Code4Thought

Trust and Quality in the Era of Software 2.0

26/11/2020

Yiannis Kanellopoulos

Authority is increasingly expressed algorithmically

"Already today, 'truth' is defined by the top results of the Google search."

Yuval Noah Harari, "21 lessons for the 21st century"

Recent Headlines

MIT Technology Review

Artificial Intelligence Oct 25

A biased medical algorithm favored white people for health-care programs

Artificial Intelligence Dec 20

A US government study confirms most face recognition systems are racist

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

PUBLIC RELEASE: 9-APR-2015

Who's a CEO? Google image results can shift gender biases

UNIVERSITY OF WASHINGTON

SHARE

Harvard Business Review

TECHNOLOGY

When Algorithms Decide Whose Voices Will Be Heard

by Theodora (Theo) Lau and Uday Akkaraju

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO MORE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James \

Artificial Intelligence Nov 11

Apple Card is being investigated over claims it gives women lower credit limits

Artificial Intelligence

There's an easy way to make lending fairer for women. Trouble is, it's illegal.

Goldman Sachs defended itself in the Apple Card scandal by saying it did not consider gender when calculating creditworthiness. If it did, that could actually mitigate the problem.

by Karen Hao | a month ago

Technology as part of history

Microsoft Urged To Follow Amazon And IBM: Stop Selling Facial Recognition To Cops After George Floyd's Death

Thomas Brewster Forbes Staff

Cybersecurity

Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Technology

Microsoft won't sell police its facial-recognition technology, following similar moves by Amazon and IBM

Big tech companies back away from selling facial recognition to police. That's progress.

After IBM, Amazon, and Microsoft upend their facial recognition businesses, attention turns to federal lawmakers.

By Rebecca Heilweil | Updated Jun 11, 2020, 5:02pm EDT

What keeps us at night

- Our team has spent the better part of two decades analyzing and evaluating large scale software systems in order to help corporations address any potential risks and flaws related to them.
- By doing so we realised that the produced technology is the mirror of its organisation.
- At Code4Thought, we're turning all this expertise into a technology that will ensure AI/ML models are:
  - Fair,
  - Accountable,
  - Transparent.

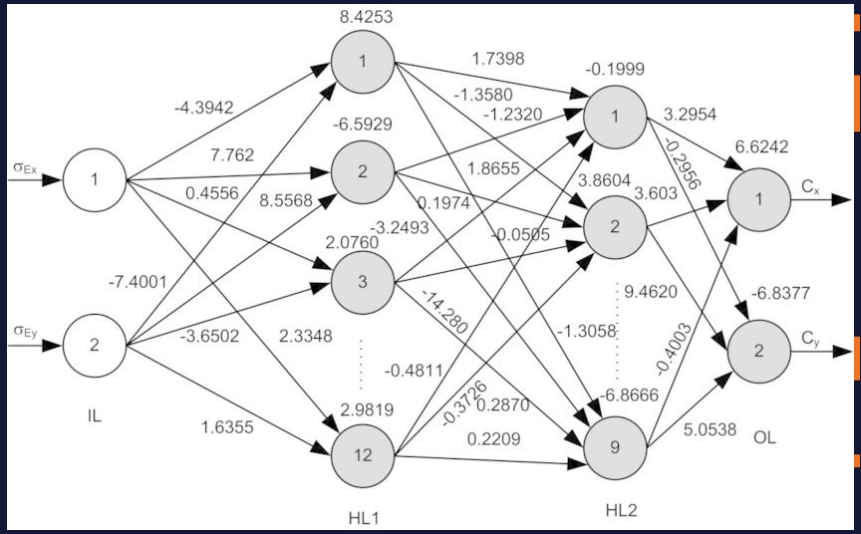
The software types

Deterministic (Code Driven)

```
In [7]:
down = 0
up = 100
for i in range(1,10):
    guessed_age = int((up+down)/2)
    answer = input('Are you ' + str(guessed_age) + " years old?")
    if answer == 'correct':
        print("Nice")
        break
    elif answer == 'less':
        up = guessed_age
    elif answer == 'more':
        down = guessed_age
    else:
        print('wrong answer')

Are you 50 years old?less
Are you 25 years old?more
Are you 37 years old?less
Are you 31 years old?less
Are you 28 years old?more
Are you 29 years old?correct
Nice
```

