelligent Systems 10, 4 (October 2020).

EAI The Institute for Experiential AI
Northeastern University

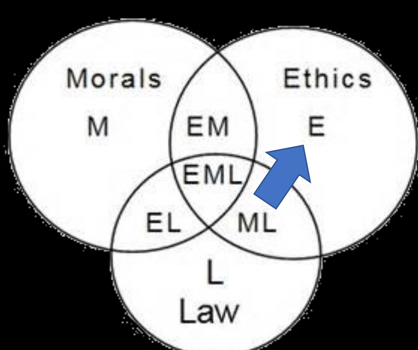
Legal and Ethical Colonialism



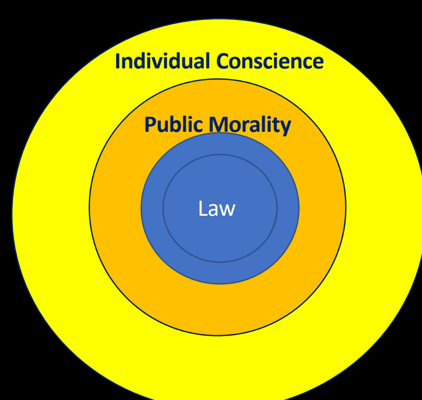
Technological Humanism

JuriGlobe - World Legal Systems Research Group, Univ. of Ottawa, Canada
http://www.juriglobe.ca/eng/rep-geo/cartes/monde.php

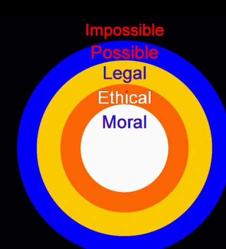
Religious Differences



Christian



Muslim



???

Cultural Differences

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 / / DONATE / NEWSLETTER / PSYCHE / SIGN IN

aeon

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

Abeba Birhane

7 April 2017

"Humanity" in Bantu languages

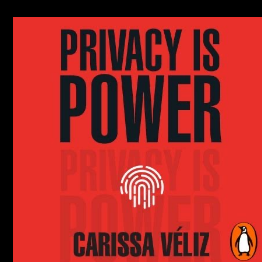
Language	Word	Countries
Chewa	umunthu	Malawi, Zambia
Zulu and Xhosa	ubuntu	South Africa
Sesotho	botho	South Africa
Shona	unhu, hunhu	Zimbabwe
Swahili	utu	Kenya, Tanzania
Meru	munto[a]	Kenya
Kikuyu	umundu[a]	Kenya
Heroro	omundu	Namibia
Tswana	muthu	Botswana
Kongo	gimuntu	Angola
Tonga	vumuntu	Mozambique

EAI The Institute for Experiential AI
Northeastern University

Legal Issues

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



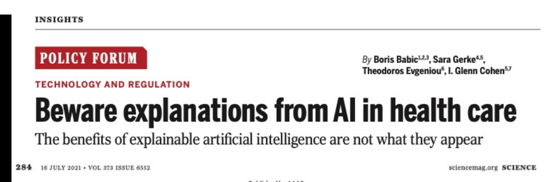
EAI The Institute for Experiential AI
Northeastern University

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**.

The Institute for Experiential AI
Northeastern University

What this Means?


INSIGHTS

POLICY FORUM

TECHNOLOGY AND REGULATION

By Boris Babic[1,2,3], Sara Gerke[4,5], Theodoros Evgeniou[6], I. Glenn Cohen[5,7]

Beware explanations from AI in health care

The benefits of explainable artificial intelligence are not what they appear

284 16 JULY 2021 • VOL 373 ISSUE 6552 sciencemag.org SCIENCE

Published by AAAS

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 for transparency
  - 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.

The Institute for Experiential AI
Northeastern University

Legal Issues

GDPR in Action

- Competence
- Consent
- Proportionality

- One Size Fits All
  - All human rights, domains, sizes, etc.
- Technological solutionism vs normative solutionism
  - [Jaume-Palasi, personal communication]

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

Feb 28, 2020 | Luana Pascu

The Institute for Experiential AI
Northeastern University

Legal Issues

Accountability

- Who is responsible?

future tense

Uber's Self-Driving Car Killed Someone. Why Isn't Uber Being Charged?

BY JESSE HALFON OCT 20, 2020 • 9:00 AM

Uber reaches settlement with family of woman killed by self-driving car

The family of Elaine Herzberg, 49, killed by a self-driving Uber vehicle in Arizona reached a settlement with Uber Technologies Inc.

Uber self-driving car operator charged in pedestrian death

By Matt McFarland, CNN Business
Updated 11:09 AM ET, Fri September 18, 2020

Phoenix New Times SUPPORT US Phoenix's independent source of local news and culture ACCOUNT

| POLICE |

Was the Backup Driver in an Uber Autonomous Car Crash Wrongfully Charged?

