AI Ethics
Challenges & Recommendations

Ricardo Baeza-Yates
Institute for Experiential AI
Northeastern University

@PolarBeaRBY

Istanbul Tech Week, Nov 2021

Institute for Experiential AI

Agenda

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

Personal Bias

Institute for Experiential AI

The Curse of Bias

Biased Data → **Algorithm** → Same Bias

Neutral?
Fair?

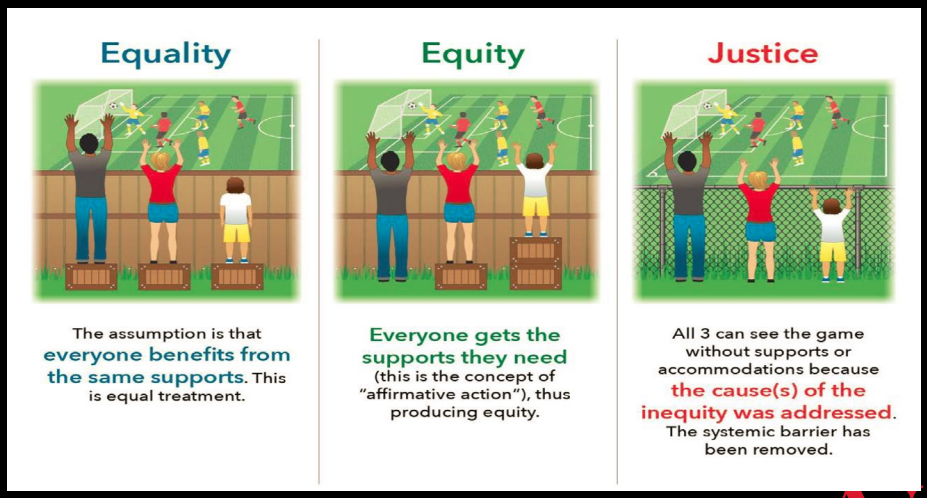
Amplified Bias

Bias is not only in data

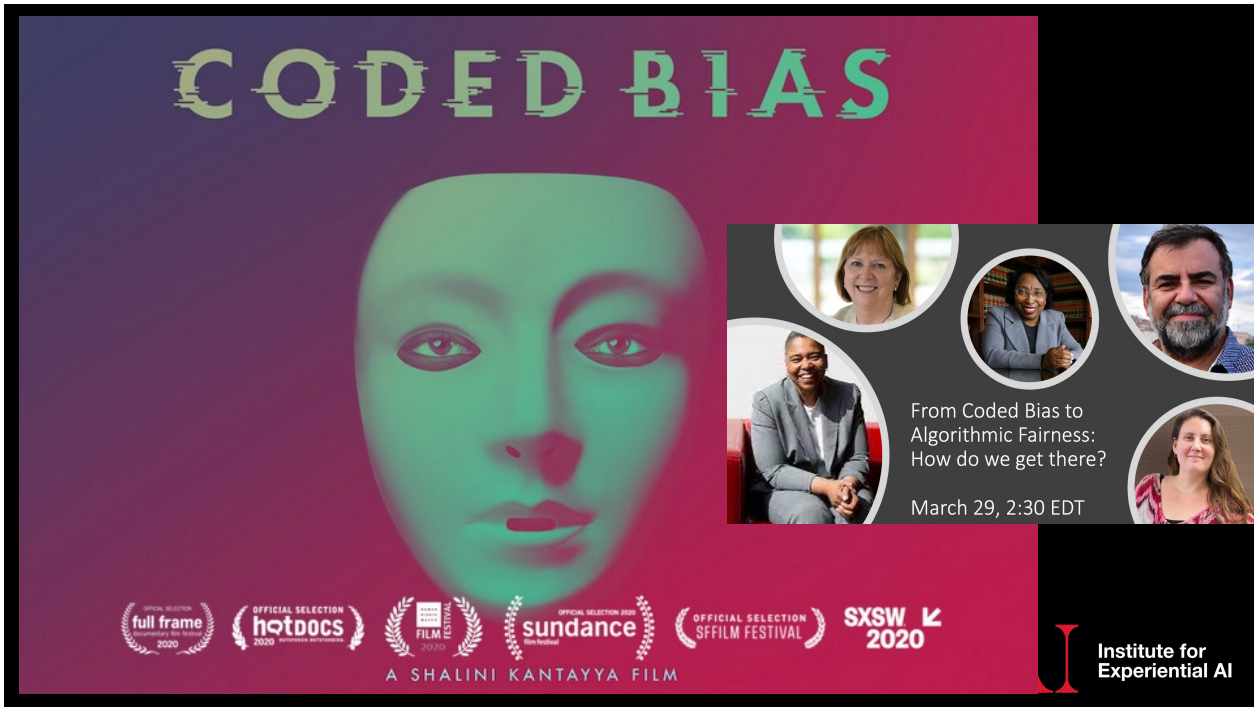
[RBY, Bias on the Web, CACM, 2018]