Probabilistic (Data Driven)*

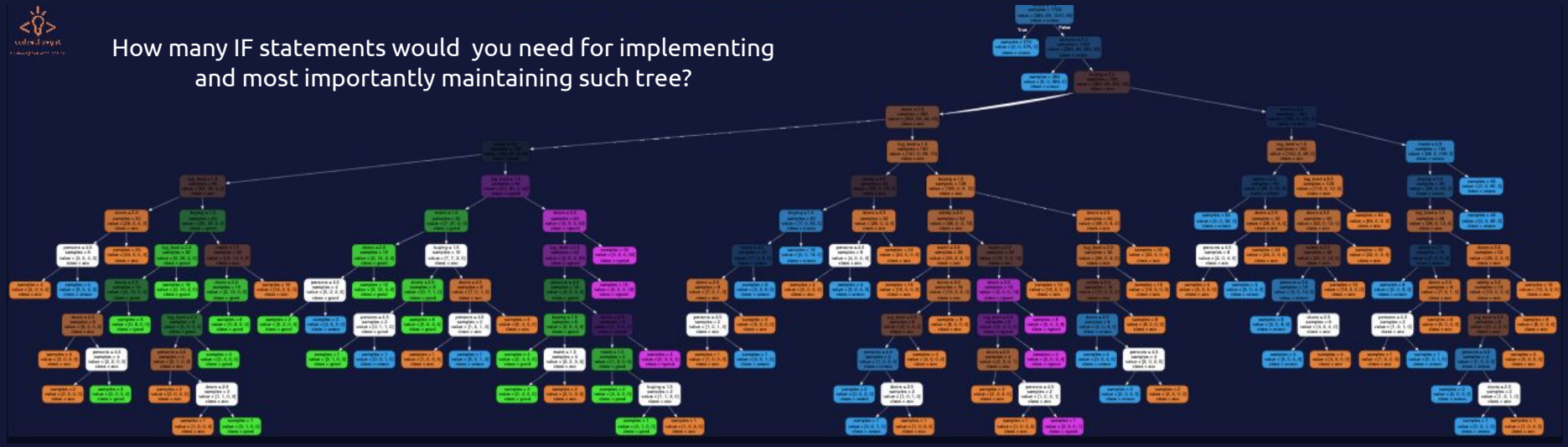


*

code4thought

Code-driven vs Data-driven

How many IF statements would you need for implementing and most importantly maintaining such tree?



From Software Quality to AI Behavior

	Code-Driven	Data-Driven
Existence of Industry Standards and Certifications	√	X
Formal Training and Professional Certifications	√	-
Methodologies, Tooling, Processes	√	-
Regulations, Legal Requirements	√	-

X Doesn't exist

√ Fully exist

\- Partially exist or enforced

Example of a Regulation: Principles of GDPR

Article 5 GDPR

(1) Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject:
- lawfulness
- **fairness**
- **transparency**
- ...

(2) “The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1”
- **accountability**

Challenges for a successful AI/ML implementation

- Choosing the right solution (i.e. suitable model, algorithm) for a given business problem,
- Creating proper training datasets (e.g. lack of labels, classes misrepresentation) for the models at hand,
- **Lack of trust to a model's results upon deployment.**

Challenges for building Trust

- Technical teams strive for accuracy and fast delivery and not so much for building trust.
- Accountability or Fairness are merely afterthoughts,
- When trust is imposed as a regulatory requirement (e.g. transparency) ad-hoc and one-off solutions are implemented.

