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

Developer Seminar Objectives

In this 10-hour 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/KiuNQACwZRMNh8H36

Seminar Overview - Day 2

Overview and questions	14.00 - 14.10
COMPAS case study discussion	14.10 - 14.40
FATE Problems	14.40 - 15.10
Break	15.10 - 15.25
FATE Solutions	15.25 - 16.25
Exercise in breakout rooms	16.25 - 17.00
Post-seminar questionnaire	17.00 - 17.30
Discussion and final thoughts	17.30 - 18.00

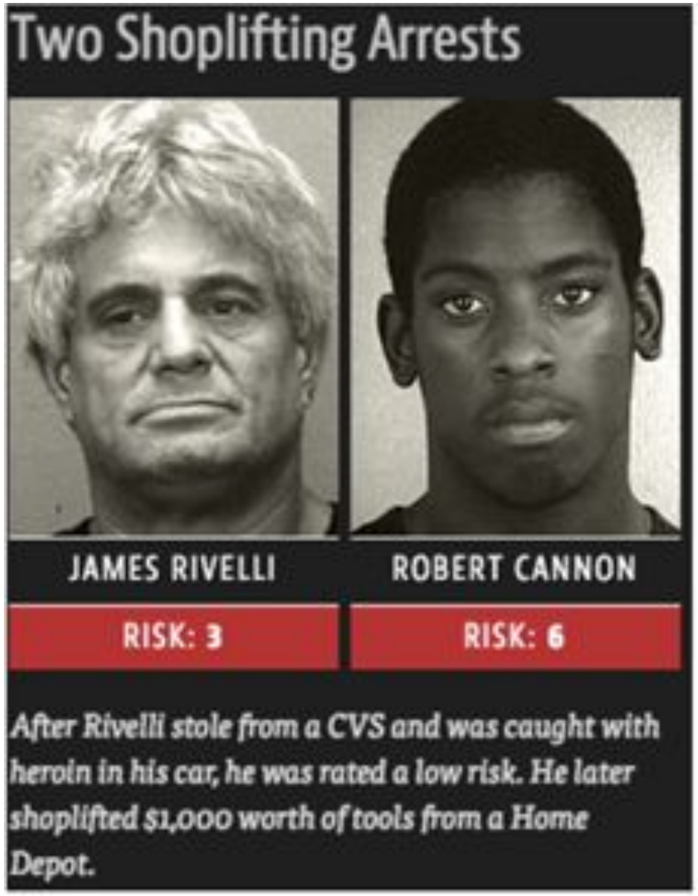
COMPAS CASE STUDY

COMPAS SYSTEM

- The COMPAS system is widely used in US courts to predict the risk of recidivism by criminal defendants.
- Intended to support judges, probation and parole officers (**system users**) to assess a criminal defendant's likelihood of becoming a recidivist.
- COMPAS provides scores from 1 (being lowest risk) to 10 (being highest risk).
- The input used for prediction of recidivism is wide-scale and uses 137 factors including age, gender, and criminal history of the defendant.
- Race is *not* an explicit feature considered by the model.

- Larson et. al analysis show that black defendants were more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn’t Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%



Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica, May*, *23*, 2016.
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS DATASET

https://github.com/propublica/compas-analysis

Story:

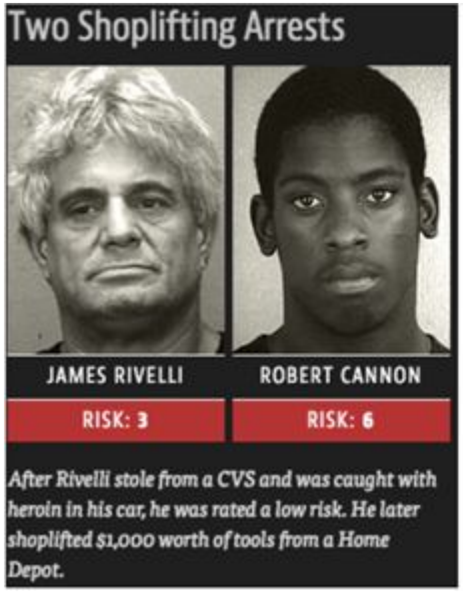
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing/

Methodology:

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/

Does the system treat different groups of defendants in a similar manner?
data journalists (Angwin et al., 2016)

How can transparency of the data and the method ensure the algorithm's fairness?
data scientists (Rudin et al., 2020)



Can a scoring algorithm respect multiple definitions of fairness?
computer scientists (Kleinberg et al., 2017)

How does the user's knowledge of statistics, as well as the justice system affect her ability to use the system?
legal scholars (Ridgeway, 2020)