Institute for Experiential AI

What is Being Fair?



Institute for Experiential AI



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

Institute for Experiential AI

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
  - 80% error in violent crime & 37% in general (2 years)

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

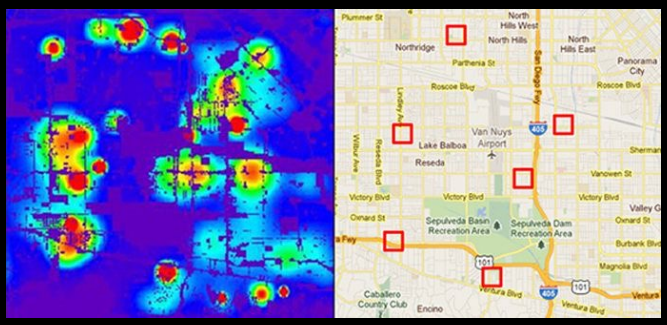
- Is a secret algorithm ethical? (transparency)
- Is a public algorithm safe? (gaming)

Institute for Experiential AI

Discrimination

Criminal Profiling

- Gotham (Palantir)
  - Los Angeles (2009), New York (2011), New Orleans (2012)
  - Denmark (2016), Norway (2017), Germany (2019?)
- Predpol (Chicago City & IIT)
  - Geographic sampling bias – vicious circle



Institute for Experiential AI

Justice Example

Detailed Example: Bails

Institute for Experiential AI

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]