Building Trust: (How to) use the F.A.T properties

- Be Simple but not simplistic,
- Be Transparent but selective,
- Use references/standards/check-lists.

F.Acc.T. checks as part of a ML pipeline

Automated Data Analysis from a Fairness perspective: helps mitigating the risk of statistical bias in the data

- Model Assessment against the FAccT principle
- Use interpretation methods for understanding model's behavior

Monitor Fairness and provide explanations constantly as defined in points #1 and #2

Multiplicity of fairness metrics

video

IBM Research Trusted AI

AI Fairness 360

link

Aequitas
Bias & Fairness Audit

link

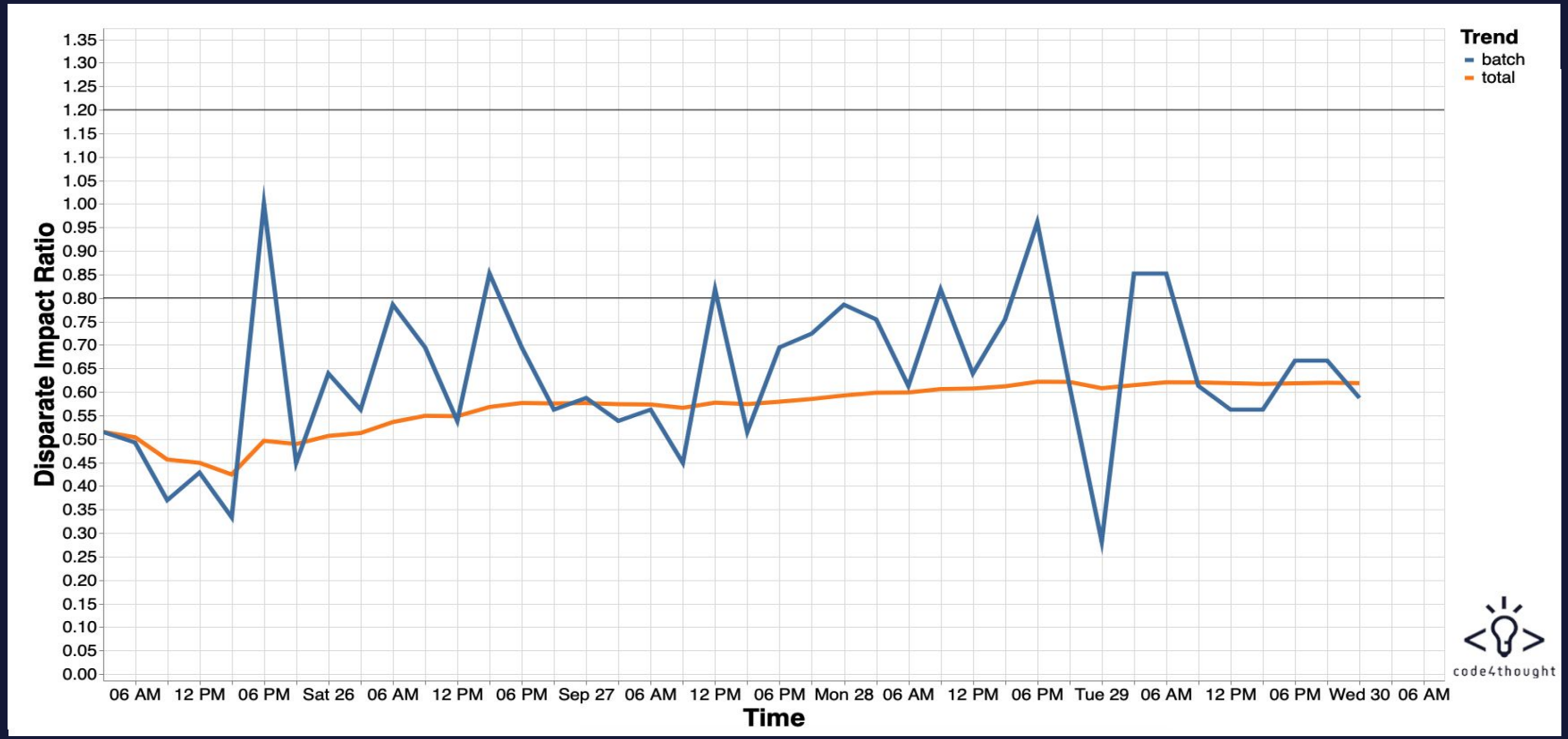
code4thought

At Code4thought we strive to:

- Pinpoint discriminatory behavior in algorithmic systems based on proportional disparities in their outcomes
  - adjacent to the *non-discimination* and *fair-balancing* concepts of GDPR

- **The metric we use**:
  - ***Disparate Impact Ratio***
  - connected with a well-known regulatory rule, the "*four-fifths-rule*" [EEOC in 29 C.F.R. § 1607.4(D)]