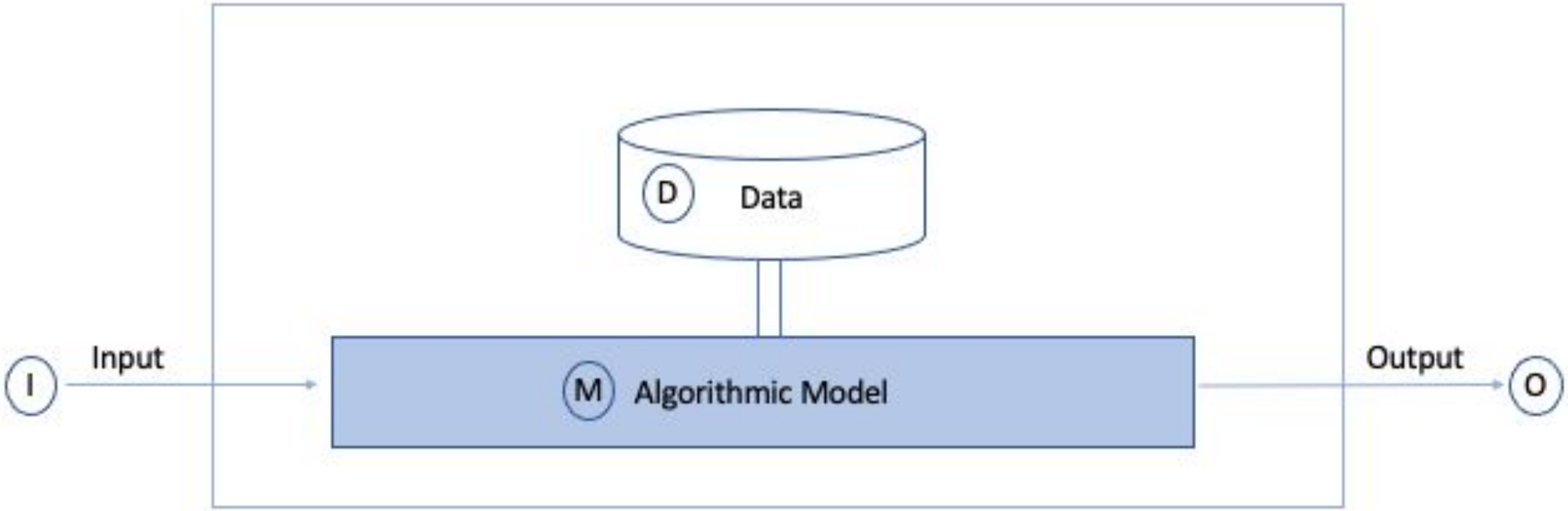
STAKEHOLDERS

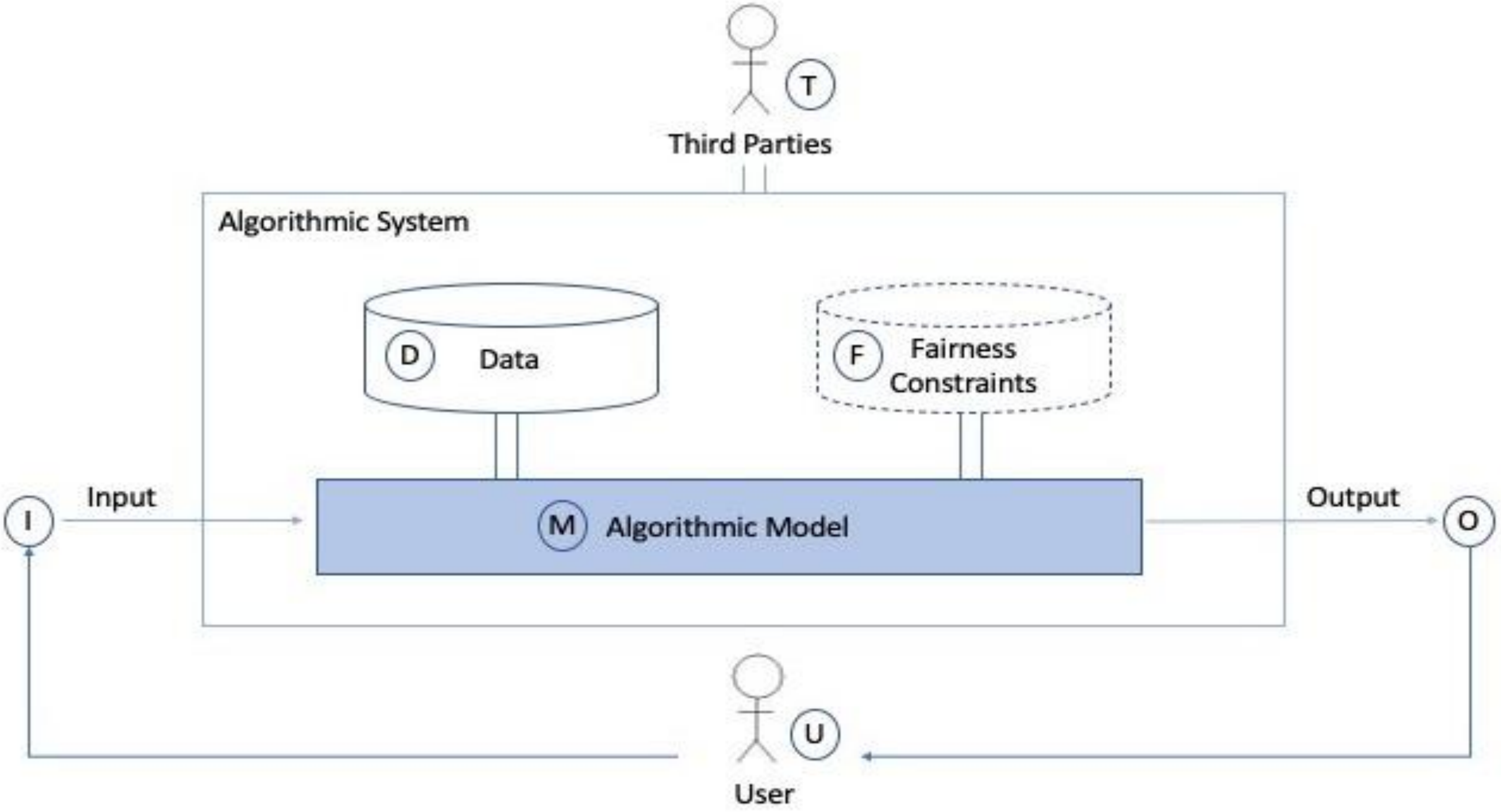
- Observers: typically have limited access to the process/system
  - Researchers
  - Journalists
  - Regulators

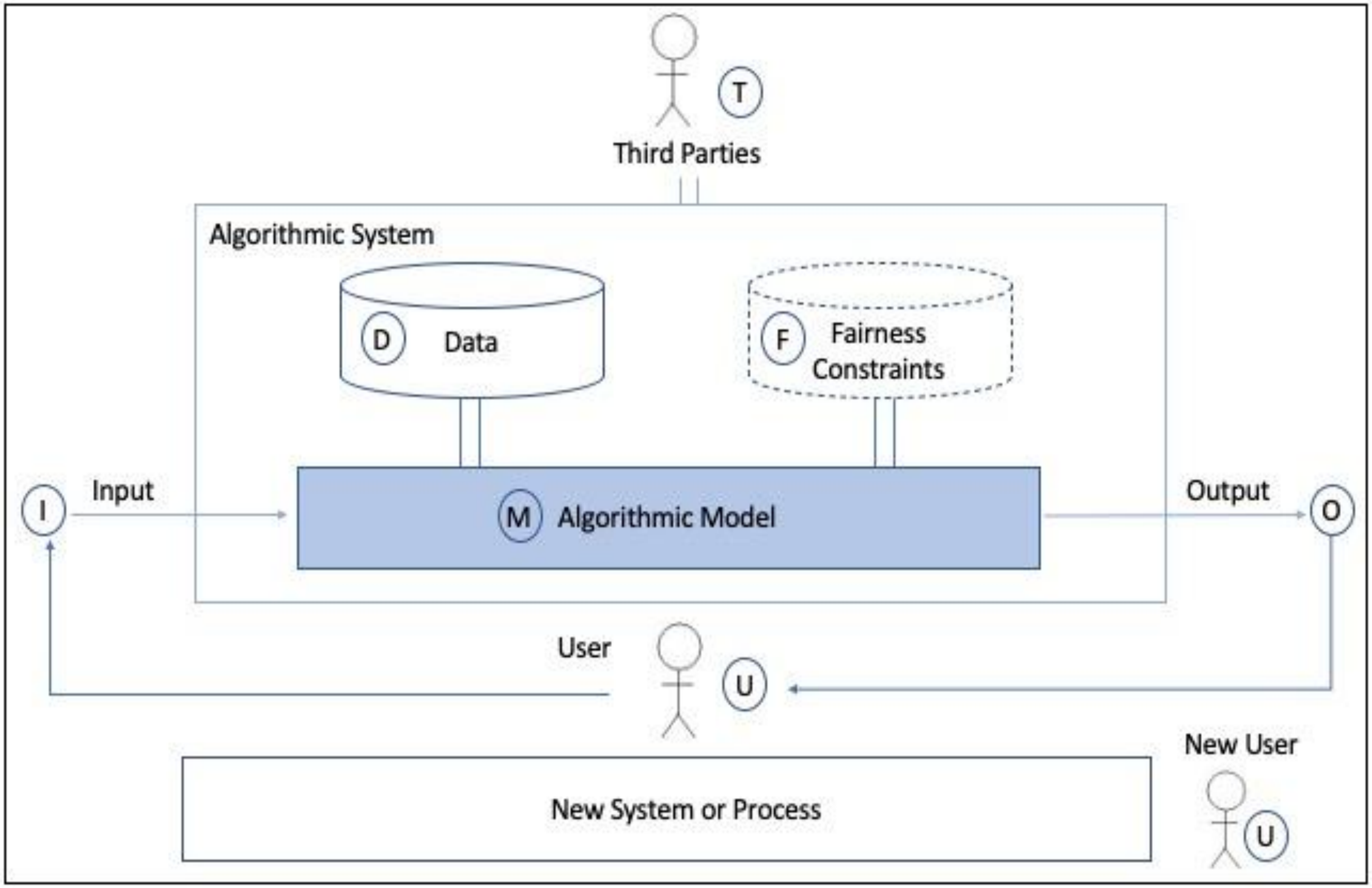
- Developers: have access to the process/system
  - ML practitioners
  - Interface designers
  - Data managers