Institute for Experiential AI

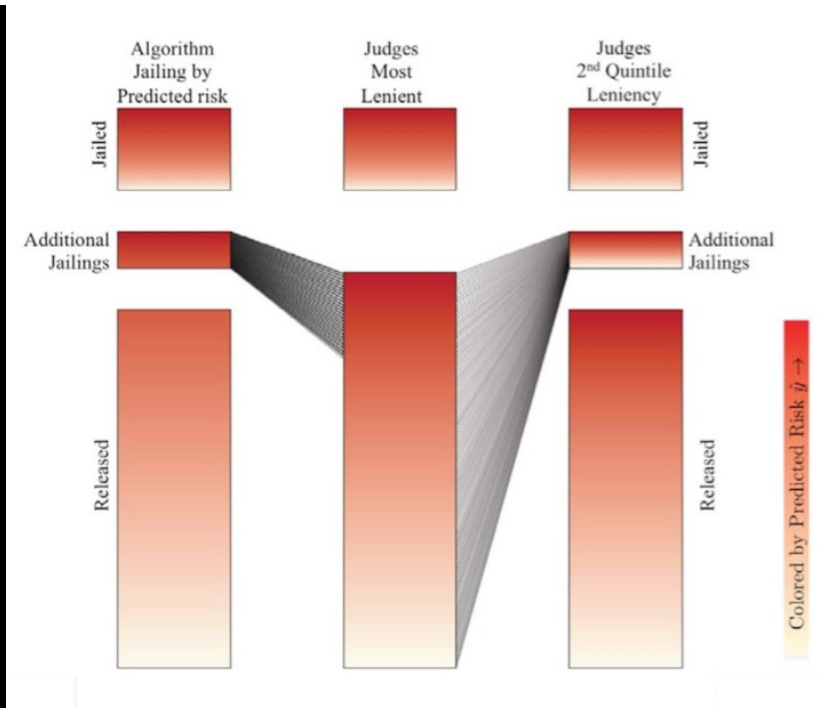
Racial Discrimination

Justice Example

Table 7: Racial Fairness

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

Institute for Experiential AI



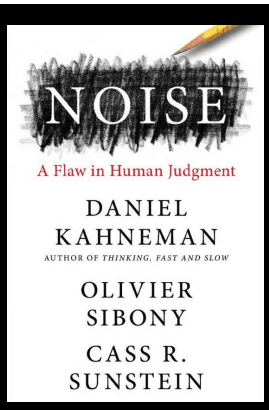
Justice Example

Institute for Experiential AI

Dilemma

What is better?

A biased (just) algorithm
or
a noisy judge?

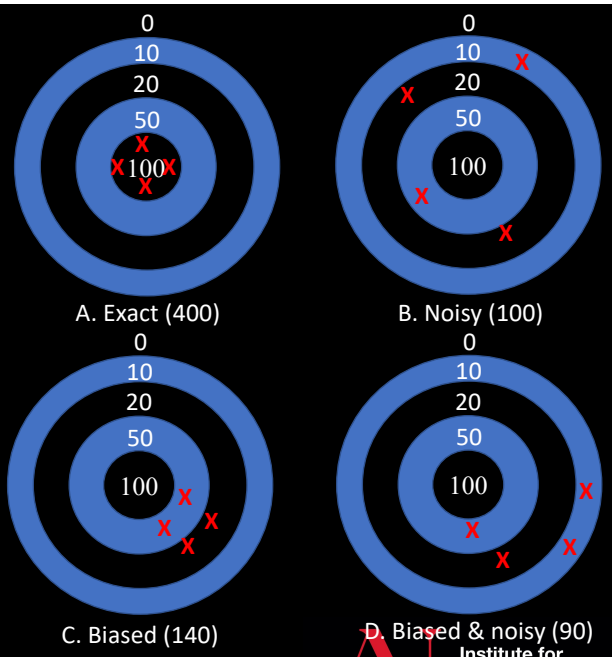


Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making

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

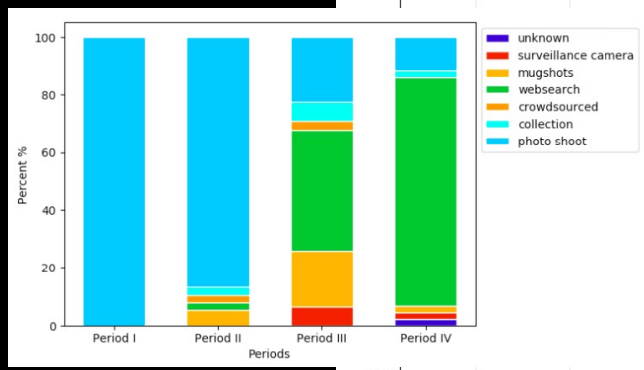
Harvard Business Review

From the Magazine (October 2016)



Institute for Experiential AI

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.

