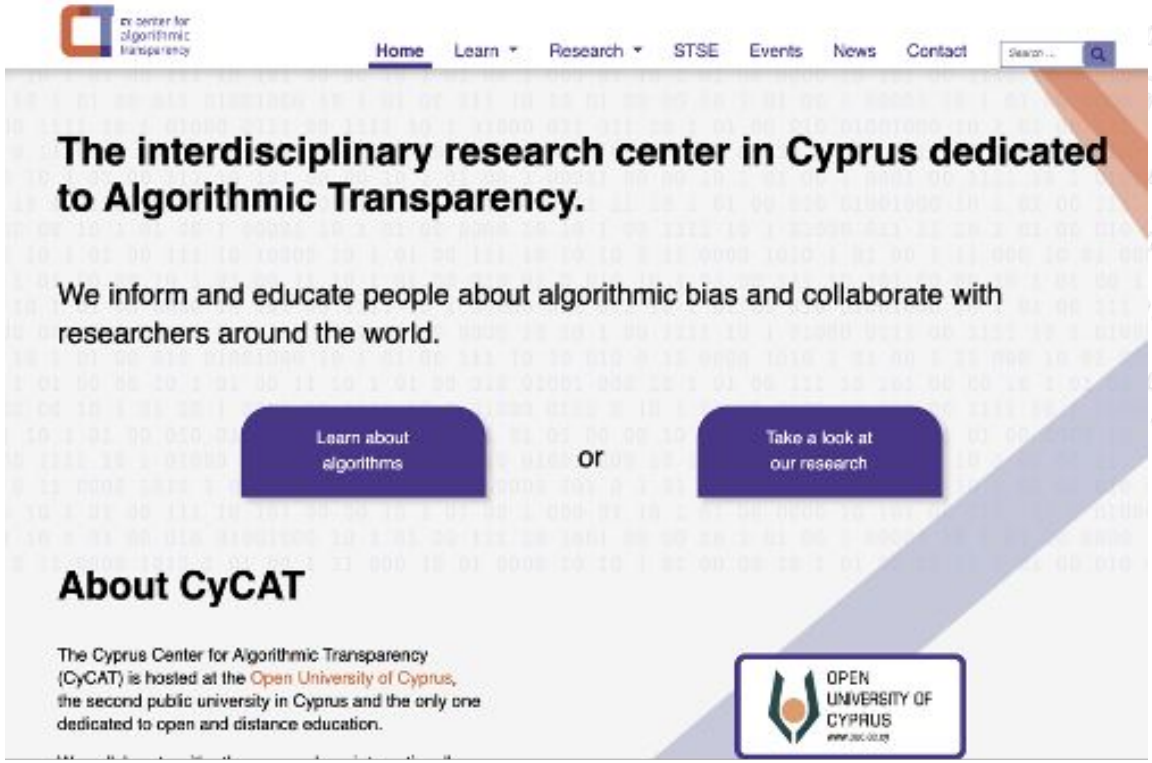
FATE: Fairness, Accountability, Transparency and Ethics

An introduction for developers

Styliani Kleanthous Loizou, Ph.D.
Kalia Orphanou, Ph.D.
Jahna Otterbacher, Ph.D.

cy. center for algorithmic transparency

ΑΝΟΙΚΤΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ
www.ouc.ac.cy



אוניברסיטת חיפה
University of Haifa
جامعة حيفا

UNIVERSITY OF TRENTO - Italy

THE UNIVERSITY of EDINBURGH

The University Of Sheffield.

Developer Seminar Objectives

In this seminar participants will:

- Become aware of FATE issues in the development of (algorithmic) process/systems
- Learn core FATE concepts related to software development
- Develop appreciation for the role that developers play in mitigating algorithmic bias and in promoting ethical practices
- Experiment for techniques for auditing services / modules used in development

Pre-seminar questionnaire

https://forms.gle/XM68VeYRJZ9agi5RA

Introduction to FATE

Fairness, Accountability, Transparency and Ethics

Nearly Half Of All 'AI Startups' Are Cashing In On Hype

Parmy Olson Former Staff
AI
AI, robotics and the digital transformation of European business.

Some 40% of firms across Europe classified as being "AI startups" showed no evidence that they used ... [+] GETTY IMAGES/ISTOCKPHOTO

It can seem that hardly a day goes by that a new technology startup hasn't raised investor cash on the hope that it uses artificial intelligence, or AI, as a key part of its business. Now however, a new report makes the surprising claim that 40% of European firms that are classified as an "AI startup" don't exploit the field of study in any material way for their business.

Out of 2,830 startups in Europe that were classified as being AI companies, only 1,580 accurately fit that description, according to the eye-opening stat on page 99 of a new report from MMC, a London-based venture capital firm. In many cases the label,

Startups labelled as being in AI attract 15% to 50% more funding than other technology firms.

One in 12 startups use AI as part of their products or services, up from one in 50 about six years ago, according to the survey. Meanwhile some 12% of large companies are using AI applications in their business, up from 4% in just the past year.

The most popular uses of AI were chatbots, followed by process automation tools that replace simple administrative tasks like processing an insurance claim and fraud detection.

https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/

Democratizing AI

For every person and every organization

Microsoft
News Center

As we think about the future of technology, it resides in the notion of intelligence. At Microsoft, we have an approach that's both ambitious and broad, an approach that seeks to democratize Artificial Intelligence (AI), to take it from the ivory towers and make it accessible for all.

And as we consider the future, it's often instructive to look to the information. With the advent of the printing press in the 1400s w event around access that made it possible for humans everywhe

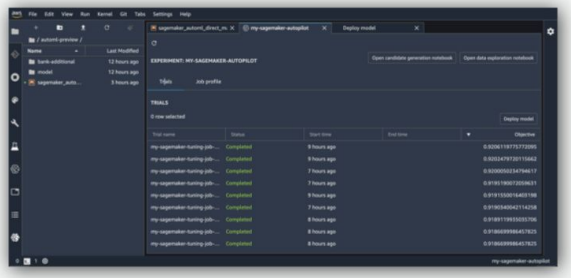
H2O.ai is leading the movement to democratize AI for Everyone

Our approach is to be open, transparent and push the bleeding edge. Our philosophy is to create a culture of makers: community, customers, partners, entrepreneurs and our own "makers gonna make". Our vision is to democratize AI for everyone. Not just a select

Automatically build, train, and tune models with full visibility and control, using Amazon SageMaker Autopilot

Amazon SageMaker Autopilot is the industry's first automated machine learning capability that gives you complete control and visibility into your ML models. Typical approaches to automated machine learning do not give you the insights into the data used in creating the model or the logic that went into creating the model. As a result, even if the model is mediocre, there is no way to evolve it. Also, you don't have the flexibility to make trade-offs such as sacrificing some accuracy for lower latency predictions since typical automated ML solutions provide only one model to choose from.