- Users: rely on / are affected by the process/system

FATE PROBLEM SPACE


D
Data
I
Input
M
Algorithmic Model
Output
O


T
Third Parties
Algorithmic System
D
Data
F
Fairness
Constraints
I
Input
M
Algorithmic Model
Output
O
U
User



Third Parties
T
Algorithmic System
D
Data
F
Fairness
Constraints
I
Input
M
Algorithmic Model
Output
O
User
U
New User
New System or Process
U

CASE STUDY 1: AD_SERV SYSTEM

1. **Input:** Data provided by the specific end user, the specific content providers and advertisers.

2. **Training Data:** The algorithm is initially trained by data provided by the advertisers. It subsequently learns from the behaviour of all users, advertisers and content providers.

3. **Third Party constraints:** These constraints are supplied by the advertisers who target their marketing to specific market segments. Other constraints may be provided by the content providers who ban or encourage certain classes of advertisers from their sites or apps.

4. **Algorithm:** The algorithm provided by the owner attempts to maximize click through rates in order to satisfy its customers (the advertisers and content providers).

5. **Output:** The algorithm provides a set of ads that are displayed on the content providers platform for a particular end user.

CASE STUDY 2: CREDIT_RATE SYSTEM

1. **Input:** Data provided about the specific end user, by the bank officer the specific content providers and advertisers.

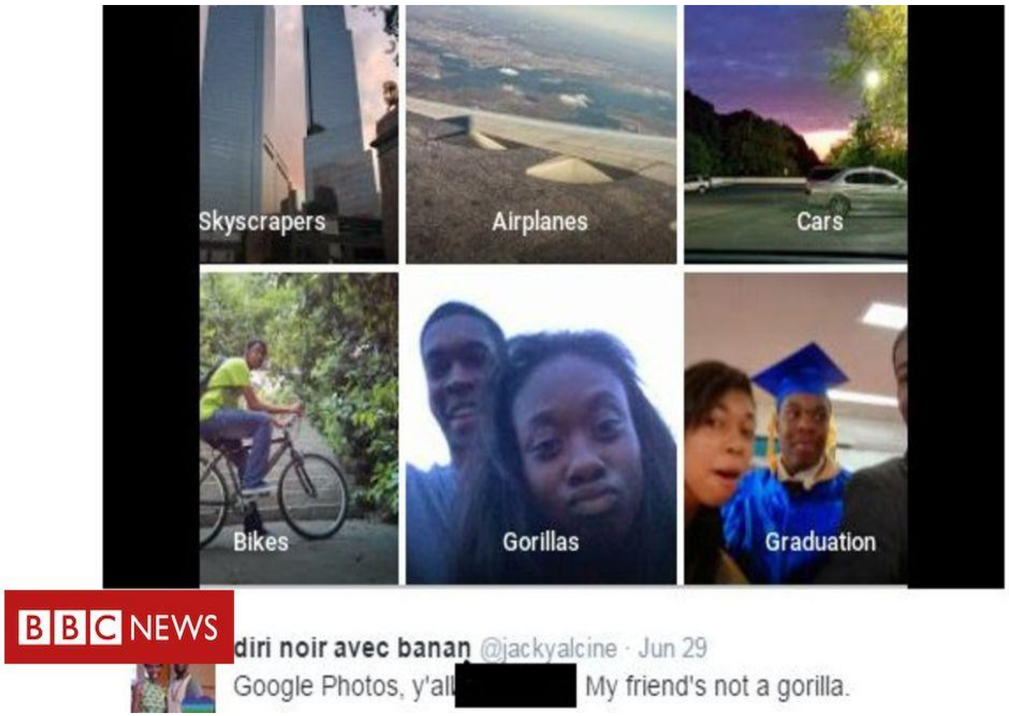
2. **Training Data:** The algorithm is initially trained using the bank's historical data.

3. **Third Party constraints:** These constraints are defined by the bank officers.

4. **Algorithm:** The algorithm implemented assesses the risks and benefits of approving the credit request given the historical data and the system configuration.

5. **Output:** The algorithm provides a decision whether to approve or deny the customer's request.