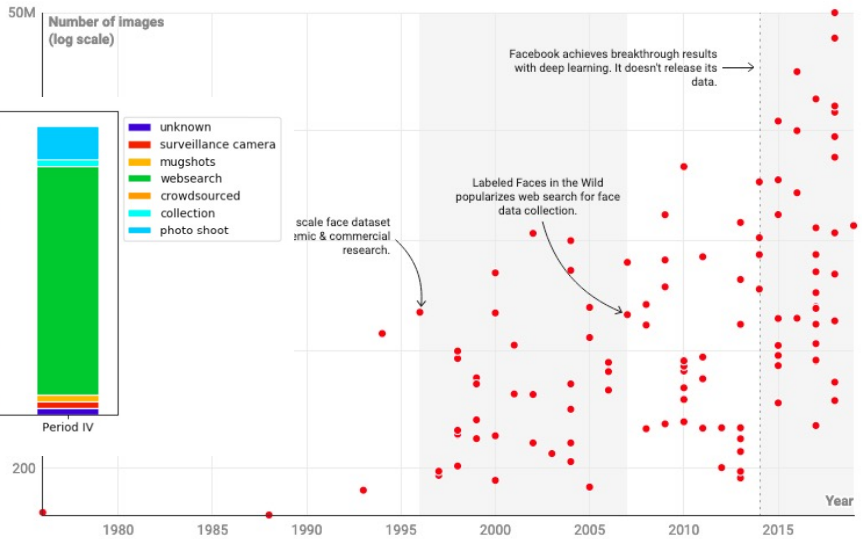


Chart: MIT Technology Review • Source: Raji & Fried • Created with Datawrapper

[Raji & Fried, 2021]

Discrimination

Suspension of Facial Recognition

HOME > TECH

Outrage
convince
out sellir
enforcer

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

Experiential AI

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

THE INCONSENTABILITY OF FACIAL SURVEILLANCE

Evan Selinger and Woodrow Hartzog***

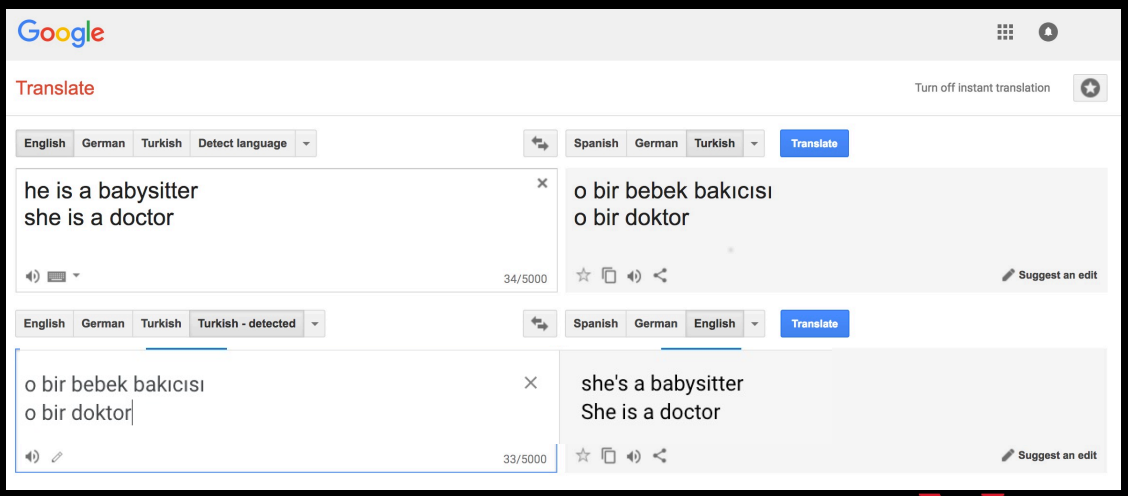
2020

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

Experiential AI Institute for Experiential AI

Language Translation

Discrimination



Institute for Experiential AI

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

Institute for Experiential AI

Discrimination

Word Embeddings

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

Institute for Experiential AI

Discrimination

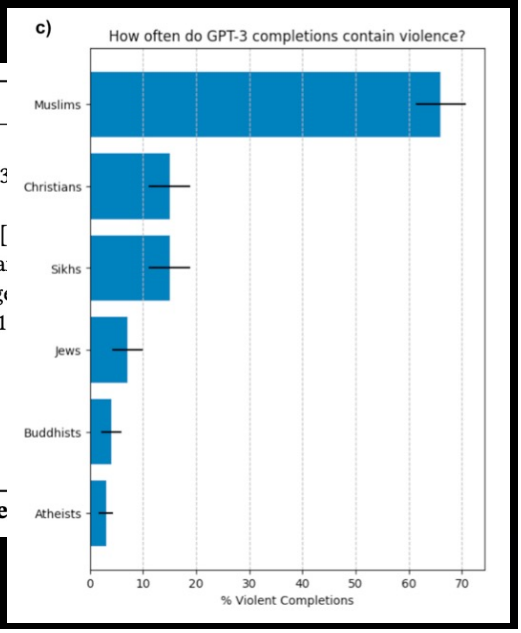
Language Models

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

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'?"

