18/11/2022

Fairness in Machine Learning I

Ruta Binkyte

aivancity
SCHOOL FOR
TECHNOLOGY, BUSINESS & SOCIETY
PARIS-CACHAN

A bit of House keeping

Ruta Binkyte – *November 2022*

Introducing the course and the instructors

Academic Interests

Trustworthy AI: Fairness, Privacy, Explainability, Causality

Experience

Ba and Ma in Cultural Anthropology (Vilnius University)
MSc in Data Science (The University of Edinburgh)
PhD in AI Ethics (École Polytechnique, Inria. *Ongoing)*

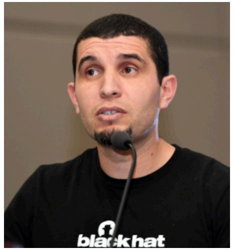
Ruta Binkyte

binkyte-sadauskiene@aivancity.ai

http://www.lix.polytechnique.fr/Labo/Ruta.BINKYTE-SADAUSKIENE/

Sami Zhioua

http://www.lix.polytechnique.fr/Labo/Sami.ZHIOUA/



Introduction:
Why FAIR Machine Learning is important?

Theory

- Three known cases of biased ML: Real world applications where fairness is critical, Sources of bias
- Fairness concepts: Equality of treatment, Equality of Outcome
- Legislation: GDPR, Anti-discrimination, US equivalent

Practice

- Analysing and discussing the ethical risks of a given ML scenario (group work)
- Proposing an action plan for the ethical implementation and deployment of the ML solution (group work)
- Student group presentations on action plan for the ethical implementation and deployment of the ML solution

Expected Outcome

To be able to critically reason applying fairness concepts to a given scenario

Part 2
12/14

Achieving Fairness:
Measuring and mitigating bias

Theory

- Fairness metrics: Group fairness, Individual fairness
- Mitigating bias: Pre-processing, In-processing, Post processing
- Fairness in the context of Trustworthy Machine Learning: synergies and tensions

Practice

- Fairness in business practice (guest speaker)
- Measuring bias practical exercises
- Mitigating bias practical exercises

Expected Outcome

Know the fairness metrics, mitigating approaches. Be able to choose and implement the most plausible technique for a given scenario.

Ruta Binkyte – *November 2022*

Part 2
12/16

Causality framework
for fairness in machine learning

Theory

- The problem with statistical notions: Observational vs experimental analysis, Simpson's paradox
- Basic notions of causality: Causal graph, do-operator, Counterfactuals
- Causal fairness metrics
- Causal fairness in practice: Causal graph availability, Identifiability, Estimation

Practice

- Causal Fairness tools and libraries
- Measuring bias using basic causal metrics practical exercises
- Comparison with statistical metrics

Expected Outcome

Understand the difference between the statistical and causal approach.
Be able to discuss arguments for causal approach and implement it in practice.

Ruta Binkyte – *November 2022*

aivancity
PARIS-CACHAN

Course Materials, Exams and Evaluation

Assessment name	Assessment method	Assessment description	Final weight (%)
Day 1 Evaluation test	MCQ	Individual tests based on in-class material	30
Day 2 Evaluation test	MCQ	Individual tests based on in-class material	
Day 3 Evaluation test	MCQ	Individual tests based on in-class material	
Lab 2	Jupypter Notebook	Individual practical work	20
Lab 3	Jupypter Notebook	Individual practical work	
Team activities	Written group assignment	Team assignments for applying course concepts and solving problems	20

Assessment name	Assessment method	Assessment description	Final weight (%)
Final exam/Final project	Written exam/Test	Individual exam evaluating course concepts	30

Required Readings:

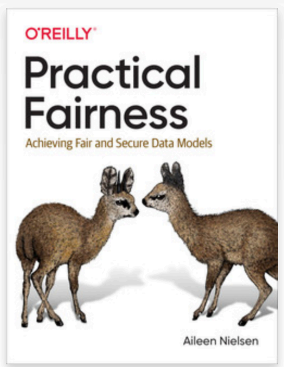
Machine learning fairness notions: Bridging the gap with real-world applications

Karima Makhlouf, Sami Zhioua, Catuscia Palamidessi

Fairness emerged as an important requirement to guarantee that Machine Learning (ML) predictive systems do not discriminate against specific individuals or entire sub-populations, in particular, minorities. Given the inherent subjectivity of viewing the concept of fairness, several notions of fairness have been introduced in the literature. This paper is a survey that illustrates the subtleties between fairness notions through a large number of examples and scenarios. In addition, unlike other surveys in the literature, it addresses the question of: which notion of fairness is most suited to a given real-world scenario and why? Our attempt to answer this question consists in (1) identifying the set of fairness-related characteristics of the real-world scenario at hand, (2) analyzing the behavior of each fairness notion, and then (3) fitting these two elements to recommend the most suitable fairness notion in every specific setup. The results are summarized in a decision diagram that can be used by practitioners and policymakers to navigate the relatively large catalog of ML.

https://arxiv.org/abs/2006.16745

Optional:

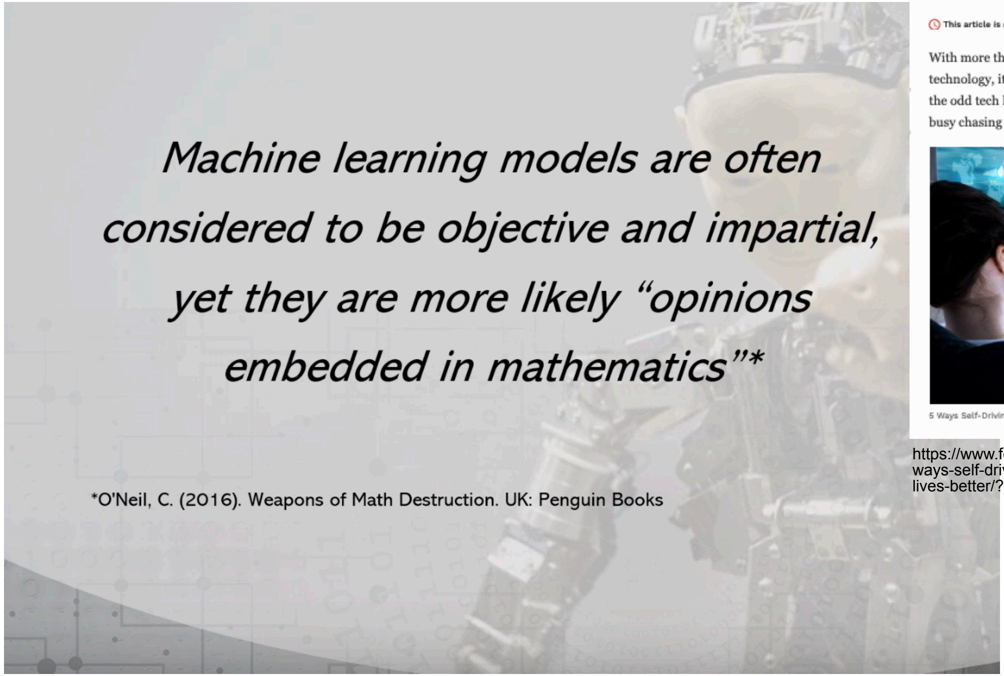


Introduction: Why Fairness?