Fairness Analysis: Checking for Bias



The “four-fifths rule”

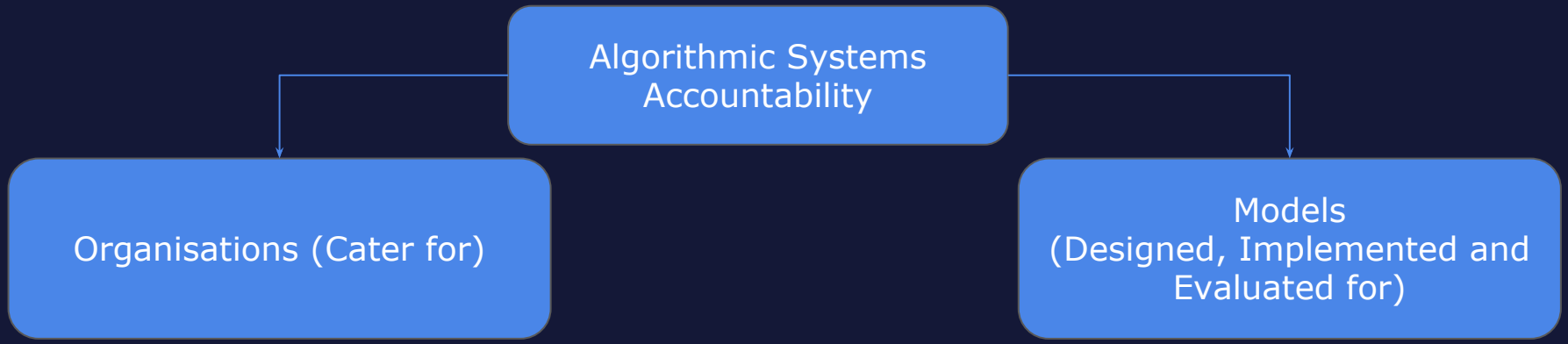
“a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or 80%) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as ***evidence of adverse impact****”*

EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (2018).

Examples of Legally recognized *sensitive* attributes

- **Race**
  (**USA:** Civil Rights Act of 1964, **EU**: Council Directive 2000/43/EC of 29 June 2000)
- **Sex**
  (**USA**: Equal Pay Act of 1963; Civil Rights Act of 1964, **EU**: European Convention on Human Rights Article 14)
- **Age**
  (**USA**: Age Discrimination in Employment Act of 1967, **EU**: Council Directive 2000/78/EC)
- **Religion, Color**
  (**USA**: Civil Rights Act of 1964, **EU**: Treaty of Amsterdam Article 13)
- **Familial Status**
  (**USA**: Civil Rights Act of 1968 Title VIII, **EU**: Equality Act 2010)
- **Disability Status**
  (**USA**: Rehabilitation Act of 1973 and Americans with Disabilities Act of 1990, **EU**: Equality Act 2010)
- **…**