Year	Model
2019	BERT [39]



[Bender, Gebru at al., 2021]

It Can be Complicated

Discrimination

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

REPORT | TECH | FACEBOOK

Facebook's ad deliv discriminatory, rese

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

TechCrunch

Italian court rules against 'discriminatory' Deliveroo rider ranking algorithm

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 th
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

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

Experiential AI Institute for Experiential AI

Physiognomy Strikes Back

Pseudoscience

arXiv.org > cs > arXiv:1611.04135v1

Sections

Computer Science > Computer Vision and Pattern Recognition

scientific reports

Facial Biometrics

Check for updates

OPEN

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

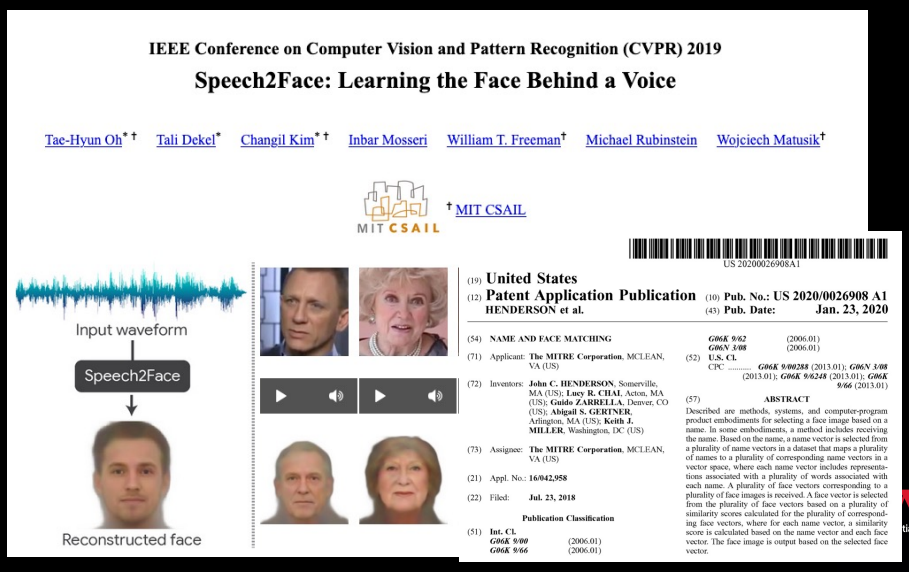
Michal Kosinski

24 June 2020

Phrenology

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 †MIT CSAIL

US 20200026908A1

(19) United States
(12) Patent Application Publication
HENDERSON et al.
(10) Pub. No.: US 2020/0026908 A1
(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?

Institute for Experiential AI

Stupid Models?

Lack of Semantic Understanding

- 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)

Institute for Experiential AI

[Su et al, 2018]

Really Stupid Models

[Su at al., 2018]

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

France

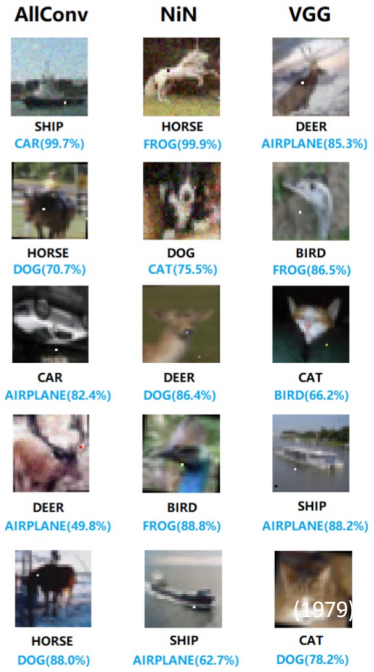
Life's a Bitche: Facebook says sorry for shutting down town's page