EXAMPLES OF BIAS



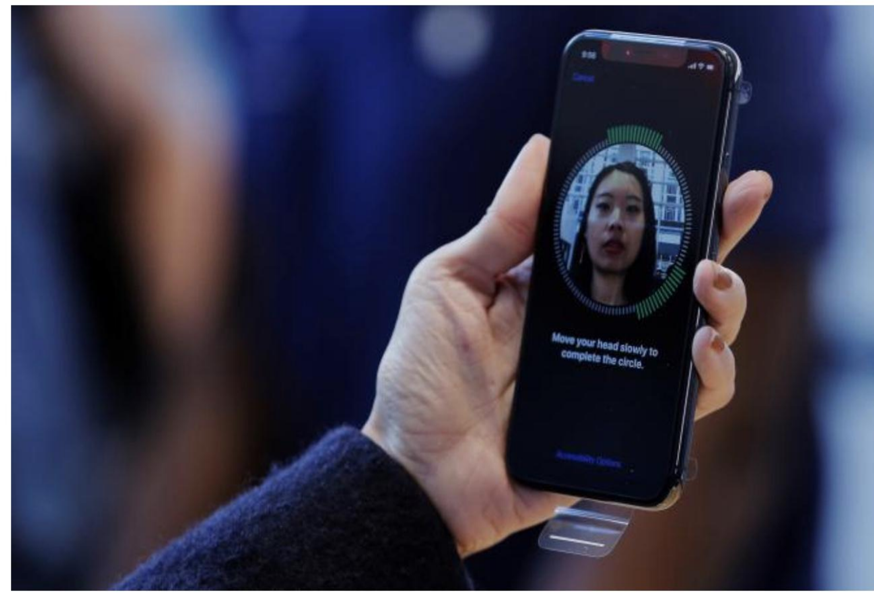
Bias in Training Data

Data Bias

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



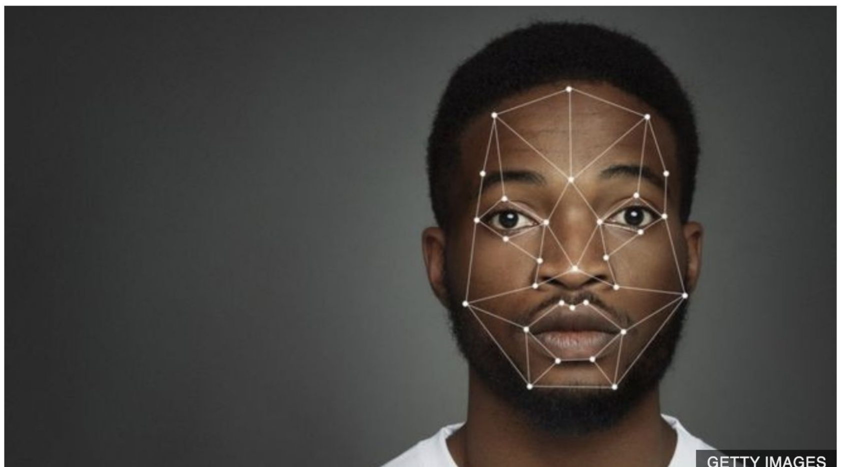
A woman sets up her facial recognition as she looks at her Apple iPhone X at an Apple store in New York, U.S., November 3. Last week a woman in China claimed that her iPhone X facial recognition could not tell her and her colleague apart.

IBM abandons 'biased' facial recognition tech

9 June 2020

Share

George Floyd death

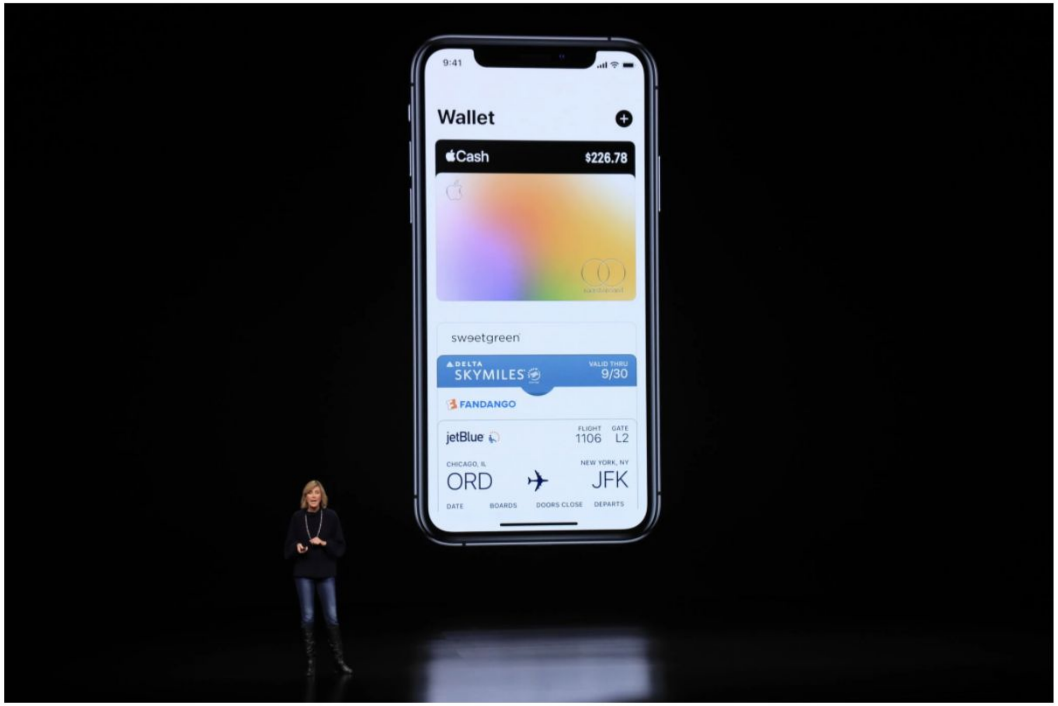


A US government study suggested facial recognition algorithms were less accurate at identifying African-American faces

Input Data Bias

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.



Algorithmic Model Bias

CY. center for algorithmic transparency

The UK used a formula to predict students' scores for canceled exams. Guess who did well.

The formula predicted rich kids would do better than poor kids who'd earned the same grades in class.

By Kelsey Piper | Aug 22, 2020, 7:30am EDT

SHARE

Protesters in London objected to the government's handling of exam results after exams were canceled due to the coronavirus outbreak. | Aaron Chown/PA Images via Getty Images