RAY STERN | JULY 9, 2021 | 10:41AM

The Institute for Experiential AI
Northeastern University

Regulation

Legal Issues

THE UNITED STATES DEPARTMENT of JUSTICE

ABOUT | OUR AGENCY | TOPICS | NEWS | RESOURCES | CAREERS

FOR IMMEDIATE RELEASE — Tuesday, October 20, 2020

Justice Department Sues Monopolist Google For Violating Antitrust Laws

Department Files Complaint Against Google to Restore Competition in Search and Search Advertising Markets

Office of Public Affairs

FEDERAL TRADE COMMISSION
PROTECTING AMERICA'S CONSUMERS

ABOUT THE FTC | NEWS & EVENTS | ENFORCEMENT | POLICY | TIPS & AD

Home » News & Events » Press Releases » FTC Sues Facebook for Illegal Monopolization

FTC Sues Facebook for Illegal Monopolization

December 9, 2020

Agency challenges Facebook's multi-year course of unlawful conduct

The New York Times

ON TECH

The Big Deal in Amazon's Antitrust Case

The claim that Amazon is crushing competition is both novel and railroad baron-style old-school.

By Shira Ovide

Published May 25, 2021 Updated May 26, 2021

- Regulate sectors or the use of specific technology?
- 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, amplified in eCommerce)
- Fair markets could be better revenue wise
  - Fairness trade-offs [Mehrotra et al., 2018; Baeza-Yates & Delnevo, to appear]

EAI The Institute for Experiential AI
Northeastern University

US Future Regulation?

Legal Issues

- **Algorithmic Accountability Act** (2019): The bill was introduced by Senators Cory Booker (D-NJ), Ron Wyden (D-OR), and Representative Yvette Clarke (D-NY). According to **Senator Wyden**, the bill would have required "companies to study the algorithms they use, identify bias in these systems and fix any discrimination or bias they find."
- **Consumer Online Privacy Rights Act** (2019): The bill, sponsored by Senator Maria Cantwell (D-WA), would have established new requirements for companies that use algorithmic decision-making to process data.
- **Justice in Policing Act** (2020): The bill was sponsored by then-Senator Kamala Harris (D-CA), Senator Cory Booker (D-NJ), and Representatives Karen Bass (D-CA) and Jerrold Nadler (D-NY). It would have been the first federal restriction on facial recognition technology.
- **Facial Recognition and Biometric Technology Moratorium Act** (2020): Sponsored by Senator Edward Markey (D-MA) and Jeff Merkley (D-OR), along with Representatives Pramila Jayapal (D-WA) and Ayanna Pressley (D-MA). The bill would have established a five-year moratorium on police 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

CATO INSTITUTE

ABOUT | EXPERTS | EVENTS | PUBLICATIONS | BLOG | DO

Constitution and Law | Economics | Politics

FEBRUARY 5, 2021 3:37PM

Algorithmic Bias Under the Biden Administration

CNBC | MARKETS | BUSINESS | INVESTING | TECH | POLITICS | CNBC TV | WATCHLIST

TECH

Lawmakers unveil major bipartisan antitrust reforms that could reshape Amazon, Apple, Facebook and Google

PUBLISHED FRI, JUN 11 2021-2:40 PM EDT | UPDATED FRI, JUN 11 2021-4:29 PM EDT

Lauren Feiner
@LAUREN_FEINER

SHARE

Northeastern University

Legal Issues

EU Proposal (April 21, 2021)

- Forbidden uses
- High and low-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

EAI The Institute for Experiential AI
Northeastern University

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
(b) 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[62] and punishable 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.

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.

Problem:

Risk is a continuous variable

The Institute Northeastern

Harvard Business Review

The Dangers of Categorical Thinking

We're hardwired to sort information into buckets—and that can hamper our ability to make good decisions. by Bart de Langhe and Philip Fernbach

From the Magazine (September–October 2019)

Legal Issues

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

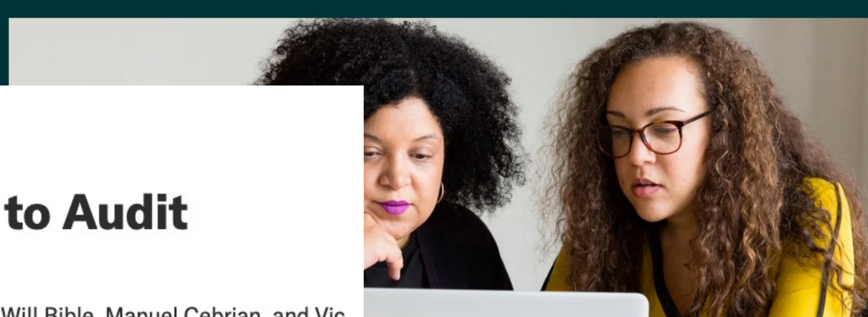
Khari Johnson @kharijohnson September 28, 2020 11:41 AM

Legal Issues

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

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

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 construct... have better incentive to participate in thi... watchdog groups can be better prepared t...