Accountability Evaluation: Organisations + Models



Responsibility/Human Involvement
Explainability
Accuracy
Auditability
Fairness

Algorithmic Presence
Data
Algorithm Input
Performance Evaluation
Inferencing

Accountability Evaluation*: The value of checklists

Not priorities

Unsupervised model

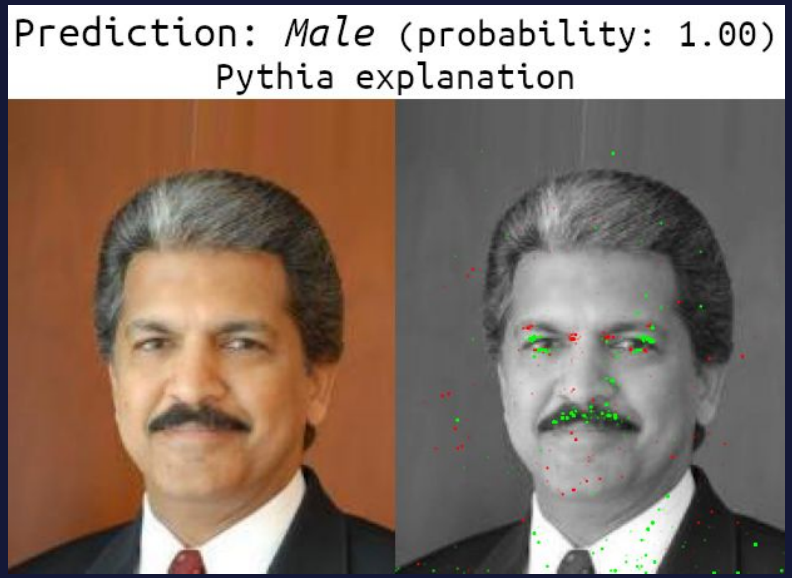
No annotations

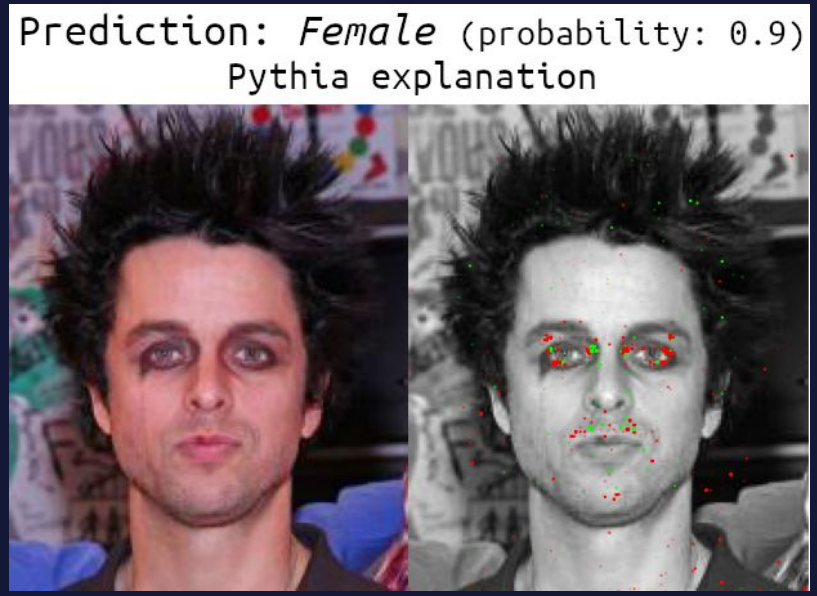
*Yiannis Kanellopoulos, "Accountability of Algorithmic Systems: How We Can Control What We Can't Exactly Measure" https://www.cutter.com/offer/accountability-algorithmic-systems-how-we-can-control-what-we-can't-exactly-measure Cutter Business Technology Journal, March 2019.