Output Bias

MOST READ

Democrats are cheering a Supreme Court ruling

BIAS IN SEARCH ENGINES

DATA BIAS

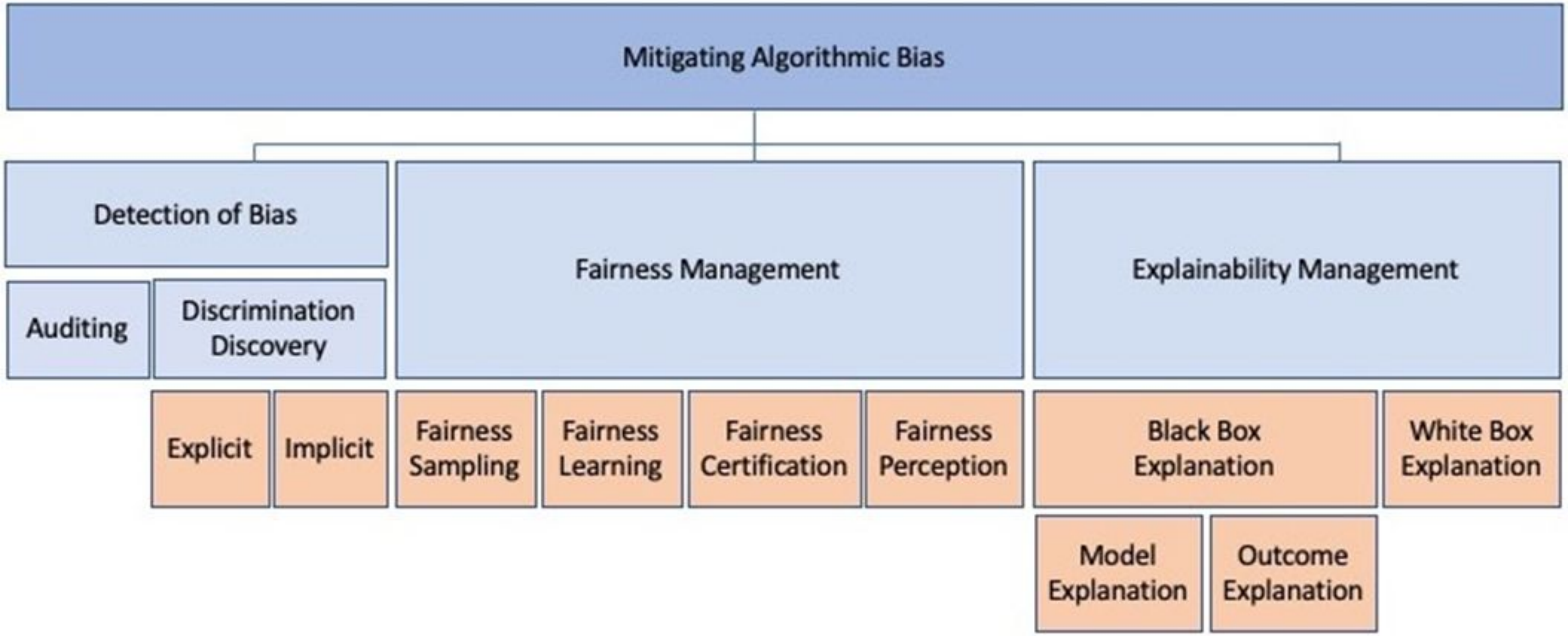
MODEL PROCESSING BIAS

OUTPUT BIAS

Break (15 minutes)

FATE SOLUTION SPACE

SOLUTION SPACE



DETECTION OF BIAS

- **Auditing**
  - **Within-system:** to discover how outputs may differ *for certain categories of inputs* in one system.
  - **Cross-system audit:** to discover how all outputs of one system may differ from outputs of other systems, for the same input.
  - Automatic Auditing tools
- **Discrimination Discovery**
  - **Explicit** (direct) **Discrimination Discovery:** The ability to identify discrimination which is caused by both data biases and inappropriate use of sensitive attributes in algorithms [Hannák et al. 2017].
  - **Implicit** (indirect) **Discrimination Discovery:** The ability to identify discrimination which is caused by algorithmic processing biases and human biases due to the fact that some insensitive attributes are very informative about sensitive attributes [Speicher et al. 2018].

AUDITING EXAMPLES

Problem	Stakeholder	Approach for Auditing	Research Domains
Data or Output Bias	User/Observer	Submit queries to search engines/Twitter	IR
Output or Model Bias	User	Analyzing system behavior	HCI
Data Bias	User/Observer	Auditing data from an application system	RecSys
Data or Model Bias	Developer	Auditing tools	ML

AUDITING TOOLS IN MACHINE LEARNING

FairML	A python toolbox for auditing machine learning models for bias
Aequitas	An open source bias audit toolkit to audit machine learning models for discrimination and bias
Audit-AI	A Python library that implements fairness-aware machine learning algorithms

DISCRIMINATION DETECTION EXAMPLES

Problem	Stakeholder	Approach for Discrimination Discovery	Research Domains
Data bias or Third party constraints	User	Crowdsourcing studies	HCI / IR
Data	Developer	Statistical metrics for discrimination e.g. absolute measures, conditional measures or statistical tests.	ML
Data / Output	User	Analysis of web logs	IR
Output	User	Discrimination detection in advertising recommender systems	RecSys
Model/Output	Developer	Discrimination detection in evaluation metrics	RecSys

FAIRNESS MANAGEMENT

- **Fairness Sampling:** Processing the data in a manner that promotes fairness.
- **Fairness Learning:** Mitigating bias in model processing for promoting fairness.
- **Fairness Certification:** Test algorithmic models for possible disparate impact, "certifying" those that do not exhibit evidence of unfairness.
- **Fairness Perception:** concerns the perception of users with the decision making outcome and it can be measured through questionnaires and statistical tests.

FAIRNESS SAMPLING SOLUTIONS - EXAMPLES

Problem	Stakeholder	Approach for Fairness Sampling	Research Domains
Data (imbalanced data)	Developer	Data balancing using data mining techniques (cross validation, imbalanced techniques) or re-sampling using statistics	ML/IR/HCI
Data (Missing important features)	Developer/Third Party	Add new features	ML/HCI/IR
Data	Developer	Remove protected attributes (e.g. race) from the input data	ML
Data	Developer	Automated generated data	HCI/ML