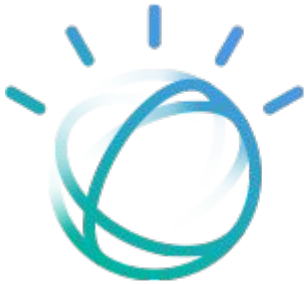
SageMaker Autopilot automatically inspects raw data, applies feature processors, picks the best set of algorithms, trains and tunes multiple models, tracks their performance, and then ranks the models based on performance, all with just a few clicks. The result is the best performing model that you can deploy at a fraction of the time normally required to train the



Automatically create machine learning models and pick the one that best suits your use case. For example, review the leaderboard to see how each option performs and pick the model that meets your model accuracy and latency

Microsoft
Face API

Google Cloud
Vision API



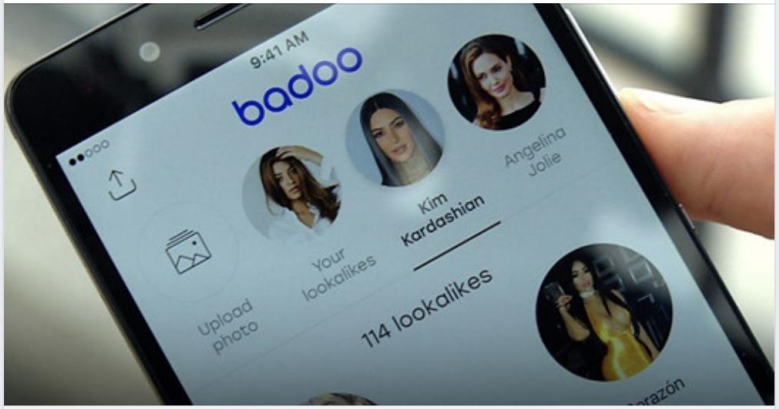
IBM
Watson™

clarifai

amazon
web services
Rekognition

Badoo
11 July 2017 ·

Have you tried our Lookalikes feature yet? 💜👌



BBC.CO.UK

Dating app tech finds celebrity lookalikes - BBC News

Popular dating app Badoo adds facial recognition technology to let...

554 241 comments 23 shares

Share

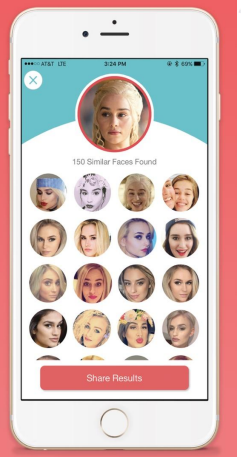
Dating.ai

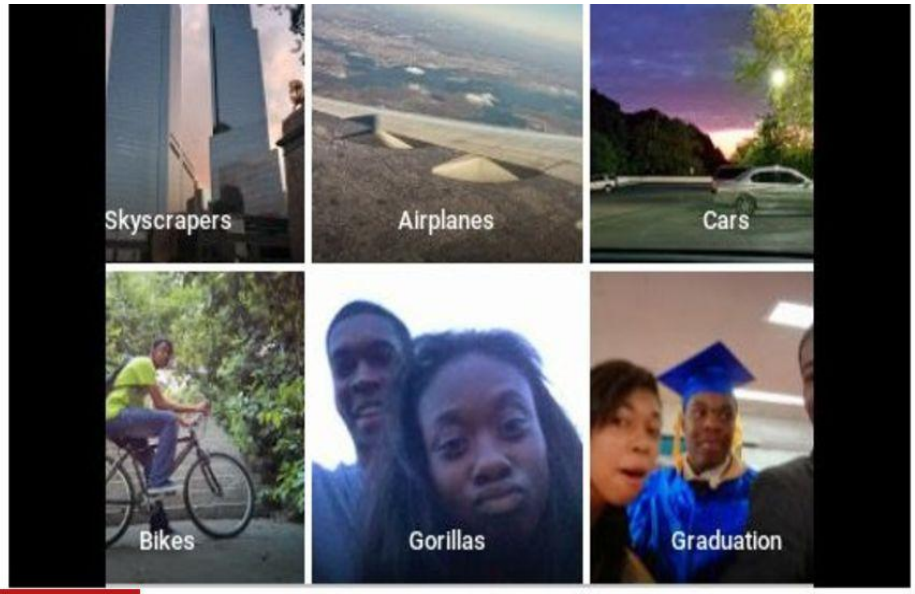
Search Dating Apps For Any Face.

Dating AI is the first dating app with Face Search. This powerful feature lets you instantly see the people you are really interested in meeting.

Download from the App Store

Download from the Play Store





BBC NEWS

diri noir avec banan @jackyalcine · Jun 29

Google Photos, y'all [redacted] My friend's not a gorilla.

Google's solution to accidental algorithmic racism: ban gorillas

Google's 'immediate action' over AI labelling of black people as gorillas was simply to block the word, along with chimpanzee and monkey, reports suggest

▲ A silverback high mountain gorilla, which you'll no longer be able to label satisfactorily on Google Photos. Photograph: Thomas Mukoya/Reuters

After Google was criticised in 2015 for an image-recognition algorithm that auto-tagged pictures of black people as "gorillas", the company promised "immediate action" to prevent any repetition of the error.

That action was simply to prevent Google Photos from ever labelling any image as a gorilla, chimpanzee, or monkey - even pictures of the primates themselves.

https://www.bbc.com/news/technology-33347866
https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours

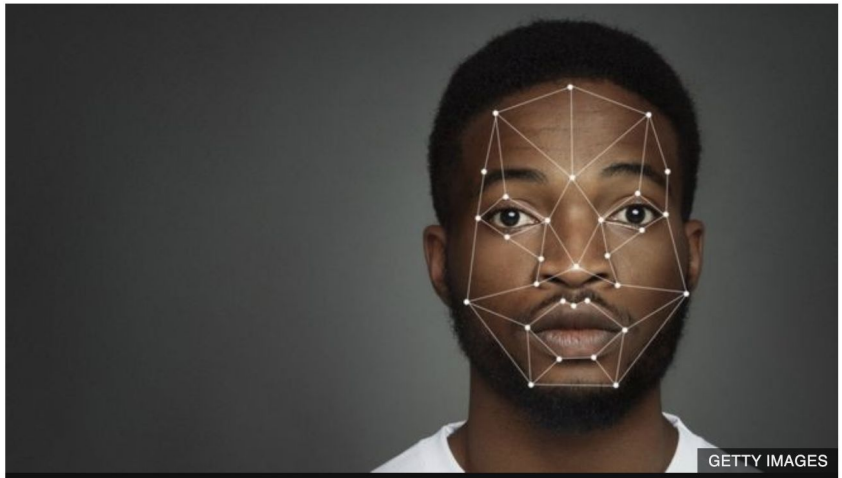
https://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/
https://www.bbc.com/news/technology-52978191

IBM abandons 'biased' facial recognition tech

9 June 2020

Share

George Floyd death

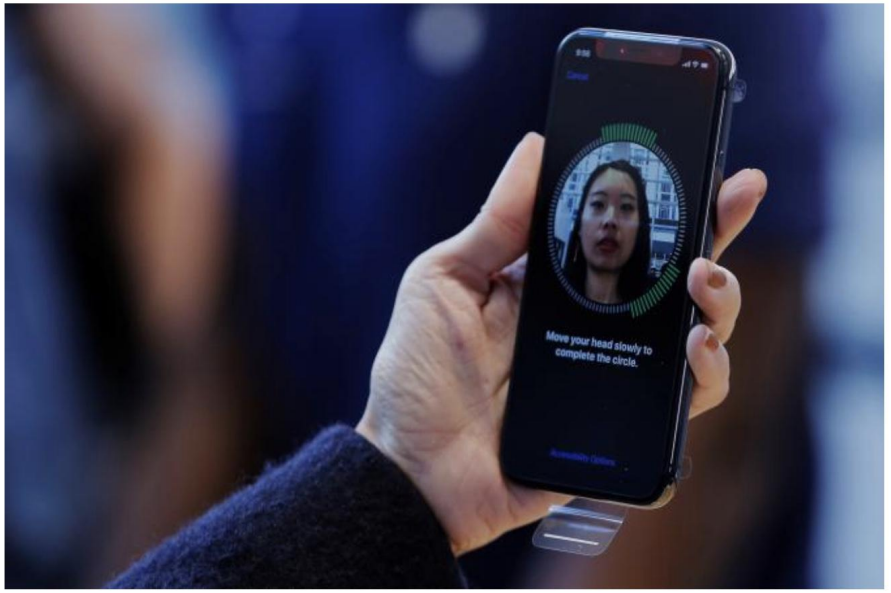


A US government study suggested facial recognition algorithms were less accurate at identifying African-American faces