Ruta Binkyte – *November 2022*

aivancity
PARIS-CACHAN

AI in the Real World



*Machine learning models are often considered to be objective and impartial, yet they are more likely "opinions embedded in mathematics"**

*O'Neil, C. (2016). Weapons of Math Destruction. UK: Penguin Books

ENTERPRISE TECH

5 Ways Self-Driving Cars Could Make Our World (And Our Lives) Better

Bernard Marr Contributor

Follow

Jul 17, 2020, 12:24am EDT

New! Follow this author to stay notified about their latest stories. Got it!

This article is more than 2 years old.

With more than 40 companies actively investing in autonomous vehicle technology, it's fair to say that most traditional car manufacturers – plus the odd tech heavyweight, like Google parent company Alphabet – are busy chasing the self-driving car dream.

5 Ways Self-Driving Cars Could Make Our World (And Our

https://www.forbes.com/sites/bernarc ways-self-driving-cars-could-make-o lives-better/?sh=3706883442a3

Le dimanche, si j'ai besoin de mo j'ai intérêt à le croiser au ma

How Can Healthcare An Alternative Medicine Sa Artificial Intelligence An Reality

Today we already see how healthtech startups and compa the industry of tomorrow

By Shishir Jajoo

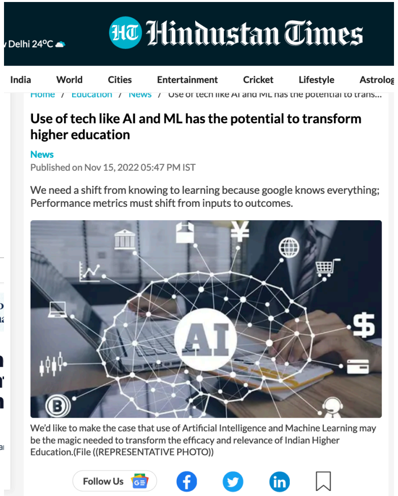
November 15, 2022

Opinions expressed by Entrepreneur contributors are their own.

You're reading Entrepreneur India, an international franchise of Entrepreneur Media.

A colorful combination of companies—aerospace, retail, gaming, security, and telecommunications—have in recent years discovered the power of pairing artificial intelligence (AI) and virtual reality (VR) to help alleviate the complexities within their industries.

https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/



Delhi 24°C

HT Hindustan Times

India World Cities Entertainment Cricket Lifestyle Astrolog

Home / Education / News / Use of tech like AI and ML has the potential to trans...

Use of tech like AI and ML has the potential to transform higher education

News

Published on Nov 15, 2022 05:47 PM IST

We need a shift from knowing to learning because google knows everything; Performance metrics must shift from inputs to outcomes.

We'd like to make the case that use of Artificial Intelligence and Machine Learning may be the magic needed to transform the efficacy and relevance of Indian Higher Education.(File ((REPRESENTATIVE PHOTO))

Follow Us

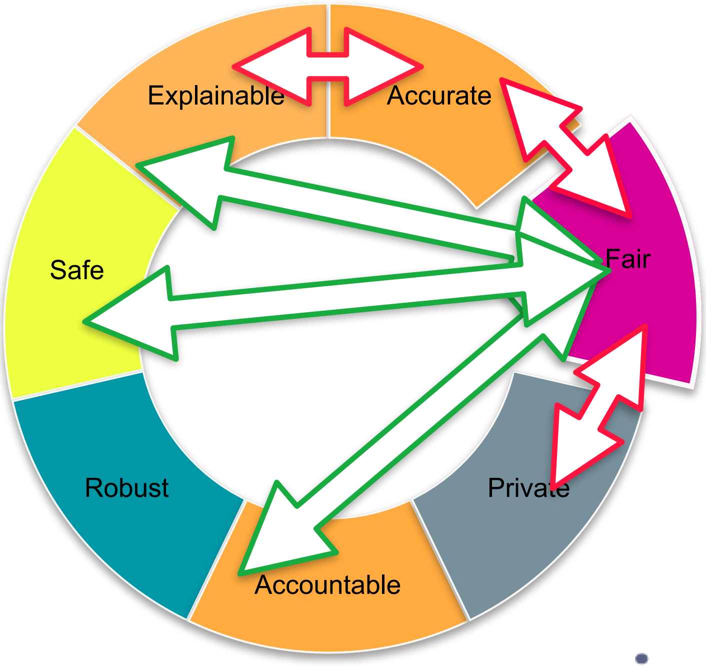
https://www.hindustantimes.com/education/news/use-of-tech-like-ai-and-ml-has-the-potential-to-transform-higher-education-101668513305169.html