Solution: Add new features

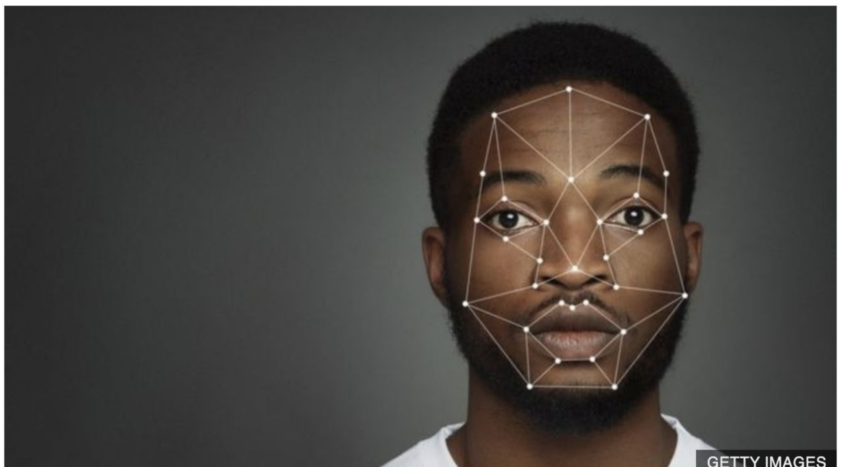
Solution: Re-balancing data

IBM abandons 'biased' facial recognition tech

9 June 2020

Share

George Floyd death



GETTY IMAGES

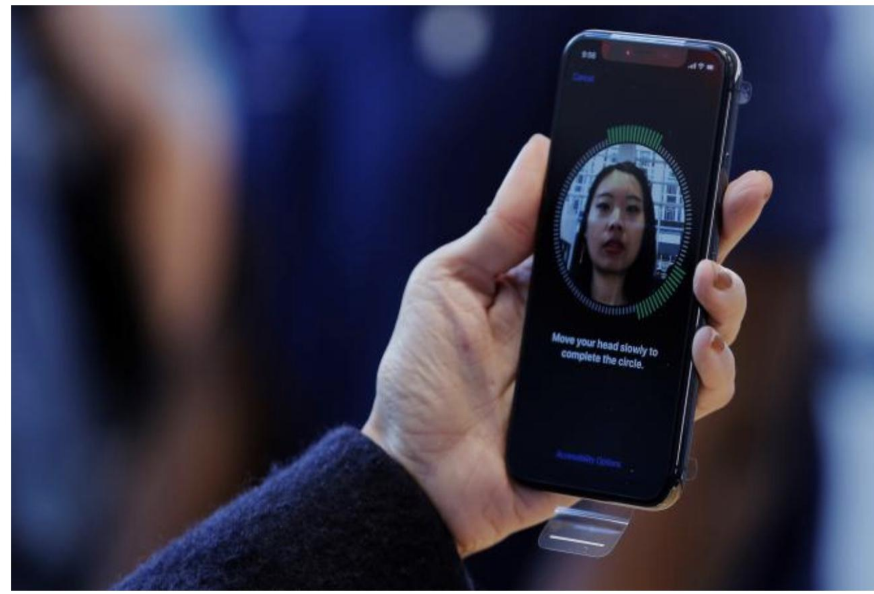
A US government study suggested facial recognition algorithms were less accurate at identifying African-American faces

Solution: Re-balancing data

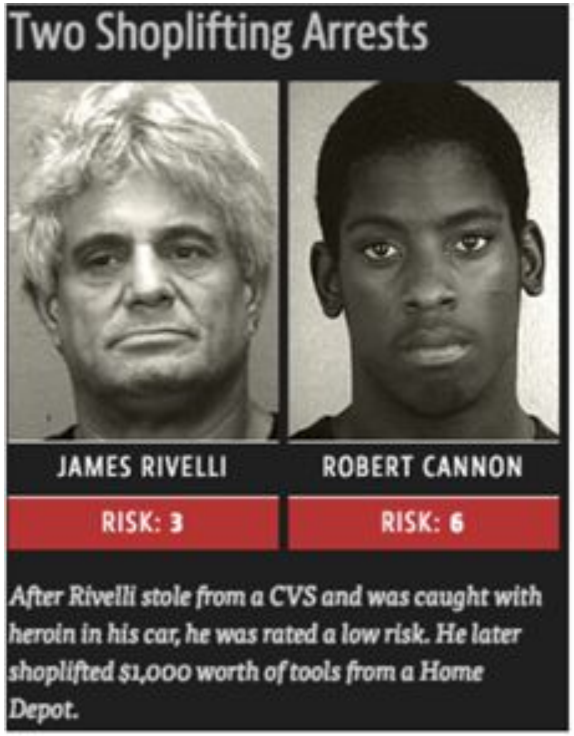
Newsweek

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.



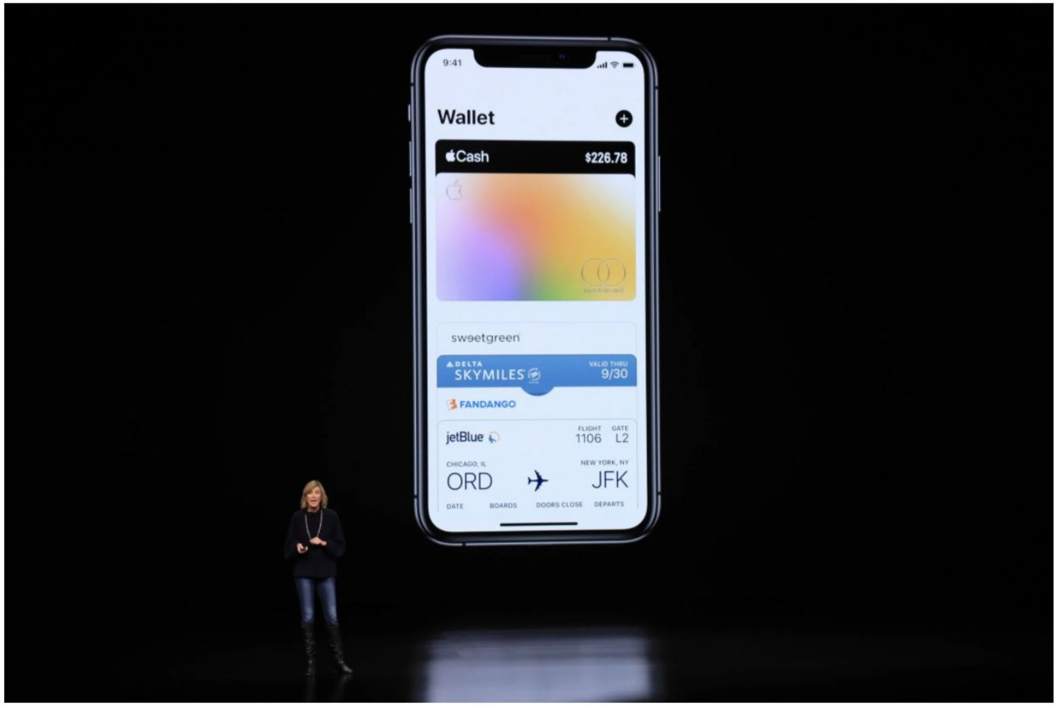
Solution: Adding fairness constraints

FAIRNESS LEARNING SOLUTIONS - EXAMPLES

Problem	Stakeholder	Approach for Fairness Learning	Research Domains
Model	Third party/Developer	Fairness constraints / fairness metrics	ML
Model/Output	Developer	Regularization approach	ML
Data/Model	Developer	Encrypted version of sensitive data	ML
Model	Developer/User	Human in the loop approach	HCI
Model	Developer/Third party	Fairness metrics to mitigate search engine bias	IR
Model/Output	Developer/Third party	Optimization approaches	RecSys

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.



Solution: encrypted version of sensitive data

The UK used a formula to predict students' scores for canceled exams. Guess who did well.

The formula predicted rich kids would do better than poor kids who'd earned the same grades in class.