** Helen Tagiou, Yiannis Kanellopoulos, Christos Makris, Christos Aridas, "A tool supported framework for the Assessment of Algorithmic Accountability", in *International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2019.

code4thought

Explanations can scrutinize algorithms for systematic bias





This gender-classifier predicts based on the the appearance of moustache and make-up in the eyes
(green pixels contribute towards predicting *Male* and red pixels towards predicting *Female*)

* A. Messalas, Y. Kanellopoulos, C. Makris, "Model-Agnostic Interpretability with Shapley values," in *International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2019

Our case study for PyThia: Twitter AI bias controversy

CODE4THOUGHT

Home About Solutions News & Events Blog Contact

Is Twitter biased against BIPOC? Maybe it's not what you think it is.

October 2, 2020 Bias Explainability Fairness Machine Learning Twitter XAI

Methodology

Specialized dataset containing images of faces of different racial groups, which is balanced for all groups

4,009 photo-collages (300 tweets/3 hours)

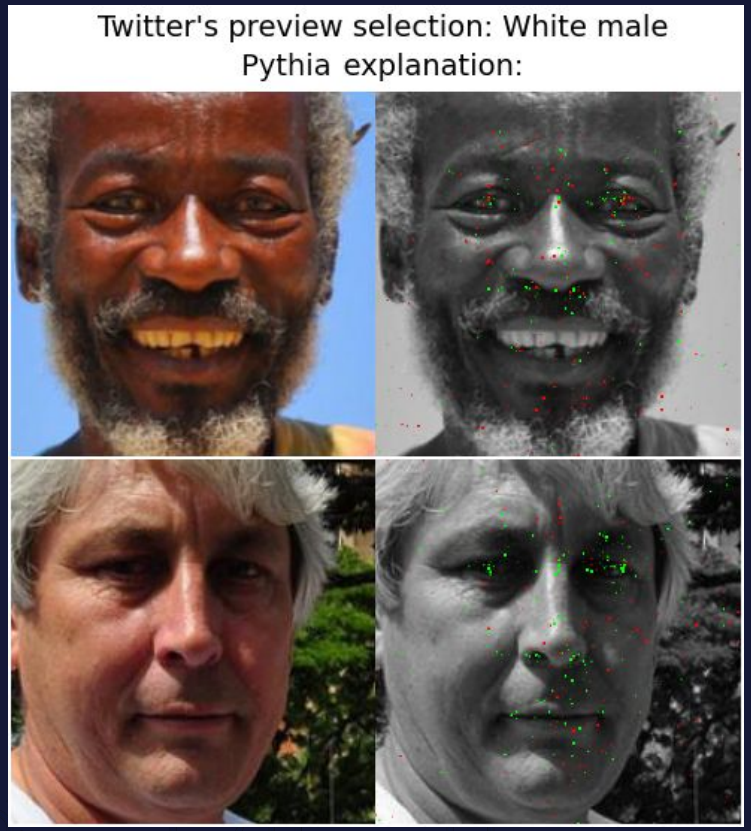
Data: 8,018 single photos
Label: selected in preview?