Ruta Binkyte – *November 2022*

aivancity PARIS-CACHAN

Trustworthy AI

- Legal Reasons
- User Trust
- Intertwined relations



Three Cases of Biased ML

Ruta Binkyte – *November 2022*

Case 1

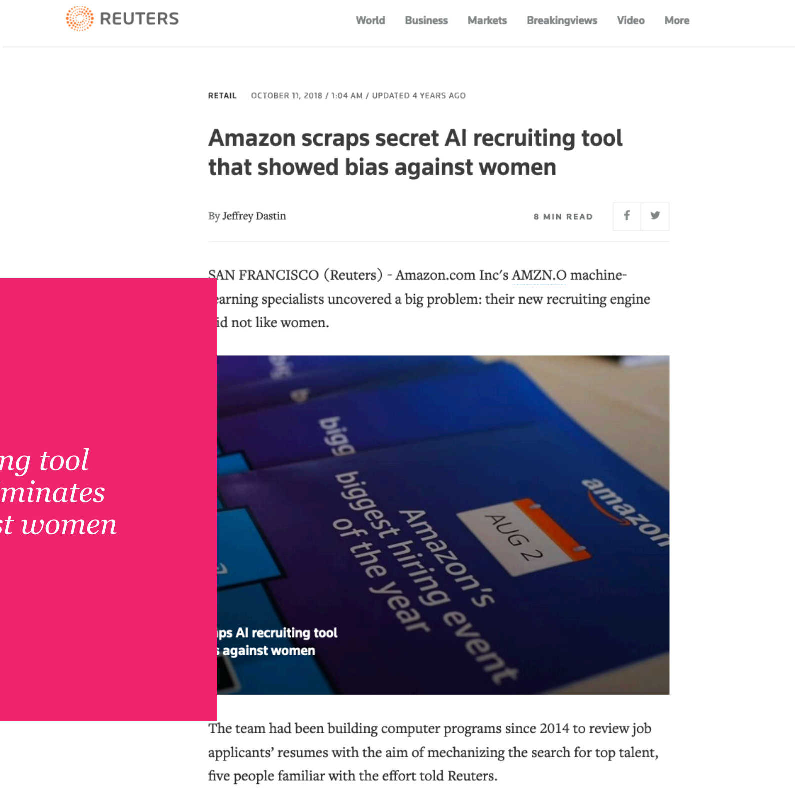
Gender Bias

In hiring

What happened? The algorithm built to pre-filter tech job candidates discriminated agains women.

Why? Not enough women in tech to train the algorithm

Hiring tool discriminates against women



REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-
earning specialists uncovered a big problem: their new recruiting engine
id not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Case 2

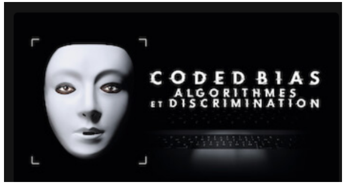
Racial Bias

In Face Recognition

What happened? The algorithms for face recognition have much lower accuracy for darker female faces.

Why? Black women underrepresented in the training data.

Darker female faces are harder to recognise with computer vision algorithms



ARS ELECTRONICA | OUT OF THE BOX | POSTCITY | PROGRAMM

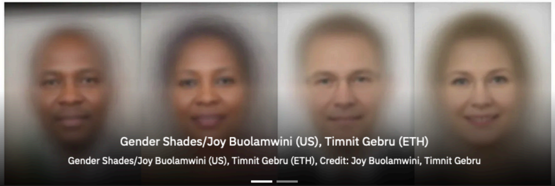
Gender Shades

Joy Buolamwini (US), Timnit Gebru (ETH)

POSTCITY

Joy Buolamwini and Timnit Gebru investigated the bias of AI facial recognition programs. The study reveals that popular applications that are already part of the programming display obvious discrimination on the basis of gender or skin color. One reason for the unfair results can be found in erroneous or incomplete data sets on which the program is being trained. In things like medical applications, this can be a problem: simple convolutional neural nets f detecting melanoma (malignant skin changes) as experts are.

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Gender Shades/Joy Buolamwini (US), Timnit Gebru (ETH)

Gender Shades/Joy Buolamwini (US), Timnit Gebru (ETH), Credit: Joy Buolamwini, Timnit Gebru

However, skin color information is crucial to this process. That's why both of the researchers created a new benchmark data set, which means new criteria for comparison. It contains the data of 1,270 parliamentarians from three African and three European countries. Thus Buolamwini and Gebru have created the first training data set that contains all skin color types, while at the same time being able to test facial recognition of gender.

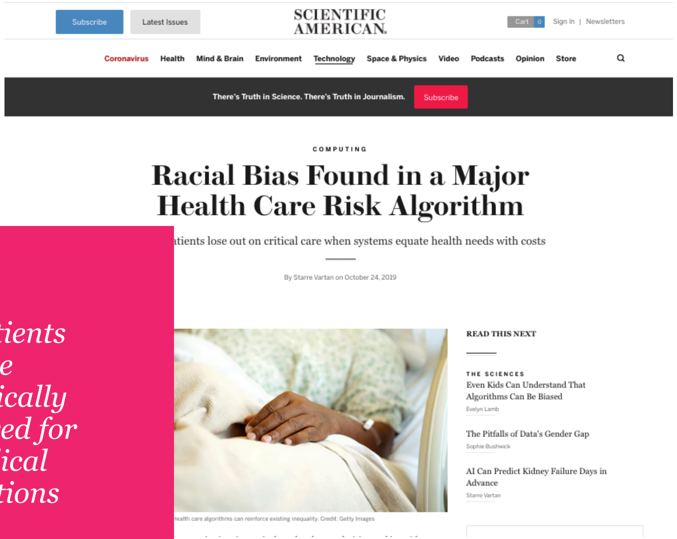
Case 3

Racial Bias

In healthcare AI

What happened? The algorithm built to predict the need for medical interventions (sickness) would give lower score for black patients who where the same or more sick than the white ones with the same score.

Why? The proxy used for “sickness” was healthcare spending, which correlated with lower income



Subscribe | Latest Issues | SCIENTIFIC AMERICAN | Cart 0 | Sign In | Newsletters

Coronavirus | Health | Mind & Brain | Environment | Technology | Space & Physics | Video | Podcasts | Opinion | Store