By Kelsey Piper | Aug 22, 2020, 7:30am EDT

SHARE

Protesters in London objected to the government's handling of exam results after exams were canceled due to the coronavirus outbreak. | Aaron Chown/PA Images via Getty Images

Solution: Removal of sensitive attributes

MOST READ

Democrats are cheering a Supreme Court ruling

cy. center for algorithmic transparency

FAIRNESS CERTIFICATION SOLUTIONS - EXAMPLES

Problem	Stakeholder	Approach for Fairness Learning	Research Domains
Output	Developer/ third party	Altering of labels	ML
Output	User	Raise user awareness	IR
Output	User	Perceived fairness management	HCI

EXPLAINABILITY MANAGEMENT

- Black-box explanation
  - Model explanation
  - Outcome explanation
- White-box explanation

EXPLAINABILITY IN ML SYSTEMS: EXAMPLES

Problem	Stakeholder	Approaches for Explainability
Model	Developer	Decision tree mimic a black-box model
Data/Model	Developer	Feature-based explanation
Model	Developer	Decision rules explaining black-box model
Output	Developer/User	Visualization methods
Model/Output	Developer	Automatic tools

EXPLAINABILITY TOOLS IN ML

Tool	Link
LORE: Local rule-based explanations	https://www.ai4eu.eu/resource/lore-local-rule-based-explanations
LIME: Local-Interpretable Model Agnostic Explanations	https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf https://github.com/marcotcr/lime.
AI Explainability 360	https://aix360.mybluemix.net/
DeepLIFT (Deep Learning Important FeaTures)	https://github.com/kundajelab/deeplift
Microsoft InterpretML	https://github.com/interpretml

EXPLAINABILITY IN HCI SYSTEMS - EXAMPLES

Problem	Stakeholder	Approaches for Explainability
Output	Developer/User	Feature-based explanation
Output	User	Explanation styles
Output	User	Raise user’s awareness

EXPLAINABILITY IN RECOMMENDER SYSTEMS - EXAMPLES

Problem	**Stakeholder**	**Approaches for Explainability**
Model	Developer/User	Taxonomy of concepts
Output	User	Based on user’s opinions
Output	User	Matrix factorization

CASE STUDY 1: AD_SERV SYSTEM - DISCRIMINATION DISCOVERY

1. **Explicit discrimination** in AD_SERVE may be observed due to third party constraints (e.g. Do not show my ad to male end-users - may be legitimate if the advertiser is promoting female cosmetics.)
2. **Implicit discrimination** Sweeney (2013) note that ads for services providing criminal records on names were significantly more likely to be served if the name search was on a typically black first name.

CASE STUDY 1: MITIGATION OF FAIRNESS AND TRANSPARENCY RISKS

1. The system provides Explainability Management in the form of a response to the question “Why am I seeing this ad?”. The response could be a simple “Inspired by your browsing history” which is a **Black Box Outcome Explanation.**
2. Fairness Management **could be implemented for sensitive ads like those offering research into criminal records or other ads with potential for discriminatory display**

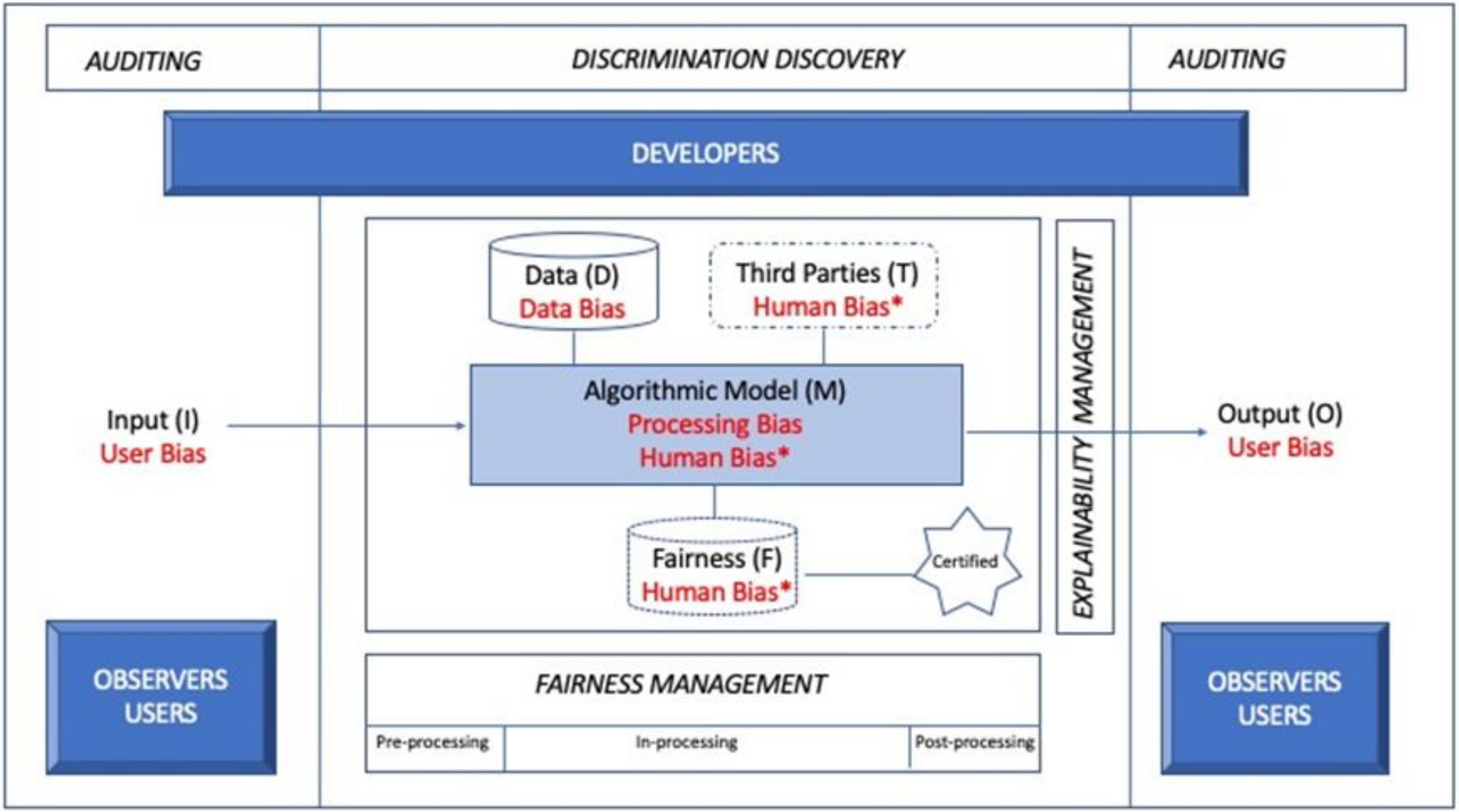
CASE STUDY 2: RISKS TO FAIRNESS AND TRANSPARENCY

1. **Explicit discrimination** may appear in the system has been configured to consider specific protected or proxi attributes as part of its reasoning (if this information is provided as input).
2. **Implicit discrimination** may appear if the training set used by the system includes protected or proxi attributes and it is biased in the sense that these attributes correlate with final decisions.