Newsweek

IS THE IPHONE X RACIST? APPLE REFUNDS DEVICE THAT CAN'T TELL CHINESE PEOPLE APART, WOMAN CLAIMS

BY CHRISTINA ZHAO ON 12/18/17 AT 12:24 PM



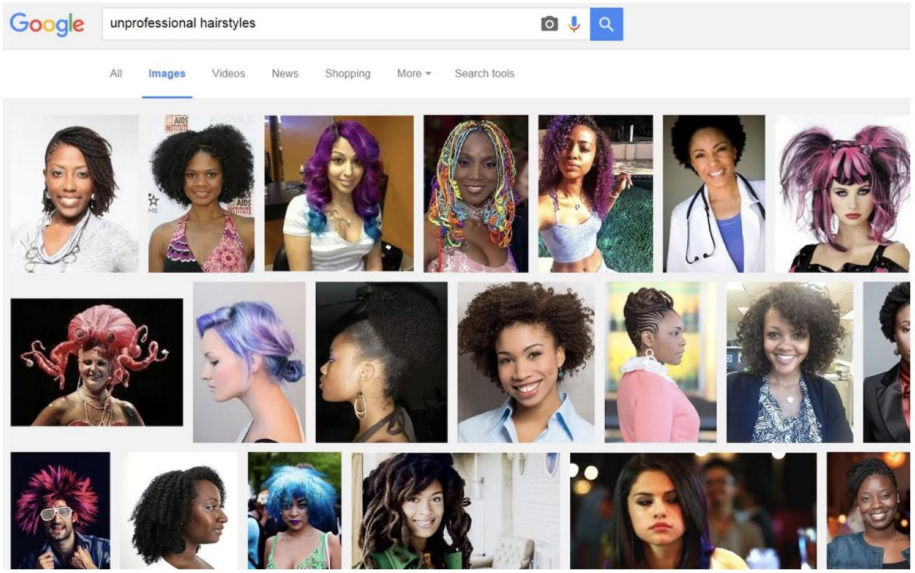
A woman sets up her facial recognition as she looks at her Apple iPhone X at an Apple store in New York, U.S., November 3. Last week a woman in China claimed that her iPhone X facial recognition could not tell her and her colleague apart.

Bias in Information Access?

› Technology Intelligence

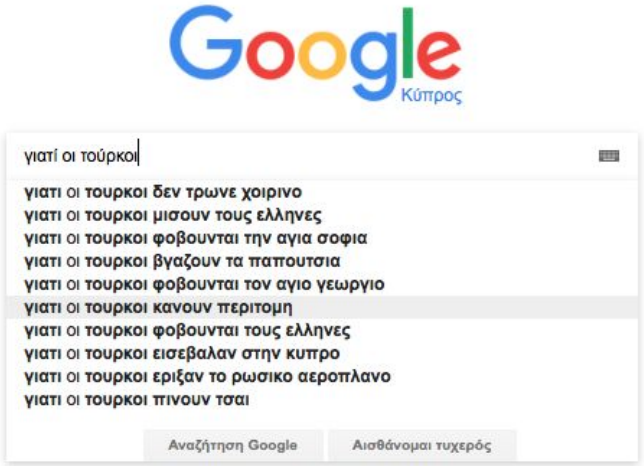
Google under fire over 'racist' image search results for 'unprofessional hair'

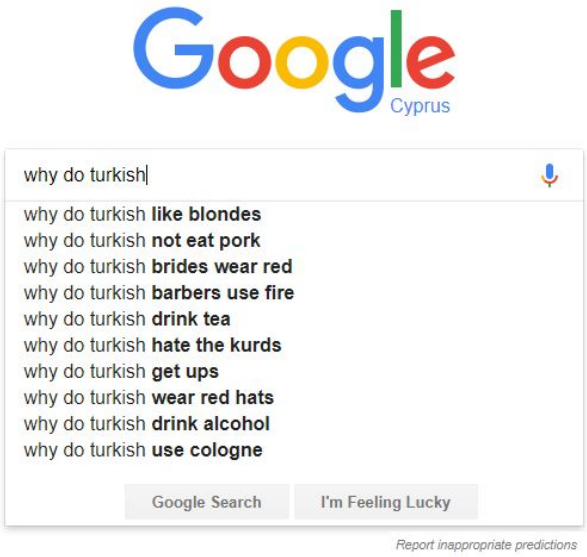
share



Google Image search results for 'unprofessional hair'

Bias in Information Access?



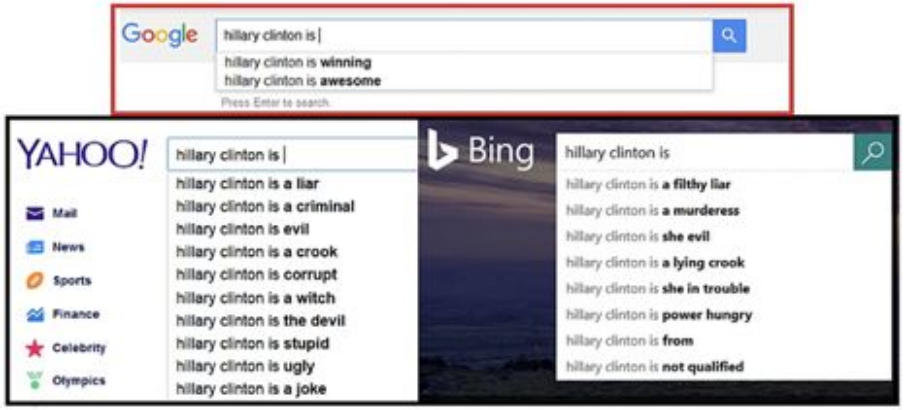


All Systems have a Slant

Bias in information system is not a new problem!

1. Results are slanted in *unfair discrimination* against particular persons or groups
2. That discrimination is *systematic*

[Friedman & Nissenbaum, 1996]



Response: Government / Regulators

EU: General Data Protection Regulation

- Is there a “right to an explanation”?
  - The right not to be subject to automated decision-making and safeguards enacted thereof (Article 22, Recital 71)
  - Notification duties of data controllers (Articles 13-14, Recitals 60-62)
  - The right to access (Article 15, Recital 63)

EU General Data Protection Regulation (GDPR)

The GDPR affects the routine use of machine learning algorithms:
Article 22 “Automated individual decision-making, including profiling.”

“any form of automated processing of personal data consisting of the use of personal data to evaluate certain aspects relating to a natural person.”