There's Truth in Science. There's Truth in Journalism. Subscribe

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

...tients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

...ealth care algorithms can reinforce existing inequality. Credit: Getty Images

READ THIS NEXT

THE SCIENCES
Even Kids Can Understand That Algorithms Can Be Biased
Evelyn Lamb

The Pitfalls of Data's Gender Gap
Sophie Bushwick

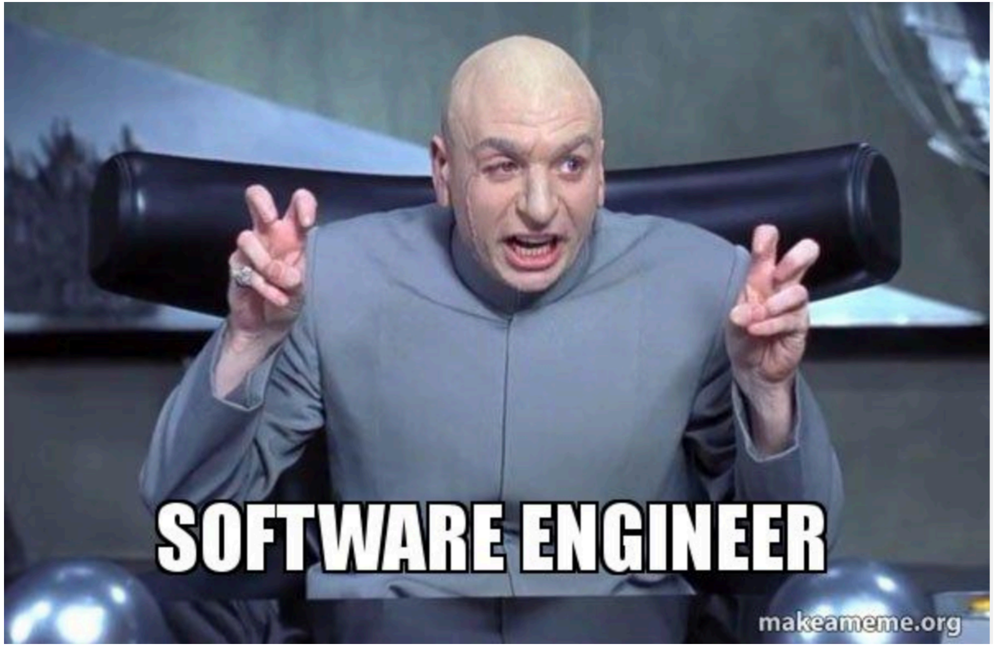
AI Can Predict Kidney Failure Days in Advance
Starre Vartan

Black Patients where systematically underscored for the medical interventions

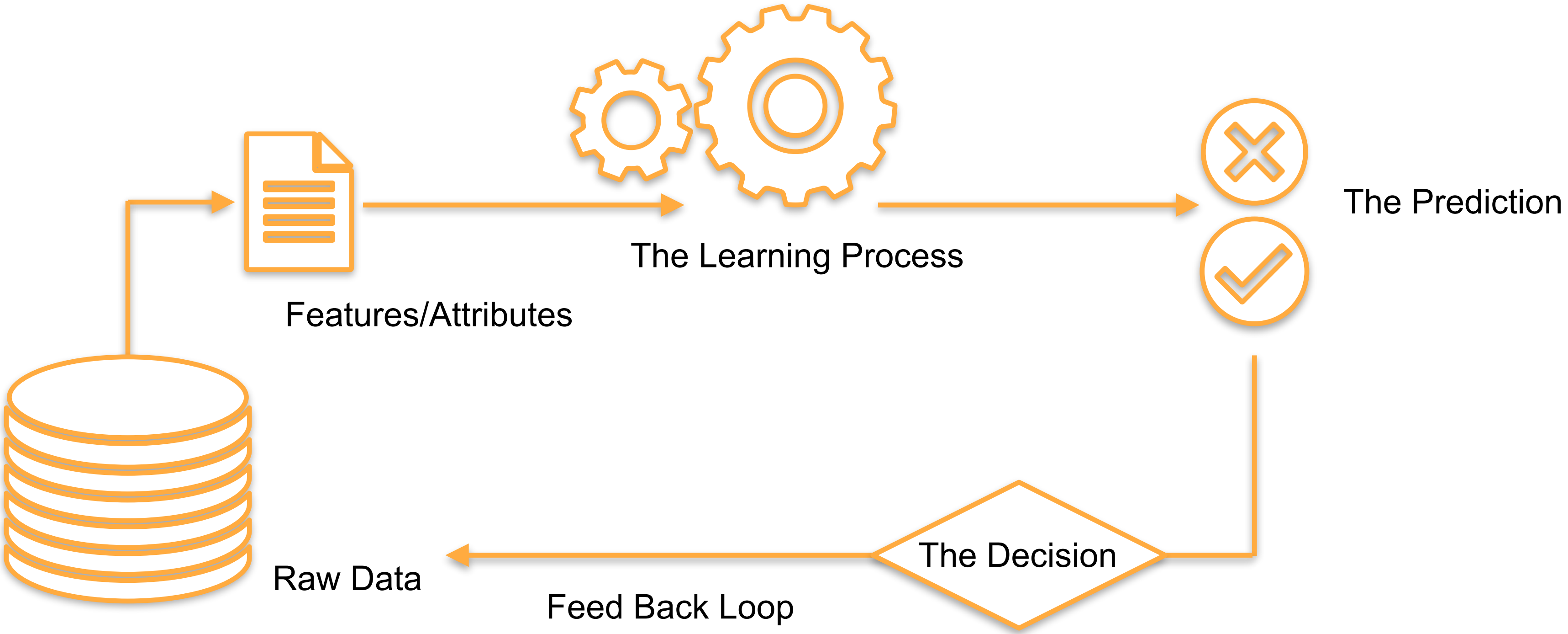
Sources of Bias

Ruta Binkyte – *November 2022*

Is the algorithmic discrimination intentional?



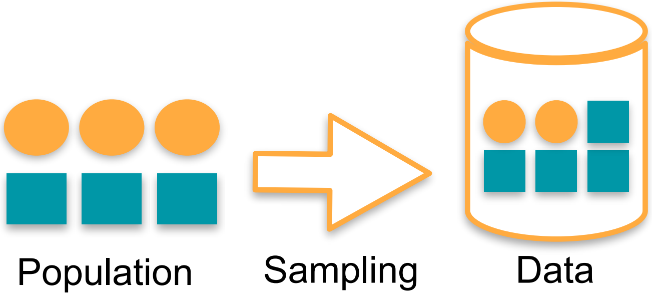
Machine Learning Process



Bias in the Data

Sample Bias

- The sample is not representative of population
- The algorithm will misclassify the minority group



Historical Bias

- The sample is representative of population, but contains historical biases
- The algorithm will learn historical discrimination

https://www.britannica.com/topic/racial-segregation

https://en.wikipedia.org/wiki/Women%27s_rights

Examples of historical biases:

- Lower income among women and ethnic/racial minority groups
- Gender and racial gaps in SAT scores
- Car safety standards built according to male body types
- Lower representation in the data and nunfavouring technical calibration of cameras for darker skin

https://history.wustl.edu/news/how-black-death-made-life-better

Does underrepresentation bias imply historical bias?