CASE STUDY 2: MITIGATION OF FAIRNESS AND TRANSPARENCY RISKS

1. The system provides **Explainability Management** in the form of an explanation of its decision as it is a black box, hence **Black Box Outcome Explanation** is provided**.**
2. **Fairness Management** could be implemented for ensuring group and individual parity

HIGH-LEVEL VIEW



CY. center for algorithmic transparency

Exercise

Explanation

Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithms and the specific decision that are made.

Try to understand MovieLens (https://movielens.org) explanations on the movie recommendations. Sign in, define a profile, rate a few movies and check your suggested recommendations. Explain why they were suggested by MovieLens and elaborate on the reasons/facts as you understand them. Provide suggestions on improving their algorithm, and what else can be taken into consideration while creating explanations.

Explanation (2)

Variation:

You might also investigate explanations in other recommender systems that you use (e.g., Amazon, Netflix, etc.)

It is also interesting to compare explanations of the recommendations you receive over time, as your user profile evolves over time.

POST-SEMINAR QUESTIONNAIRE

https://forms.gle/SuV24weHP1h34JHZ8

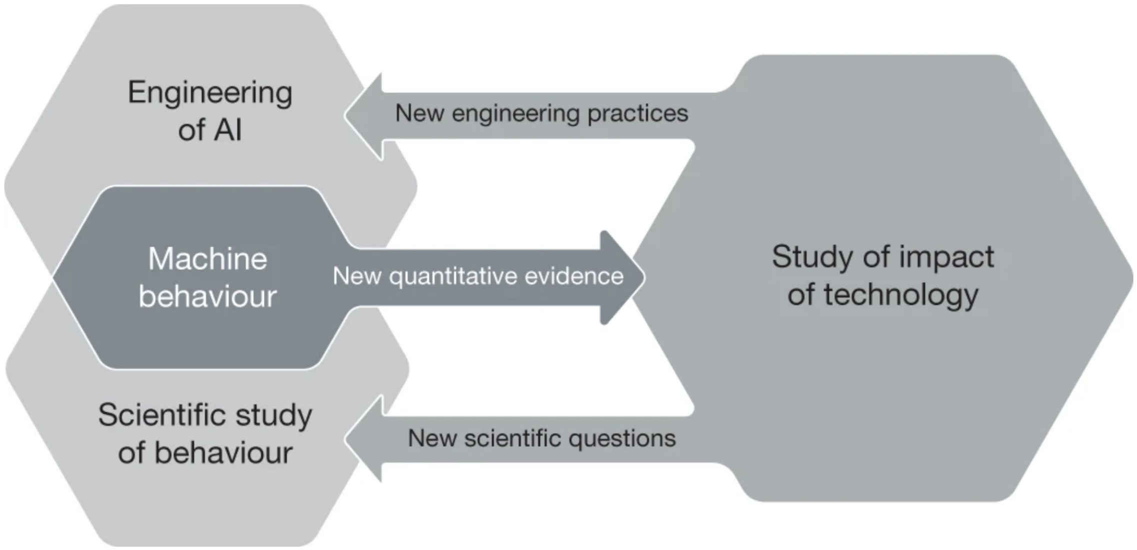
Conclusion

Review | Published: 24 April 2019

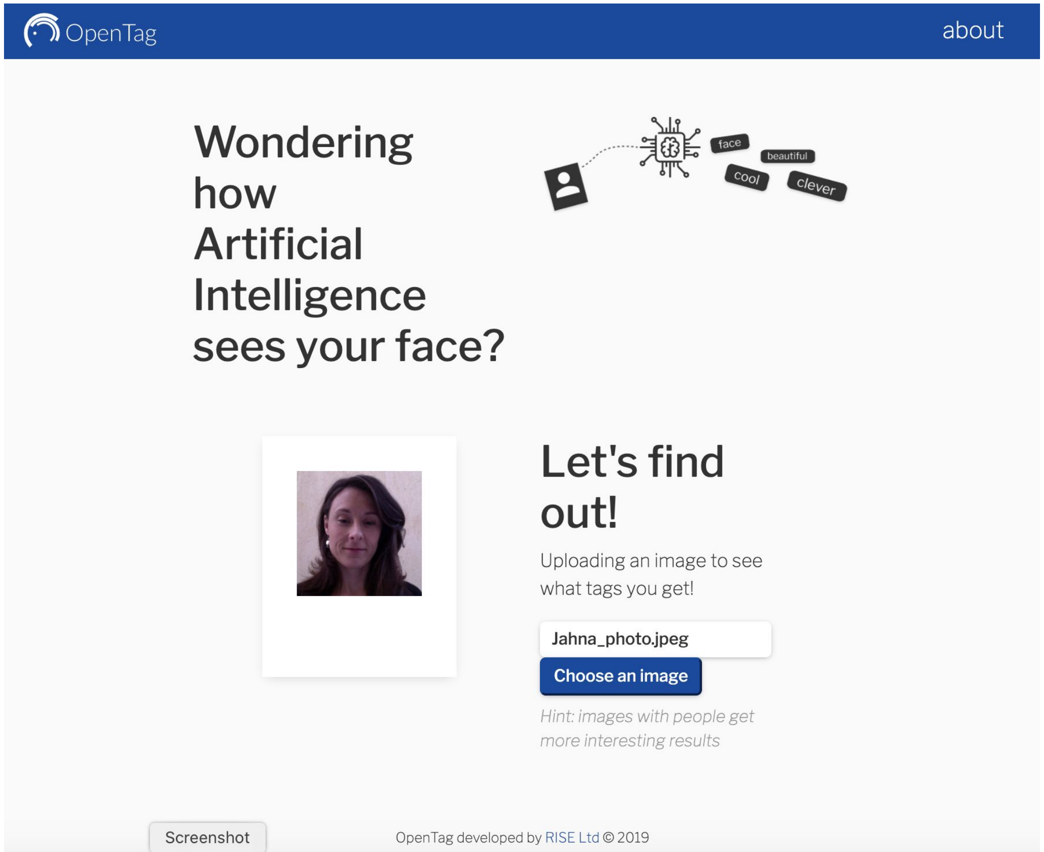
Machine behaviour

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



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).

Exam Question

Αρκετές μελέτες έδειξαν ότι υπάρχει ημεροληψία (bias) στα αποτελέσματα των εικόνων μιας μηχανής αναζήτησης, κυρίως ως προς το φύλο και εθνικότητα (gender and racial bias).

α) Ποιοι είναι οι κύριοι ενδιαφερόμενοι (stakeholders) που επηρεάζονται άμεσα ή έμμεσα από την ημεροληψία του συγκεκριμένου συστήματος;

β) Σε ποιο/α συστατικό/α του συστήματος διακρίνονται τα συγκεκριμένα είδη ημεροληψίας ;

γ) Ποιος είναι ο ρόλος του προγραμματιστή (developer) σχετικά με το μετριασμό της ημεροληψίας στη μηχανή αναζήτησης;