EU: GDPR

Just a few challenges…

- Vague language
  - “meaningful information/explanation”
  - “logic involved”
  - “significance”
  - “envisaged consequences”
- What kinds of “meaningful explanations”?
  - Global vs. local explanations
  - Explanation for whom?
  - Issues of algorithmic and digital literacy

EC: Trustworthy AI

Next Steps

Based on fundamental rights and ethical principles, the Guidelines list **seven key requirements** that AI systems should meet in order to be trustworthy:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and Data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

National AI Strategies

ΚΥΠΡΙΑΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ
ΜΕΤΑΦΟΡΩΝ, ΕΠΙΚΟΙΝΩΝΙΩΝ ΚΑΙ ΕΡΓΩΝ

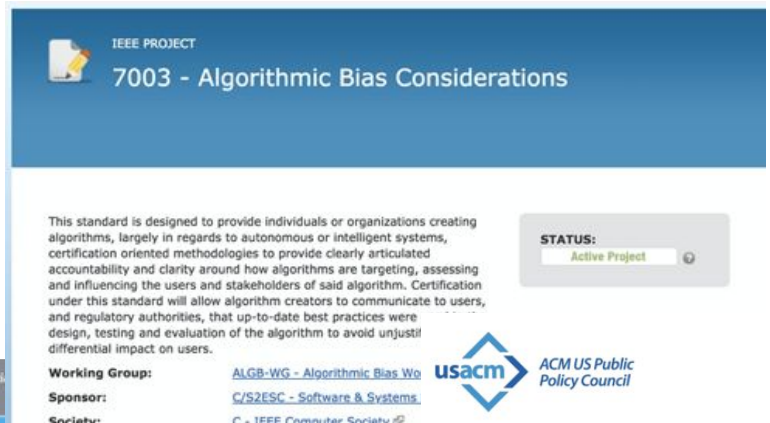
ΤΜΗΜΑ
ΗΛΕΚΤΡΟΝΙΚΩΝ ΕΠΙΚΟΙΝΩΝΙΩΝ
ΛΕΥΚΩΣΙΑ 2048

Τίτλος Έργου	:	**Εθνική Στρατηγική Τεχνητής Νοημοσύνης (ΤΝ): Δράσεις για την Αξιοποίηση και Ανάπτυξη της ΤΝ στην Κύπρο**
Υπηρεσία	:	**Τμήμα Ηλεκτρονικών Επικοινωνιών, Υπουργείο Μεταφορών Επικοινωνιών και Έργων**
Έκδοση	:	**1.6**
Ημερομηνία	:	**13/01/2020**

5 Ανάπτυξη Ηθικής και Αξιόπιστης ΤΝ

Βρισκόμαστε μόλις στην πρώτη φάση προώθησης της ΤΝ και είναι αναγκαίο να συνεχιστεί ο διάλογος με όλους τους εμπλεκόμενους φορείς. Οι επιπτώσεις είναι δύσκολο να προβλεφθούν για δύο κυρίως λόγους: ο πρώτος λόγος είναι ο απρόβλεπτος ρυθμός της τεχνολογικής ανάπτυξης και ο δεύτερος λόγος είναι ότι η τεχνολογική ανάπτυξη από μόνη της δεν καθορίζει τον τρόπο με τον οποίο η εργασία και η κοινωνία θα αλλάξουν. Ως εκ τούτου καθορίζεται η ανάγκη να κατανοήσουμε τους τρόπους με τους οποίους η ΤΝ επηρεάζει ζητήματα ηθικής και ανθρωπίνων δικαιωμάτων, ούτως ώστε να αντιμετωπιστούν ζητήματα αξιοπιστίας της ίδιας της τεχνολογίας.

Response: Industry & Professions



usacm ACM US Public Policy Council

acm Europe Council

Principles for Algorithmic Transparency and Accountability

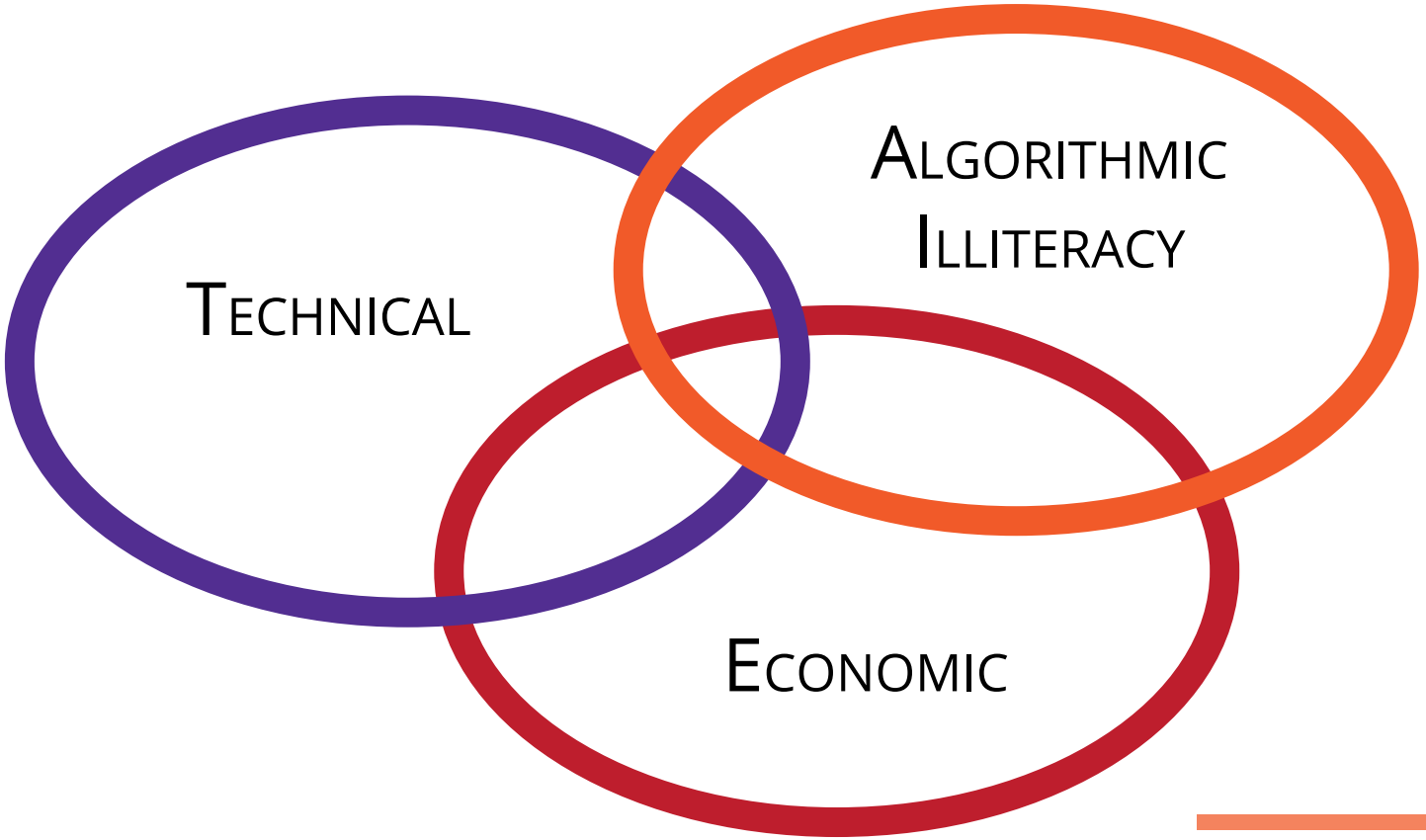
1. Awareness: Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

5. Data Provenance: A description of the way in which the training data was collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process. Public scrutiny of the data provides maximum opportunity for corrections. However, concerns over privacy, protecting trade secrets, or revelation of analytics that might allow malicious actors to game the system can justify restricting access to qualified and authorized individuals.