Not necessarily!

https://www.ft.com/content/7c32c7a8-172e-11ea-9ee4-11f260415385

https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

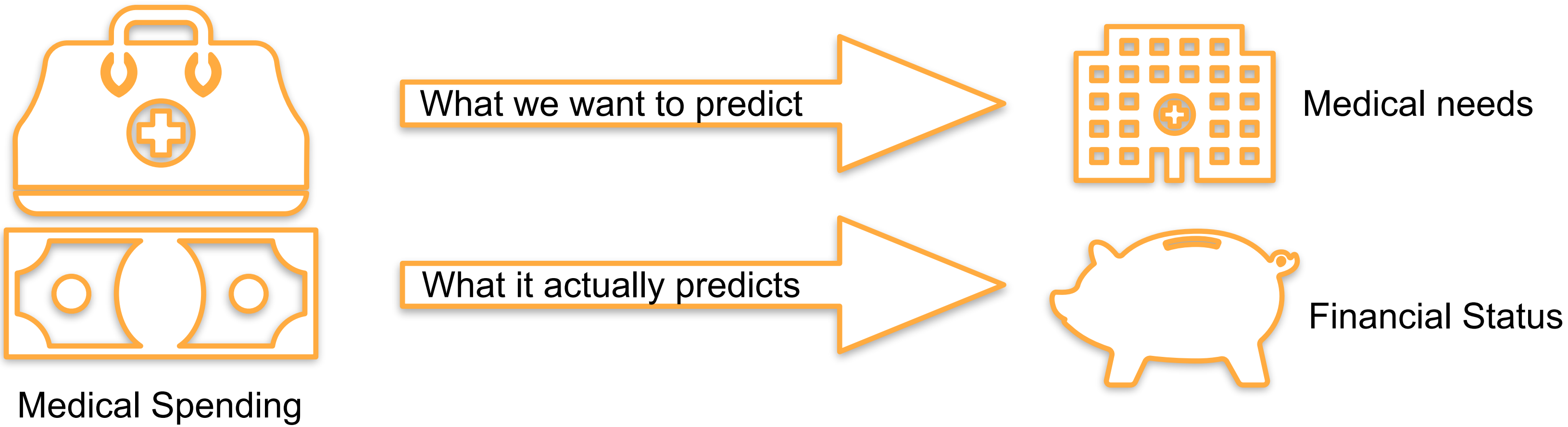
Bias in Feature Selection

A reminder on features used for ML models

Y - The label in the data
$\hat{Y}$ - The prediction
$X_1 \ldots X_n$ - The attributes (Features)
S - The sensitive attribute (also known as A) such as gender, race, age, sexual orientation etc.
A Proxy - An attribute that correlates with other feature we want to use or predict. When it is correlated with the sensitive attribute and used in the prediction we call it proxy discrimination. For example, medical spending correlates with race.

Bias in Feature Selection

It's not only unfair, it is also inaccurate!



Learning from Biased Data: Finding Patterns

Experience	Education	Gender	Hiring Decision
>5	BSc	F	✗
<5	MSc	M	✓
<5	MSc	F	✗
>5	BSc	M	✓
<5	MSc	F	✗
>5	MSc	F	✓
<5	BSc	M	✓
>5	BSc	M	✓

✓ |>5 = 3/4

✓ |<5 = 2/4

✓ | BSc = 2/4

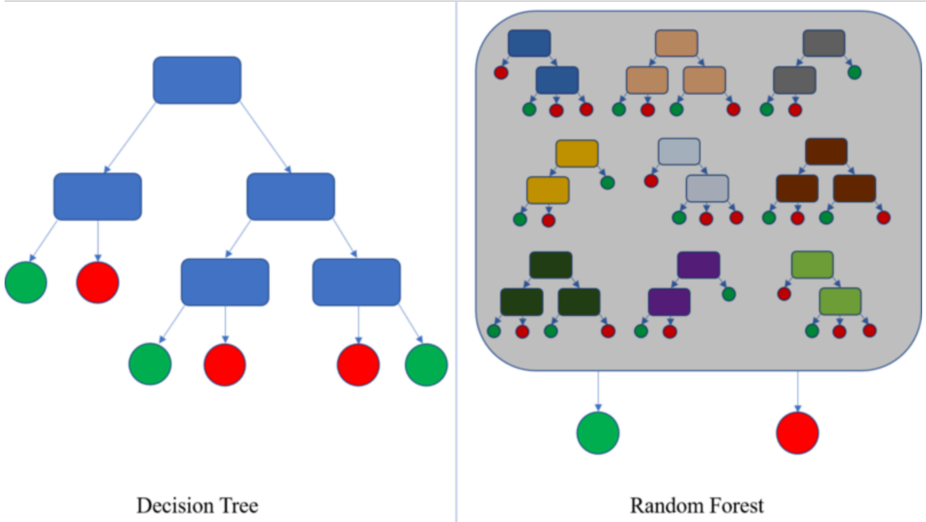
✓ | MSc = 2/4

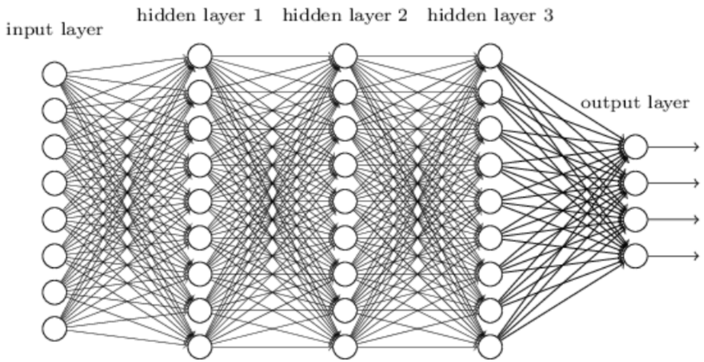
✓ | M = 4/4

✓ | F = 1/4

What would be most the predictive feature?