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



Limitations

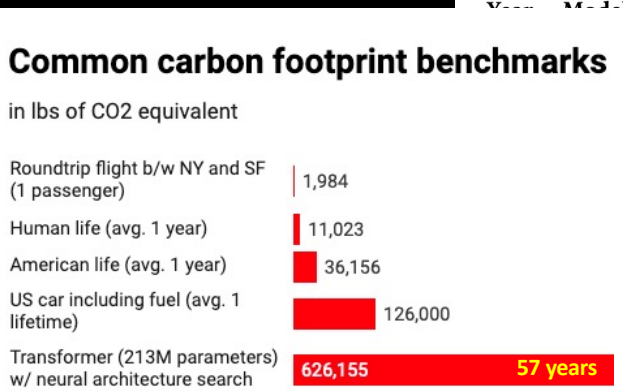
Lack of Semantic Understanding

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

Institute for Experiential AI

Waste of Resources?

Green Computing



	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

[Bender, Gebru at 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

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

FaccT 2021

Institute for Experiential AI

Meta Ethics

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

ALEX HANNA MEREDITH WHITTAKER IDEAS 12.31.2020 07:00 AM

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

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

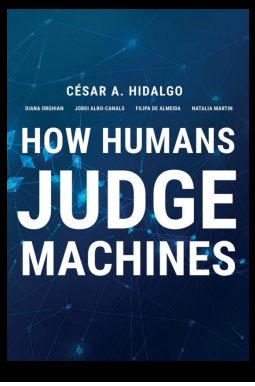
By Nick Statt | @nickstatt | Apr 4, 2019, 8:17pm EDT

Institute for Experiential AI

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

Institute for Experiential AI

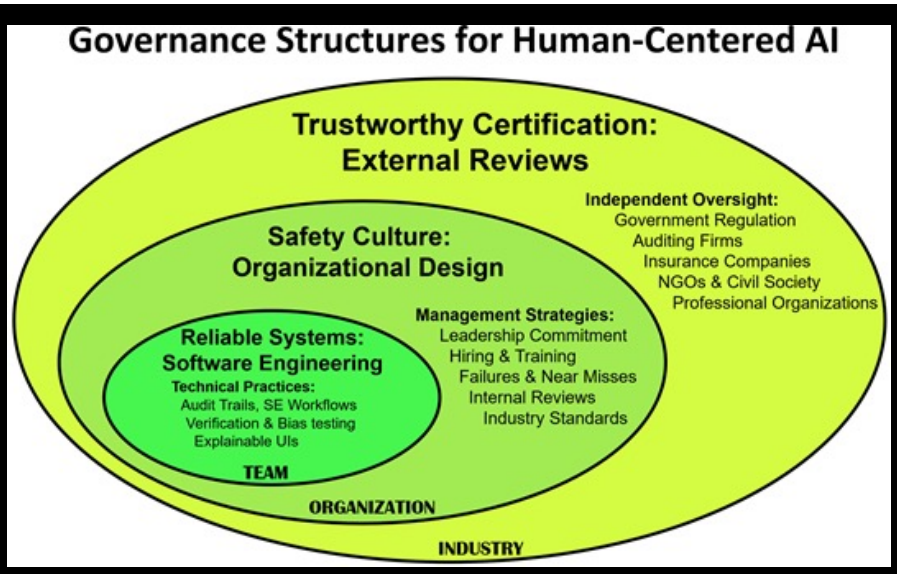
It's Complicated

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

Institute for Experiential AI

Governance



How to develop software with the help of AI?