Challenges

- What exactly does transparency mean?
- And fairness? Fair for whom?
  - 21 fairness definitions https://www.youtube.com/watch?v=jIXIuYdnyyk
- Bias - what is the baseline?
  - specific aspects of bias in ICT systems (e.g., based on age, gender, race, popularity, etc.)
- Diversity
  - different approaches and representations

Lack of Transparency

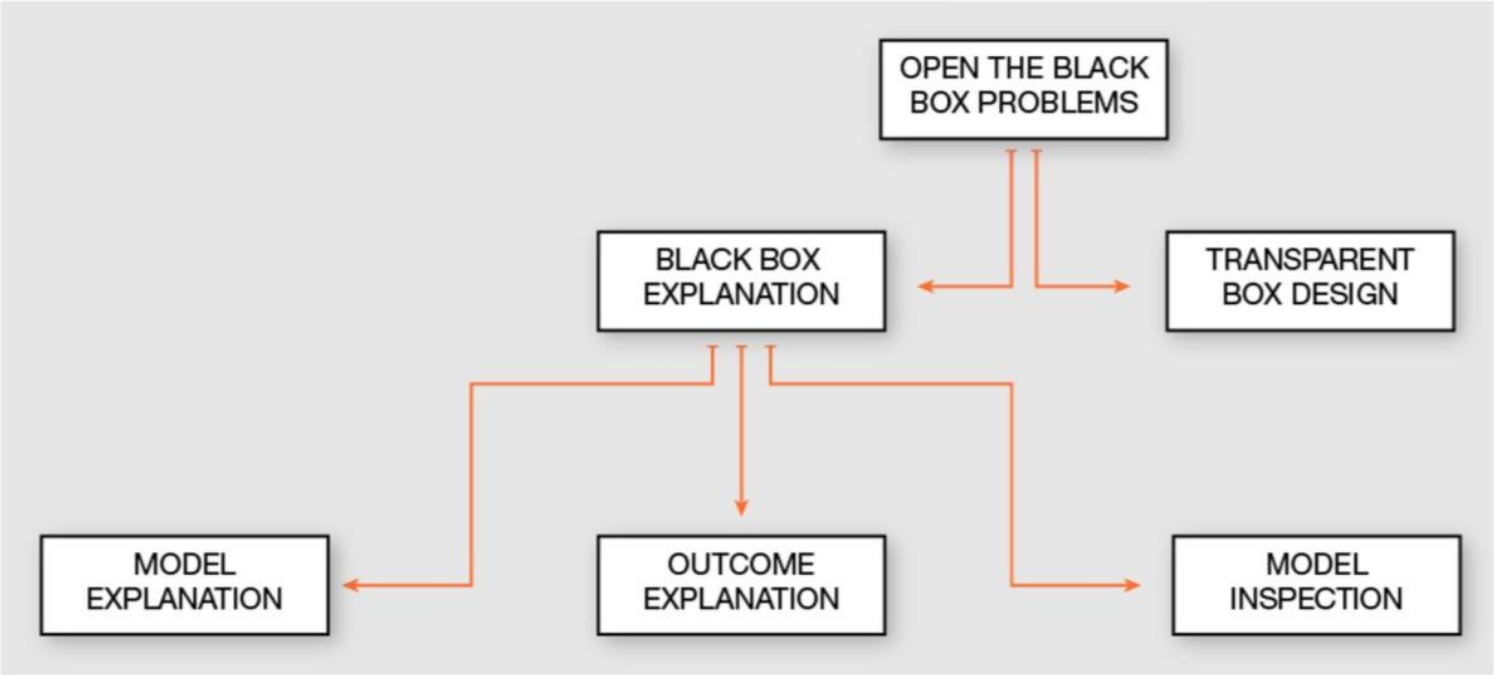


What is Algorithmic Transparency?

They carry social values; some exhibit **biases** or systematically produce results that could lead to **discrimination** against certain people.

Many users remain totally unaware that algorithms play a key role in their access to public information.

Black Box problems taxonomy
[Guidotti et al. 2018]



Model Inspection Problem

"...providing a representation for **understanding some specific property** of the black box model or of its predictions." (p. 14)

Inspection via **Partial Dependence**

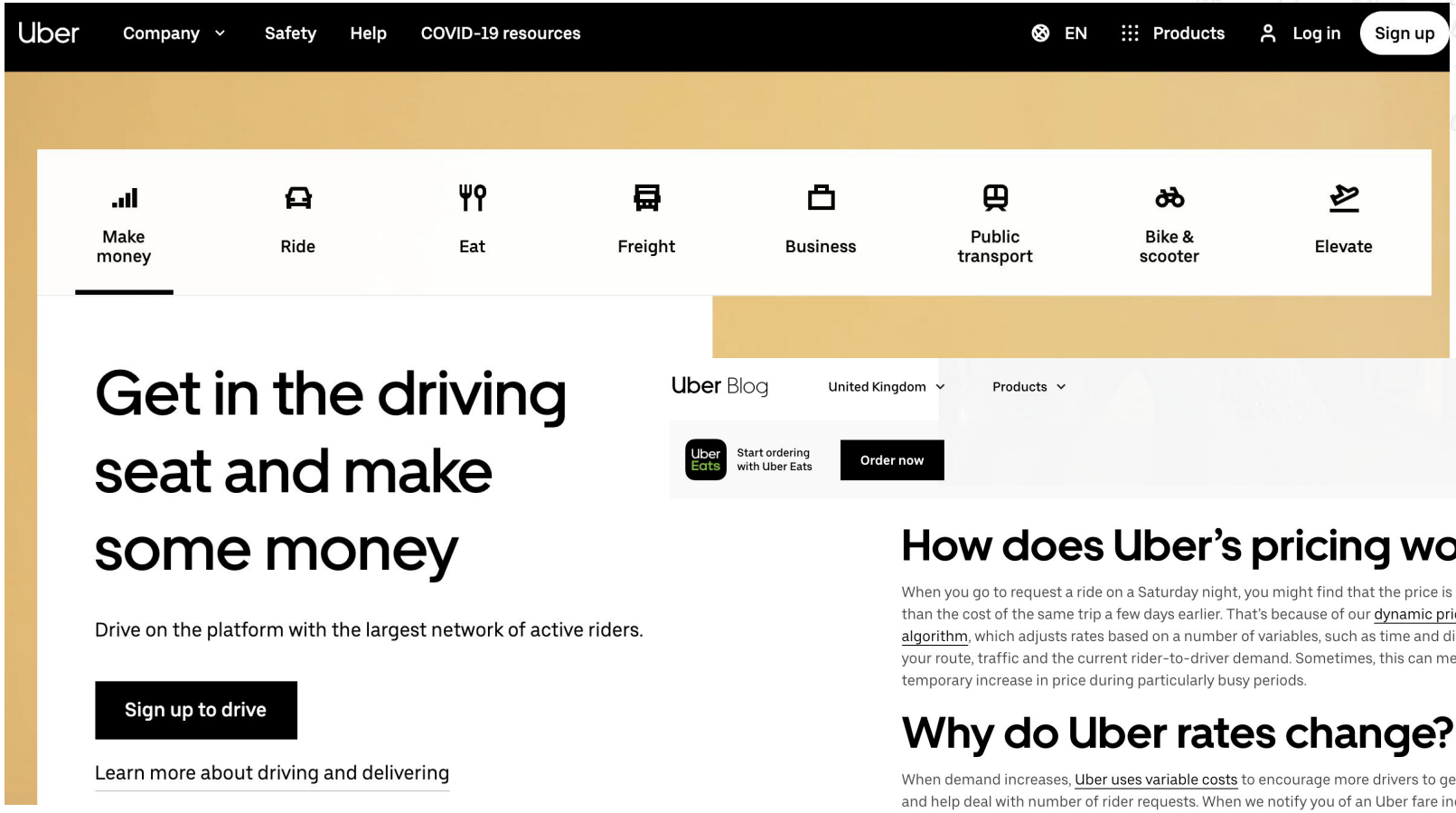
"...a tool for visualizing the relationship between the response variable and predictor variables in a reduced feature space." (p. 31)