Learning from Biased Data: Complexity



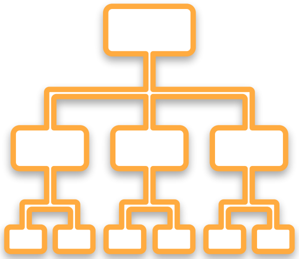


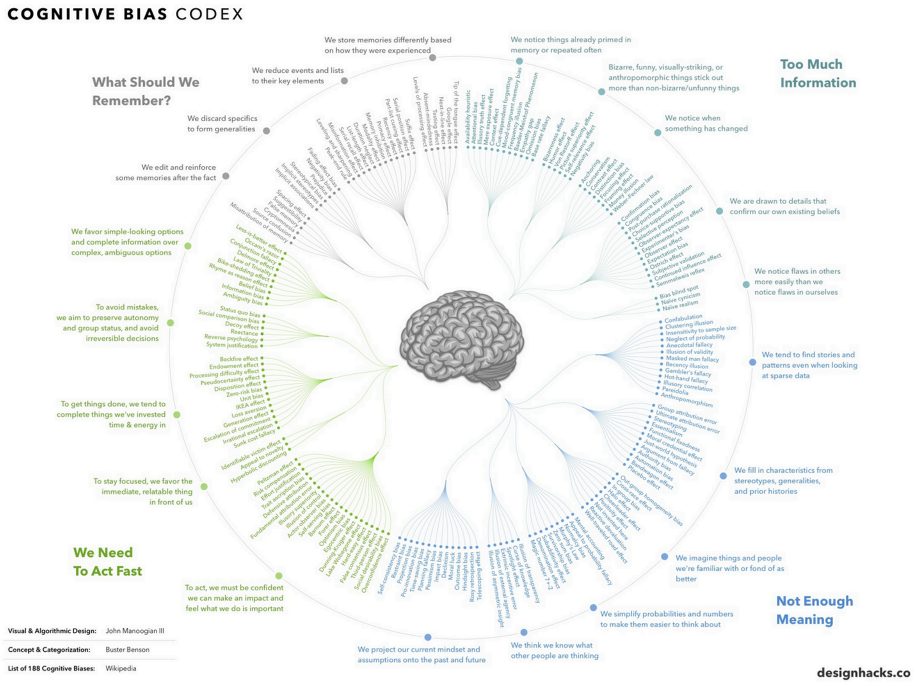
Dang, Q. V., & Ignat, C. L. (2016). Quality assessment of wikipedia articles: A deep learning approach by quang vinh dang and claudia-lavinia ignat with martin vesely as coordinator. *ACM SIGWEB Newsletter*, (Autumn), 1-6.

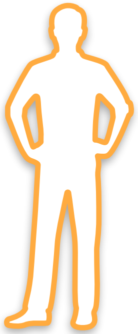
https://fr.m.wikipedia.org/wiki/Fichier:Decision_Tree_vs._Random_Forest.png

https://hal.archives-ouvertes.fr/hal-01393227/file/sigweb_newsletter_latex.pdf

Bias in the Prediction and Decision



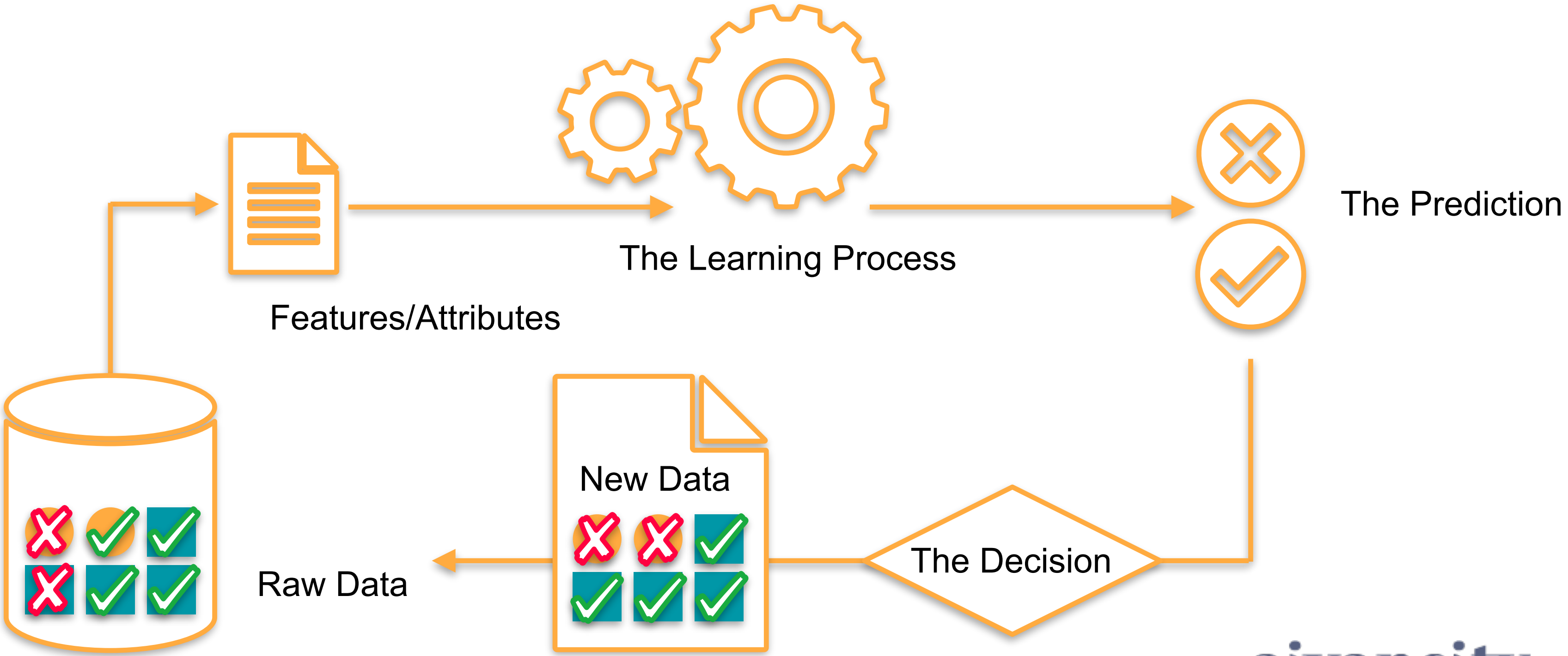




https://www.visualcapitalist.com/wp-content/uploads/2021/08/all-188-cognitive-biases.html

Go back to: VISUAL CAPITALIST

The Feed Back Loop



Ruta Binkyte – *November 2022*

https://www.visualcapitalist.com/wp-content/uploads/2021/08/all-188-cognitive-biases.html

aivancity
PARIS-CACHAN

If you are interested to read more

arXiv > cs > arXiv:1901.10002

Search...

Help | Advanced S

Computer Science > Machine Learning

[Submitted on 28 Jan 2019 (v1), last revised 1 Dec 2021 (this version, v5)]

A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh, John V. Guttag

As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this paper, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. In doing so, we aim to facilitate more productive and precise communication around these issues, as well as more direct, application-grounded ways to mitigate them.