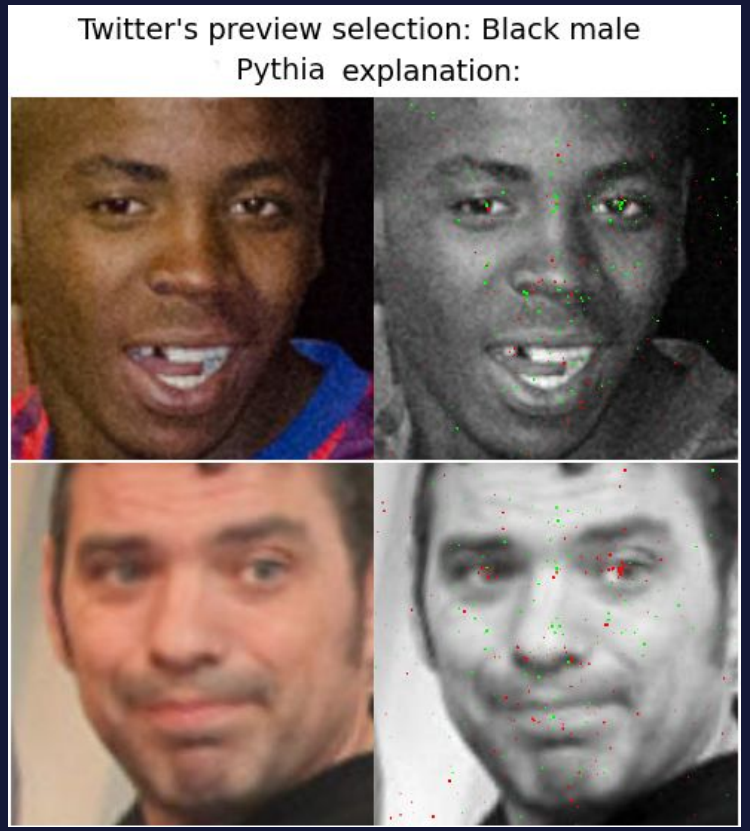
Data Collection → Data Preparation → Upload to Twitter → Manual Labeling → Fairness Analysis → Transparency Analysis

Preview label: ‘Black male’, ’White male’

0 1

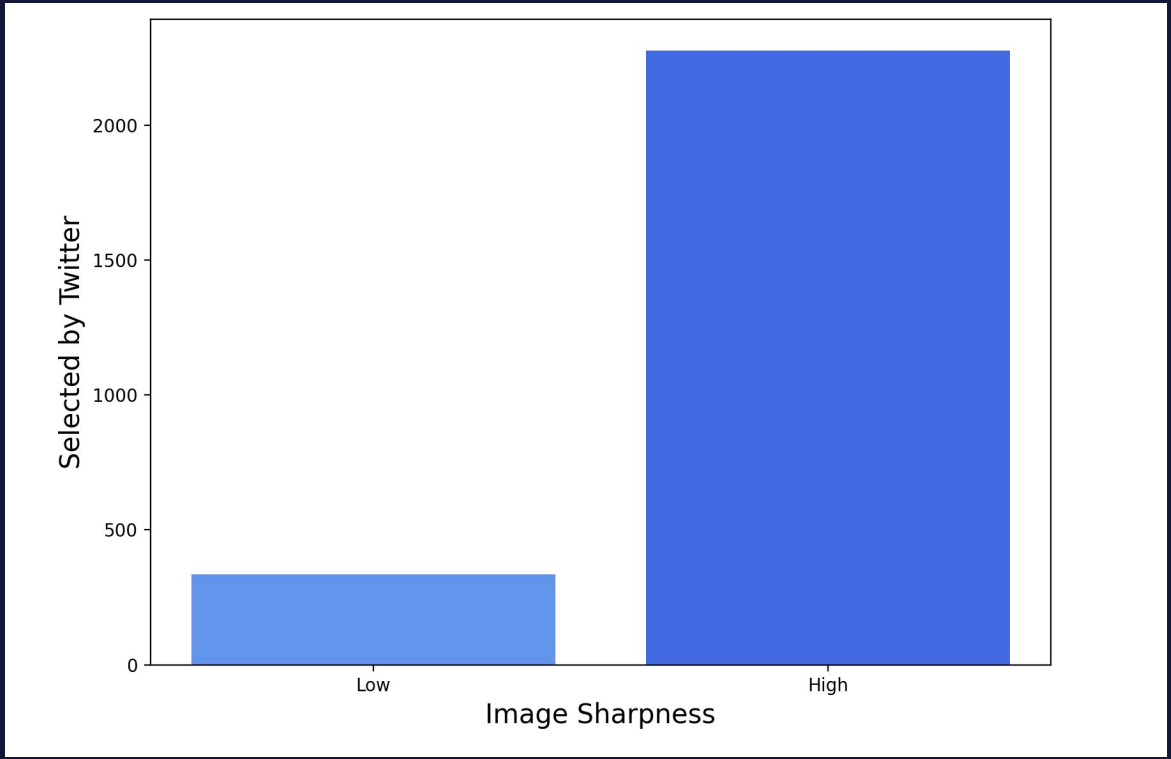
Transparency Analysis*: Open up the black box