FATE as a Scientific Field

Background: FATE Research

- Some illustrative examples
  - Uber
    Dynamic pricing algorithms
  - Fiverr & TaskRabbit freelance marketplaces
    Recommendation systems
  - Search engines
    Information access (ranking, personalization)
  - Image tagging APIs
    Computer vision

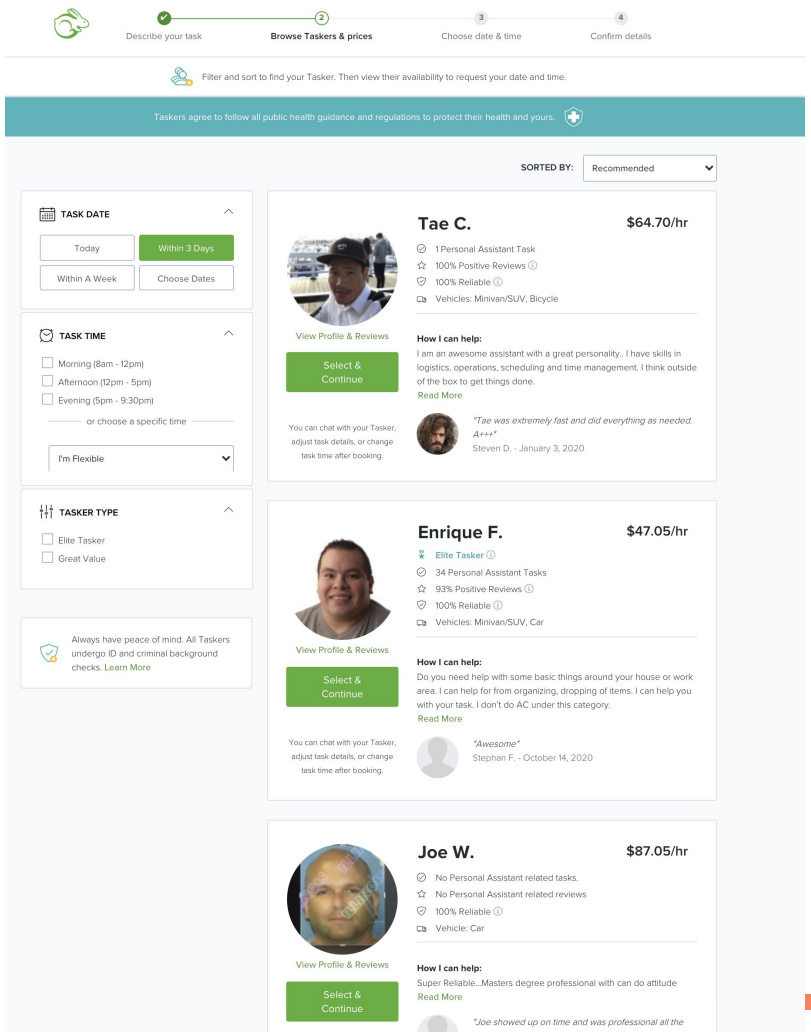
UBER


Uber
Company
Safety
Help
COVID-19 resources
EN
Products
Log in
Sign up
Make money
Ride
Eat
Freight
Business
Public transport
Bike & scooter
Elevate
Get in the driving seat and make some money
Drive on the platform with the largest network of active riders.
Sign up to drive
Learn more about driving and delivering
Uber Blog
United Kingdom
Products
Uber Eats
Start ordering with Uber Eats
Order now
How does Uber's pricing work?
When you go to request a ride on a Saturday night, you might find that the price is different than the cost of the same trip a few days earlier. That's because of our dynamic pricing algorithm, which adjusts rates based on a number of variables, such as time and distance of your route, traffic and the current rider-to-driver demand. Sometimes, this can mean a temporary increase in price during particularly busy periods.
Why do Uber rates change?
When demand increases, Uber uses variable costs to encourage more drivers to get on the road and help deal with number of rider requests. When we notify you of an Uber fare increase, we notify drivers as well. If you decide to go ahead and request your ride, you'll get an alert on the app to make sure you know that the rates have changed.
Price normalisation
Once more drivers get on the road and ride requests are taken, the demand will become more manageable and fares should revert to normal.
Share

Rosenblat, A., & Stark, L. (2016). Algorithmic labor and information asymmetries: A case study of Uber's drivers. *International Journal of Communication*, *10*, 27.

- "Uber's claims regarding its labor model, which center on *freedom, flexibility, and entrepreneurship*, are complicated and contradicted by the experience of its drivers."
- "Power and information asymmetries emerge via Uber's software-based platform through algorithmic labor logistics *shaping driver behavior*, electronic surveillance, and policies for performance targets. "
- "Through the Uber app's design and deployment, the company produces the equivalent effects of what most reasonable observers would define as a *managed labor force*."