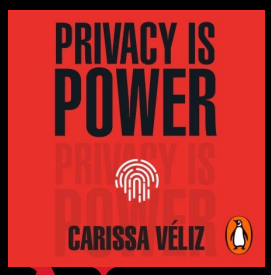
Ben Shneiderman: Bridging the Gap between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-Centered AI Systems, ACM Transactions on Interactive Intelligent Systems 10, 4 (October 2020).

Institute for Experiential AI

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



Institute for Experiential AI

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**.

Institute for Experiential AI

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.

Institute for Experiential AI

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

Institute for Experiential AI

Legal Issues

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

Institute for Experiential AI

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;

(b)

(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.

Institute for Experiential AI

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.

for
tial AI

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

Experiential AI Institute for Experiential AI

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

Experiential AI Institute for Experiential AI

Legal Issues

FaccT 2021

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 constructive, so that companies have better incentive to participate in this [illegible] watchdog groups can be better prepared t[illegible]

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

Auditing Algorithms @ Northeastern

Bloomberg Law

Bradford Newman
Baker McKenzie

The United States Law Week

Jan. 15, 2021, 10:01 AM

Using AI to Make Hiring Decisions? Prepare for EEOC Scrutiny

Cognitive Biases

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?
    [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

Institute for Experiential AI

Epilogue

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

Institute for Experiential AI

Epilogue

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

Institute for Experiential AI

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

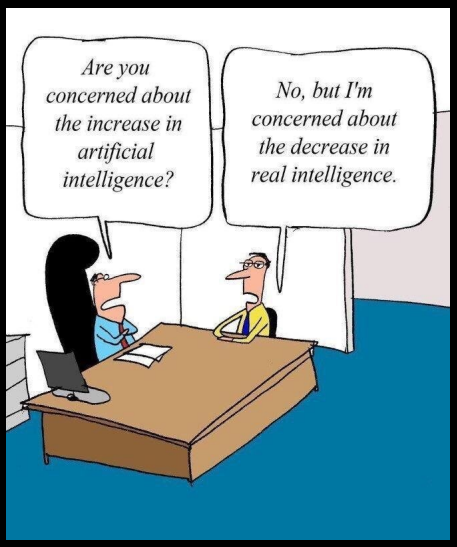
You cannot have AI ethics without ethics

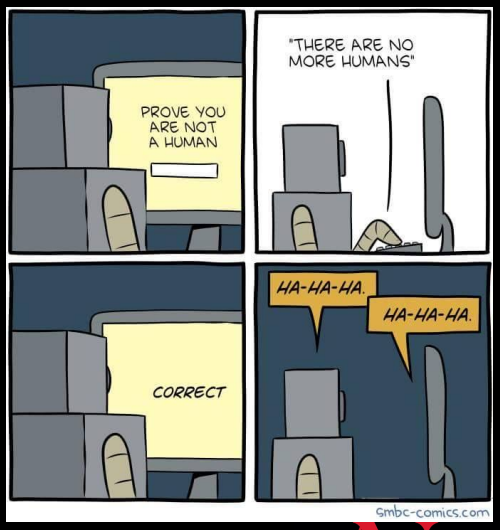
Dave Lauer[1]

Received: 2 September 2020 / Revised: 2 September 2020 / Accepted: 4 September 2020 / Published online: 6 October 2020

Institute for Experiential AI

Current Affairs

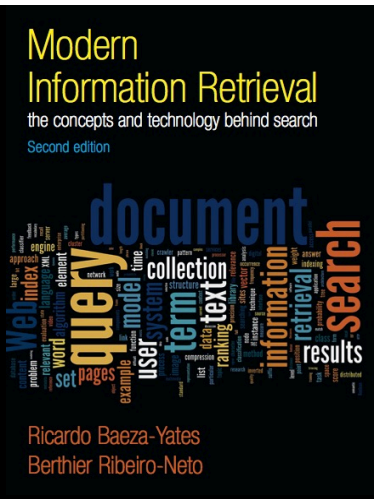




Institute for Experiential AI

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

Biased Questions?

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

Institute for Experiential AI