* A. Messalas, Y. Kanellopoulos, C. Makris, "Model-Agnostic Interpretability with Shapley values," in *International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2019

Search beyond the obvious features and explain



Conclusions

- Software 2.0 can impact our lives at scale and at unprecedented speed. Trust is paramount for its further adoption.

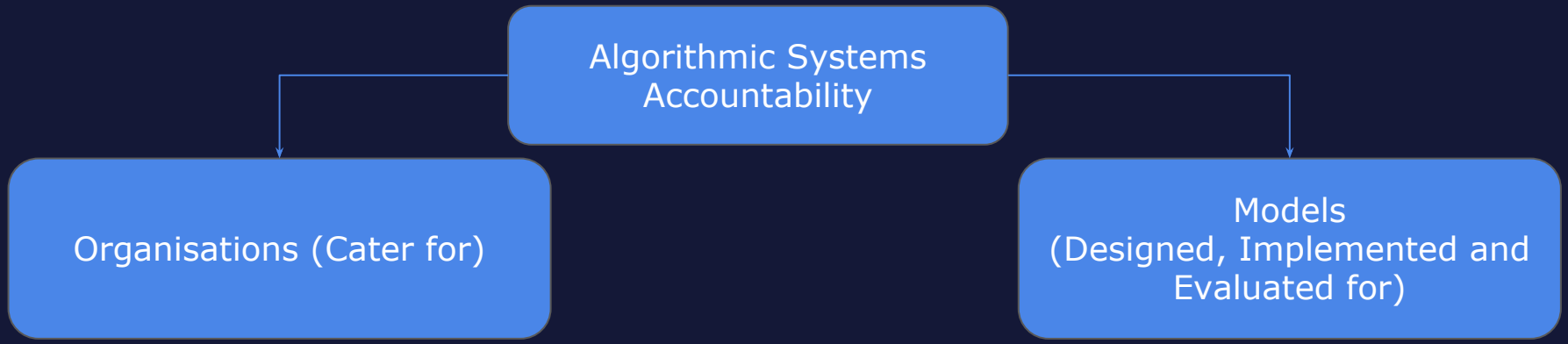
- We can use technology to control technology. The crucial factor here is humans and how they can be augmented.

- Building trust over an AI/ML models is a great opportunity for jobs' creation. It can be the catalyst for the jobs of the future.

Stay in touch

- **See:** xai.code4thought.eu
- **Contact:** yiannis@code4thought.eu
- **Follow:** @code4thought.eu

Accountability Evaluation: Organisations + Models



Responsibility/Human Involvement
Explainability
Accuracy
Auditability
Fairness

Algorithmic Presence
Data
Algorithm Input
Performance Evaluation
Inferencing

Accountability Evaluation*: The value of checklists

Not priorities

Unsupervised model

No annotations

*Yiannis Kanellopoulos, "Accountability of Algorithmic Systems: How We Can Control What We Can't Exactly Measure" https://www.cutter.com/offer/accountability-algorithmic-systems-how-we-can-control-what-we-can't-exactly-measure Cutter Business Technology Journal, March 2019.

** Helen Tagiou, Yiannis Kanellopoulos, Christos Makris, Christos Aridas, "A tool supported framework for the Assessment of Algorithmic Accountability", in *International Conference on Information, Intelligence, Systems and Applications (IISA)*, July 2019.

code4thought