Task Rabbit

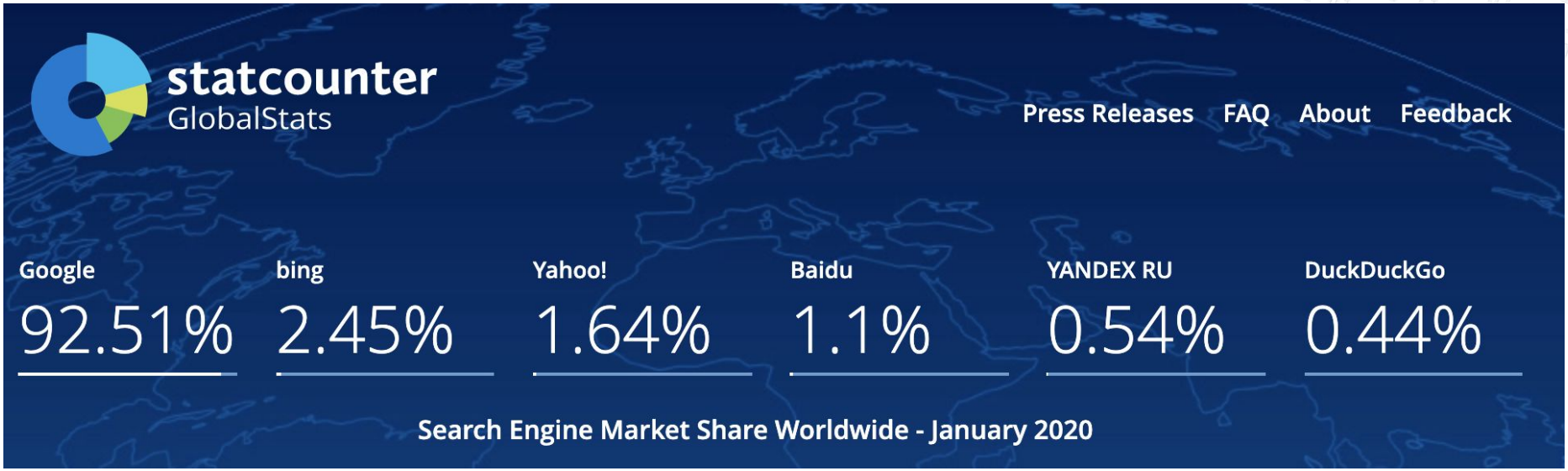

Describe your task
Browse Taskers & prices
Choose date & time
Confirm details
Filter and sort to find your Tasker. Then view their availability to request your date and time.
Taskers agree to follow all public health guidance and regulations to protect their health and yours.
SORTED BY: Recommended
TASK DATE
Today
Within 3 Days
Within A Week
Choose Dates
TASK TIME
Morning (8am - 12pm)
Afternoon (12pm - 5pm)
Evening (5pm - 9:30pm)
or choose a specific time
I'm Flexible
TASKER TYPE
Elite Tasker
Great Value
Always have peace of mind. All Taskers undergo ID and criminal background checks. Learn More
Tae C.
$64.70/hr
1 Personal Assistant Task
100% Positive Reviews
100% Reliable
Vehicles: Minivan/SUV, Bicycle
View Profile & Reviews
Select & Continue
How I can help:
I am an awesome assistant with a great personality.. I have skills in logistics, operations, scheduling and time management. I think outside of the box to get things done.
Read More
You can chat with your Tasker, adjust task details, or change task time after booking.
"Tae was extremely fast and did everything as needed. A+++"
Steven D. - January 3, 2020
Enrique F.
$47.05/hr
Elite Tasker
34 Personal Assistant Tasks
93% Positive Reviews
100% Reliable
Vehicles: Minivan/SUV, Car
View Profile & Reviews
Select & Continue
How I can help:
Do you need help with some basic things around your house or work area. I can help for from organizing, dropping of items. I can help you with your task. I don't do AC under this category.
Read More
You can chat with your Tasker, adjust task details, or change task time after booking.
"Awesome"
Stephan F. - October 14, 2020
Joe W.
$87.05/hr
No Personal Assistant related tasks.
No Personal Assistant related reviews
100% Reliable
Vehicle: Car
View Profile & Reviews
Select & Continue
How I can help:
Super Reliable...Masters degree professional with can do attitude
Read More
"Joe showed up on time and was professional all the

Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., & Wilson, C. (2017, February). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing* (pp. 1914-1933).

Audit of worker rankings & reviews

- "Workers perceived to be women, especially White women, receive 10% fewer reviews than workers perceived to be men with equivalent work experience."
- "Workers perceived to be Black, especially men, receive significantly lower feedback scores (i.e., ratings) than other workers with similar attributes."

Search Engine Bias(?)

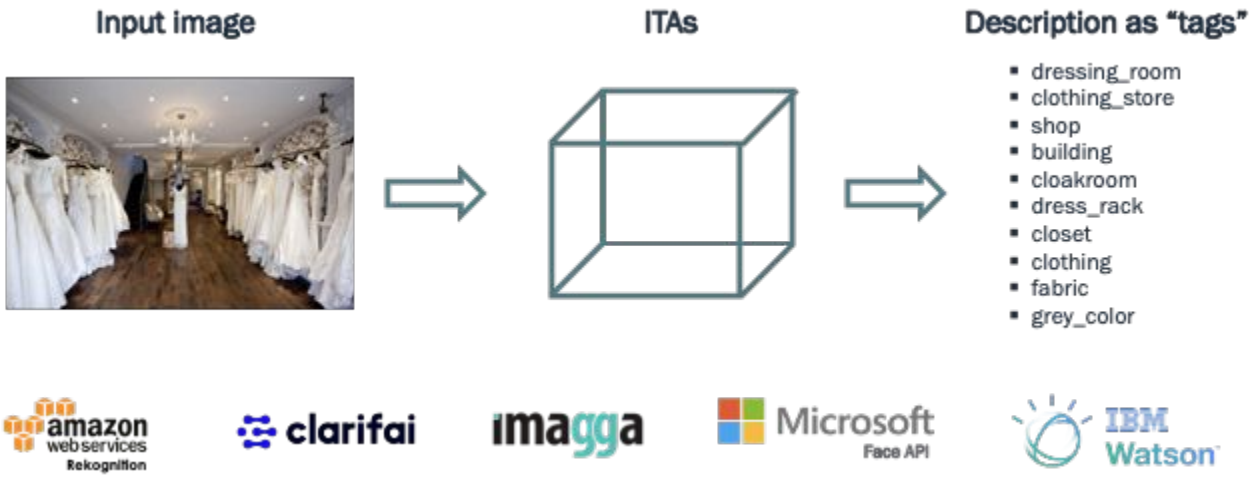


Search Engine Bias(?)

Mowshowitz, A., & Kawaguchi, A. (2005). Measuring search engine bias. *Information Processing & Management*, *41*(5), 1193-1205.

- Methodology for quantifying “bias” in search engine results, as a relative measure
- “The bias measure is designed to capture the degree to which the distribution of URLs, retrieved by a search engine in response to a query deviates from an idea of fair distribution for that query.”
- Experiments with 16 (!) search engines
- Main conclusion: lots of variance between engines, and by subject / topic

Image Tagging Algorithms



BF-231 from the Chicago Face Dataset, and tags output by the six image tagging APIs for this image

Amazon Rekognition	human, people, person, Afro, hairstyle, hair, face
Clarifai	people, one, portrait, man, wear, adult, side, pensive, profile, woman, face, isolated, child, facial, Afro, casual, fashion, athlete, adolescent
Google Cloud Vision	face, forehead, chin, eyebrow, cheek, nose, head, jaw, neck, human
Imagga Auto-tagger	Afro, man, face, portrait, male, handsome, head
Microsoft Vision	man, person, wearing, looking, necktie, standing, shirt, front, face, smiling, white, suit, posing, hair, holding, neck, young, glasses, black, head, hat, red
Watson Vision	person, woman, female, indian red color, coal black color

Auditing the Black Boxes

Kyriakou, K., Barlas, P., Kleanthous, S., & Otterbacher, J. (2019, July). Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 313-322).

Two approaches:

- **within-platform audits:** to discover how outputs may differ *for certain categories of inputs* in one system (e.g., Sweeney 2013)
- **cross-platform audits:** to discover how all outputs of *one system* may differ from *outputs of other systems*, for the same input (e.g., Eslami et al. 2017)

Are Taggers Fair?
The Short Answer: No

- "Some [taggers] offer more interpretation on images, they may exhibit less fairness toward the depicted persons, by misuse of gender-related tags and/or making judgments on physical appearance."

  - Asian females → more "attractiveness" tags
  - Black males → less interpretive tags

User Perception of Fairness

Barlas, P., Kleanthous, S., Kyriakou, K., & Otterbacher, J. (2019, June). What Makes an Image Tagger Fair?. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 95-103).

"Today, many automated tools are used to generate descriptions of images on the Web. However, some tools exhibit biases when processing images of people. Given an image and two descriptions of its content, decide which one is more fair."

"Imagine that auto-tagging is used to facilitate **searching profiles of people at a dating site**. Which of the above descriptions is **more fair**? Enter 0 if you cannot tell."

"Please **explain your answer regarding fairness**."

Description 1:

adolescent
contemporary
casual
cute
child
facial expression
eye
fine looking
fashion
fun
isolated
funny
man
one
looking
serious
people
portrait
wear
young

clarifai

Description 2:

blue eyes
woman
girl
caucasian
brown hair
front view
plain expression
short bangs
solo
sober
young
lip gloss
round face
white background
long hair

Crowdworkers

Experimental Set-up

Image	Gender	Race	“Attractiveness”	Participants (W/M)
BF-231	Woman	Black	Average	20/20
BF-233	Woman	Black	Attractive	20/20
WF-036	Woman	White	Average	20/20
WF-233	Woman	White	Attractive	20/20
BM-009	Man	Black	Average	20/20
BM-234	Man	Black	Attractive	20/20
WM-022	Man	White	Average	20/20
WM-004	Man	White	Attractive	20/20

Which is more “fair”: human or algorithm?