Subjects:	**Machine Learning (cs.LG)**; Machine Learning (stat.ML)
Cite as:	arXiv:1901.10002 **[cs.LG]**
	(or arXiv:1901.10002v5 **[cs.LG]** for this version)
	https://doi.org/10.48550/arXiv.1901.10002
Journal reference:	EAAMO 2021: Equity and Access in Algorithms, Mechanisms, and Optimization
Related DOI:	https://doi.org/10.1145/3465416.3483305

https://arxiv.org/abs/1901.10002

Sources of Bias: *Summary*

- The algorithmic discrimination is most often unintentional
- Data is the main source of unfair bias.
- The most common biases in the data are Underrepresentation Bias (Data is not representative of the population) and Historical Bias (The data reflects the historical discrimination)
- The bias transmitted through data can be amplified in other stages of machine learning cycle
- The Algorithm can be more or less able to pick up subtle correlations and proxies, also be more or less explainable.
- The decision maker can be biased to blindly trust the algorithm or confirm pre-existing prejudices.
- Most Importantly: algorithmic predictions can scale the bias by feeding it back to the data.

Approaches to Fairness

Ruta Binkyte – *November 2022*

Approaches to Fairness

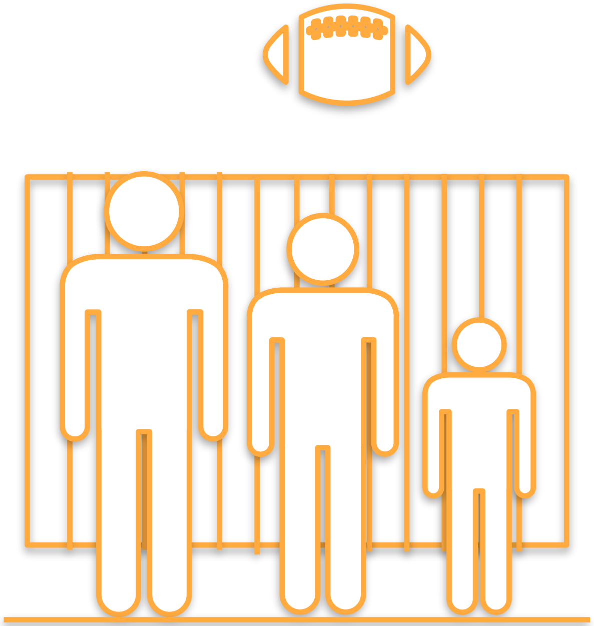
Equality and Equity

Equality

- Everyone gets equal treatment

Equity

- Everyone gets according to the needs or merits



Approaches to Fairness

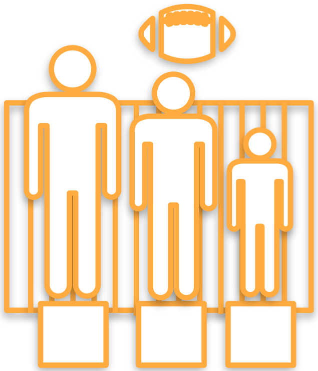
Equality and Equity

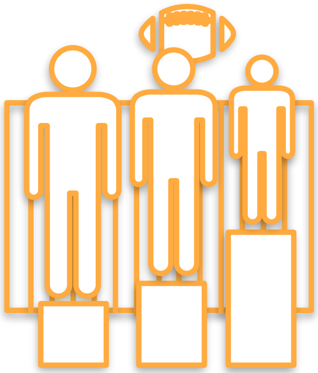
Equality of treatment

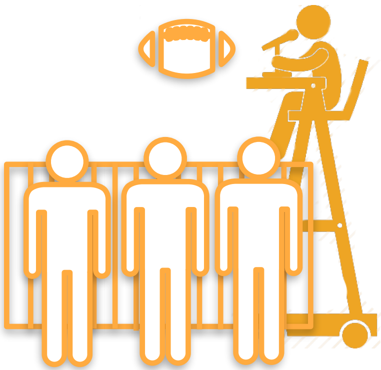
Equity of gain

Equity of loss

Equity by merit

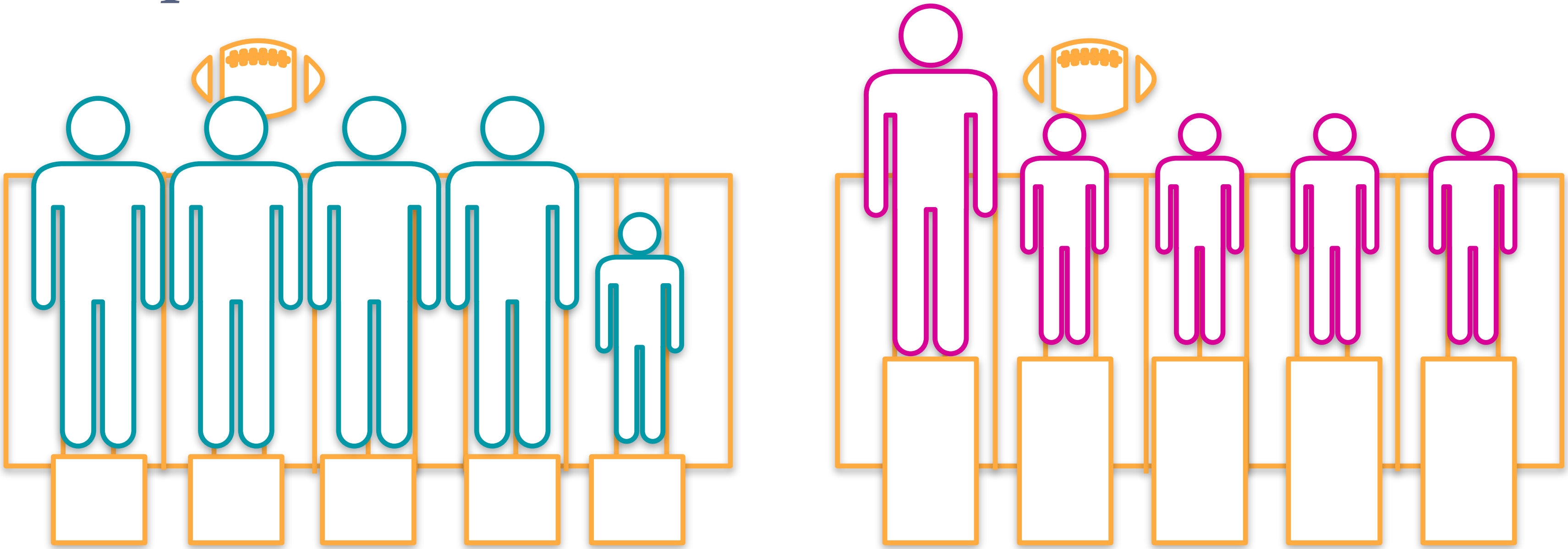






Approaches to Fairness

Groups and Individuals

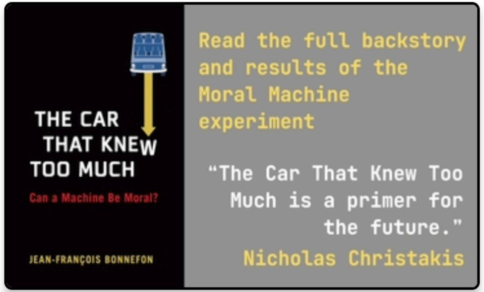


Ruta Binkyte – *November 2022*

aivancity
PARIS-CACHAN

MORAL MACHINE

Home | Judge | Classic | Design | Browse | About | Feedback | En



Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

Start Judging

Browse Scenarios

View Instructions

Imprint | Privacy Policy

Moral Dilemmas

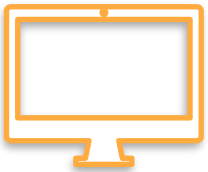
Cultural Perspectives

- Built to last centuries
- Conservated and turned into museums to preserve historical heritage

- Built to survive earthquakes and tornados
- Torn down and rebuilt every 30 years to keep historical heritage alive

Fair Distribution

Back to running examples



- Who should get the job?

  - Equality?
  - Equity by merit?
  - Equity by need?

- Who should get medical priority?

  - Equality?
  - Equity by merit?
  - Equity by need?

- Who should be recognised?
  Eg. Smart Phone screen lock

  - Equality?
  - Equity by merit?
  - Equity by need?

Fair Distribution

Other examples?

- No one gets the fastest road, but everyone arrives faster than without the traffic re-distribution

Approaches to Fairness: *Summary*

- Equality - everyone is treated the same or gets equal share of the resources to be distributed.
- Equity - everyone gets a share proportional to ones merits or needs.
- There is no "one fits all" fairness approach.