ACM Reference Format:
Christo Wilson, Avijit Ghosh, Shan Jiang, A... Janelle Szary, Kelly Trindel, and Frida Polli. ...ing Fair Algorithms: A Case Study in Candi... *ence on Fairness, Accountability, and Transpar... 10, 2021, Virtual Event, Canada.* ACM, New ... https://doi.org/10.1145/3442188.3445928

Legal Issues

FaccT 2021

Auditing Algorithms @ Northeastern

Bloomberg Law

The United States Law Week

Bradford Newman
Baker McKenzie

Jan. 15, 2021, 10:01 AM

Using AI to Make Hiring Decisions? Prepare for EEOC Scrutiny

Bad (Human) Practices

Cognitive Biases

- 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
- Discrete categories and arbitrary thresholds for continuous variables
- 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] [Andrew Ng, Unbiggen AI, IEEE Spectrum, 2022]
- Design and Implementation
  - Do systems reflect the characteristics of the designers?
  - Do systems reflect the characteristics of the coders? [Silberzahn et al., COS, Univ. of Virginia, 2015] [Johansen et al., Norway, 2020]
- Evaluation
  - 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

Big Data is Easy! Small Data is not!

- **Example:** Dyslexia screening through web game [Rello et al., 2020]
- Unbalanced data (less than 10% of people have it)

Language	Data	Accuracy
Spanish	4,000	81%
English	1,500	90%

- Cost of false negatives (not detecting dyslexia) is much higher than false positives (going to a specialist)
- Can we do it before they learn how to read & write? [Rauschenberger at al., 2018]

Solutions

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
- 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
- User Experience
  - Make sure that the user is **aware** of the biases all the time
  - Give more control to the user
- Evaluation & Deployment
  - Do not fool yourself!
  - Error & sensibility analysis (*e.g.*, synthetic data if possible)
  - Algorithms registration / External Auditing / Documentation

Solutions

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 an Ethics Board and enforce a Code of Ethics
- Improve Explainability
- 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

EAI The Institute for Experiential AI
Northeastern University

Solutions

Key Ethical Questions Before Using AI

Competence
- Political
- Scientific
- Technical

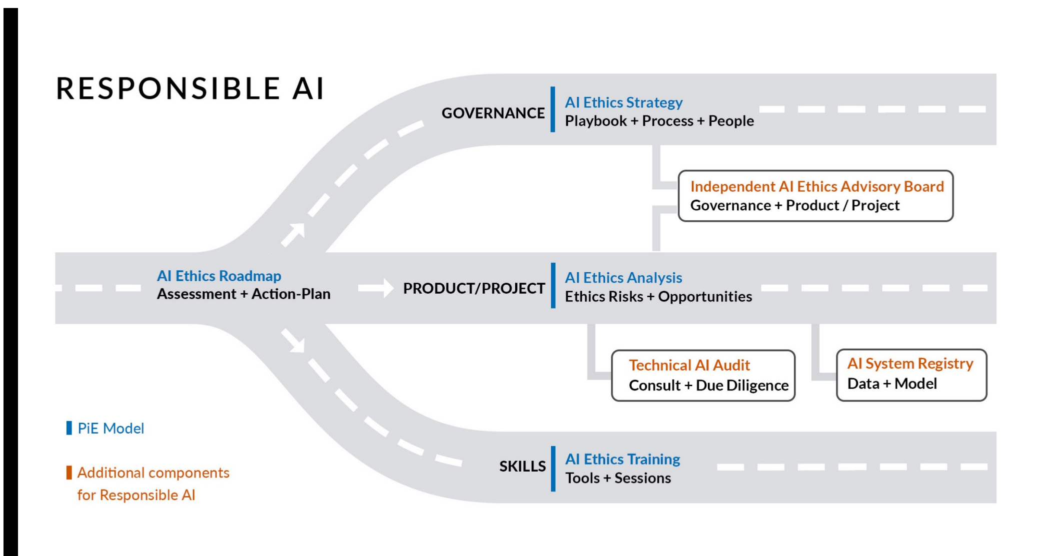
Proportionality
- Data really needed
- Data kept
- Solution

Responsibility/Awareness
- Bias
- Data Protection
- Consent/Legal Basis
- Security/Safety
- Transparency/Accountability