				Human more fair	Estimate	Z	Odds ratio
Average	Black	Woman	Intercept (BF-231)	.78	1.237	3.266**	3.44
Average	White	Woman	WF-036	.93	1.276	1.797	3.58
Attractive	Black	Woman	BF-233	.70	-3.895	-0.760	0.68
Attractive	White	Woman	WF-233	.48	-1.337	-2.708**	0.263
Average	Black	Man	BM-009	.65	-0.6177	-1.227	0.54
Average	White	Man	WM-022	.75	-1.382	-0.263	0.87
Attractive	Black	Man	BM-234	.78	-4.498	0.000	1.00
Attractive	White	Man	WM-004	.28	-2.206	-4.256***	0.110

Logit model to predict the event that human-generated tags are perceived as being more fair.

*** $p < .001$
** $p < .01$
* $p < .05$

Explaining fairness

Accuracy

"This is fair as the description is more accurate."

Physical visual characteristics

"I liked that it focused on aspects about the image, such as her hair and eye color."

Objective/Subjective

"This is more fair because it is not subjective and is accurate and less open to interpretation."

Understanding

"If someone gave me that I would be able to tell what the person looked like easier."

Political Correctness

"A lot of the words would not be described as favorable or putting the person in a good light."

Demographics

"It does not emphasize racial characteristics."

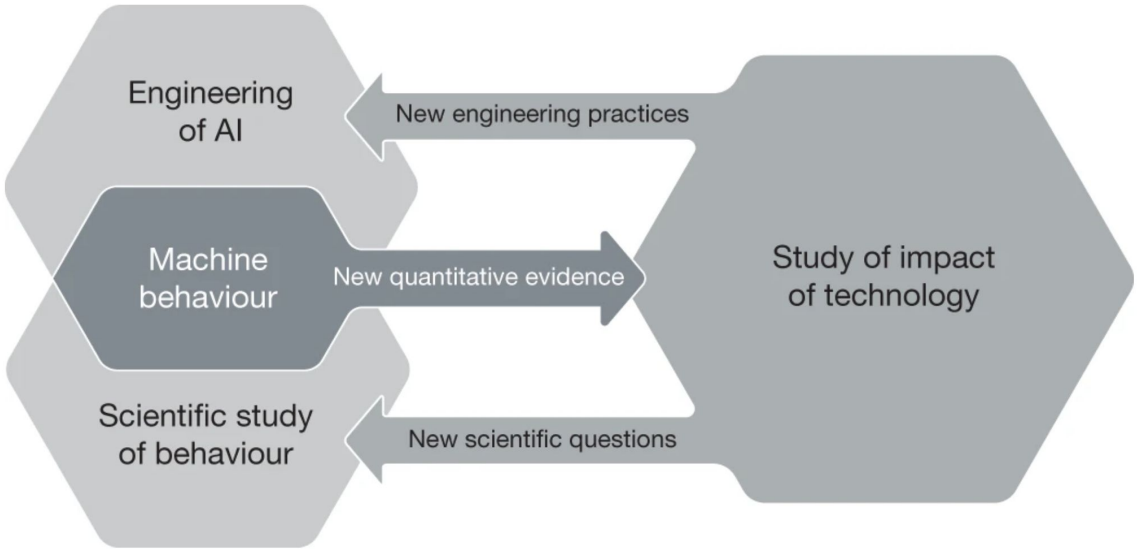
Discussion & Final Thoughts

Review | Published: 24 April 2019

Machine behaviour

Iyad Rahwan ✉, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C. Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B. Tenenbaum & Michael Wellman

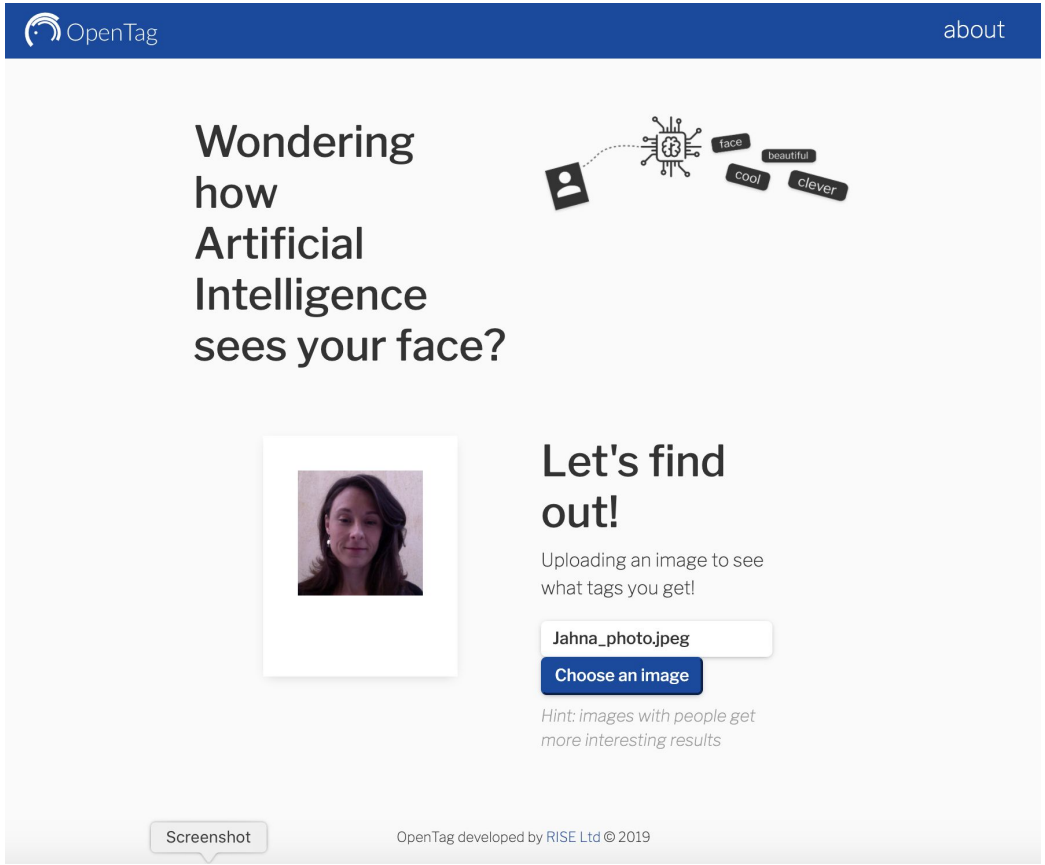
Nature **568**, 477–486(2019) | Cite this article



Post-seminar questionnaire

https://forms.gle/vijRR9FekD6r4Hox9

User Study – Invitation!



http://ec2-34-255-198-84.eu-west-1.compute.amazonaws.com/opentag

Thank you!

- www.**cycat.io**
- facebook.com/**CyCAT.EU**
- twitter.com/**CyCAT_EU**
- linkedin.com/in/**CyCAT**

This research is partially funded by the European Union's Horizon 2020 research and innovation program under grant agreements No. 739578 (RISE), 810105 (CyCAT) and the Government of the Republic of Cyprus (RISE).