The Institute for Experiential AI
Northeastern University

Baeza-Yates, Forbes Technology Forum, 2021

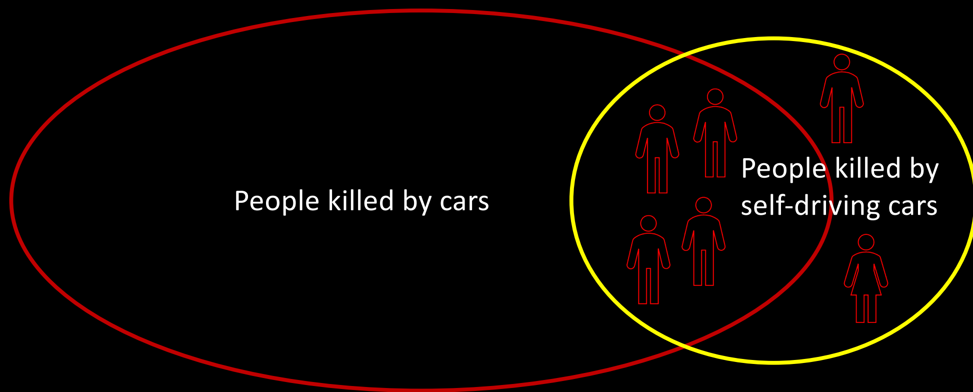
Solutions



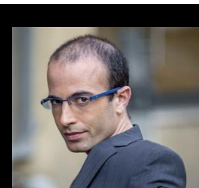
The Institute for Experiential AI
Northeastern University

AI is not Trustworthy (yet)

Ethical Risk Assessments



Dark Future?



- Infotech + Biotech [Harari, 2018]
- Free Will is an Illusion
- Humans can be hacked

Just Easy Parts (Politics?)
Emotions are predictions
[Feldman Barrett, 2017]

- Loss of Jobs

Leverage AI
More Literature & Art

- Loss of Skills

When they are better than humans

- Integrated Complex Machine Network versus Individuals
- Authority Switches to Algorithms and Owners of Our Data
- Even More Inequality
- No Sense of Purpose
- Irrelevance

Epilogue

My Future

- BIG PICTURE: Integration
- No Privacy or Complete Privacy?
- Compulsory External Ethics Committees
- Software Insurance (my worst nightmare)
- Remote Knowledge Workers: AI Teachers
- Augmented Humanity?

"Either democracy will successfully reinvent itself in a radically new form or humanity will live in 'digital dictatorships'", Harari 2018

- Still, technological change is overall good!
- Philippines 2017, China 2020?
- But, are we evolving towards Solaria?
  [*The Naked Sun*, Asimov, 1957]
- If there are nice aliens out there, please come soon!
  - See "Arrival" (2016)

EAI The Institute for Experiential AI
Northeastern University

Epilogue

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/ethics**
- Who profits/suffers technology, transhumanism vs. humanism
- Ethics is **complicated**, do not underestimate it!
- **Plenty** of open research problems! (in **small data** even more!)

SCIENCE & TECHNOLOGY

Can AI algorithms ever be ethical?

The perils of cyberspace and social media

4 FEBRUARY 2021, HAZEL HENDERSON

AI and Ethics (2021) 1:21–25
https://doi.org/10.1007/s43681-020-00013-4

OPINION PAPER

You cannot have AI ethics without ethics

Dave Lauer[1]

Received: 2 September 2020 / Revised: 2 September 2020 / Accepted: 4 September 2020 / Published online: 6 October 2020
© Springer Nature Switzerland AG 2020

EAI The Institute for Experiential AI
Northeastern University

Practice

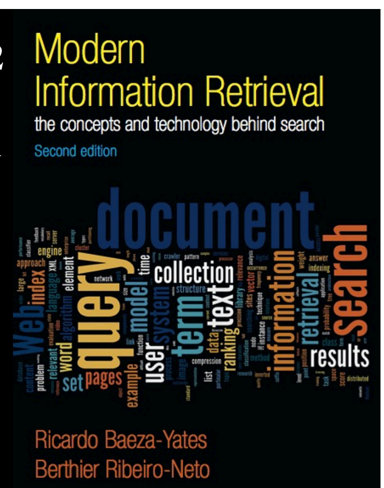
Exercise

- Go to incidentdatabase.ai
- Which fraction of cases are discrimination?
- Choose the top-5 worst examples justifying your rationale

- Irresponsible AI Atlas:
- https://ai.northeastern.edu/ai-research/rai/

Questions?

ASIST 2012
Book of the
Year Award
(Biased Ad)



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

Contact: rbaeza@acm.org
www.baeza.cl
@polarbeaRBY

Biased Questions?

EAI The Institute for Experiential AI
Northeastern University