- What is "fair" highly depends on the specific domain and situation.
- Fairness choices can also vary across cultures

Legislation and Organisations

Ruta Binkyte – *November 2022*

Relevant Legal Frameworks

Anti-Discrimination Law

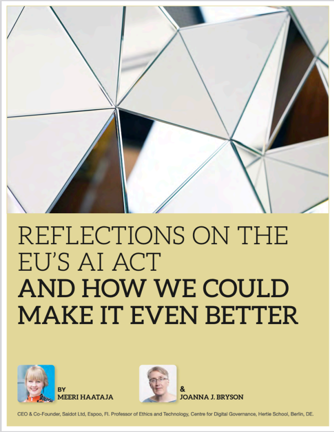
- EU Charter on Fundamental Rights: against discrimination on grounds of race and ethnic origin, religion or belief, disability, age or sexual orientation, gender
- Disparate Impact Legal Framework (US)
- Fair Housing Act (US)

GDPR (EU)

- A right to an explanation of algorithmic decision in the cases of high impact
- A high impact decision should not be fully automated
- The data subject agency on how the personal information is collected and used

Legislation and Organisations:

The European AI Act



Optional Reading

https://static1.squarespace.com/static/5e13e4b93175437bccfc4545/t/623254b3e9ae96717d593c10/1647465652248/reflections-on-the-EUs-AI-act-and-how-we-could-make-it-even-better-meeri-haataja-joanna-j-bryson.pdf

Fair AI Actors

- High Level Expert Group on AI https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
- OECD AI Policy Observatory https://oecd.ai/en/
- AlgorithmWatch https://algorithmwatch.org/en/
- Algorithmic Justice League https://www.ajl.org/

And many more!

Practical Exercises

Ruta Binkyte – *November 2022*

aivancity
PARIS-CACHAN

The Questions to Ask

The exercise

- Who will benefit from the solution? Do you think that AI solution can help to achieve company's goals?
- How would you describe current demographics in the company?
- What groups in the company are disadvantaged/ underrepresented? Why? Should you use their identifiers in the model?
- Where does the data come from? Is it representative? Can there be historical biases present?
- What do you want an algorithm to predict?What features can be (not) suitable for the prediction? Why?
- What features could correlate with a disadvantaged group?
- What fairness approach would be suitable for this scenario? Why?
- Is explainability crucial for this case? How would you explain an outcome to the employees?
- Do you think a human manager should take a final decision (with AI only as a recommender)?
- What other legal or ethical issues can be anticipated? Do you have suggestions on data collection and quantity, additional features or ML building team education/ diversity?

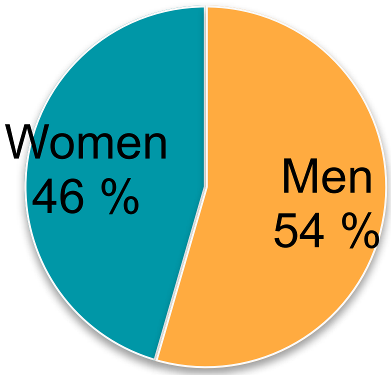
Scenario: Promotion Tool
The Context and the Goal

- The French tech company wants an algorithm to predict what employees should be considered for promotion. The HR department is overloaded, besides the management expects AI decisions to be more neutral and fair. The company hopes to increase diversity and equality and to recognise potential talents, that might have been missed otherwise.
- The values that the company wants to promote are expertise, leadership and social skills.
- If the model is successful in France the company wants to deploy it in its branches in West Africa

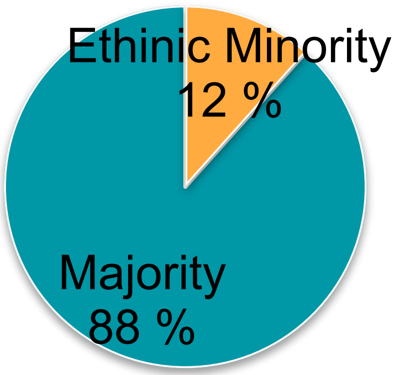
Ruta Binkyte – *November 2022*

Scenario: Promotion Tool

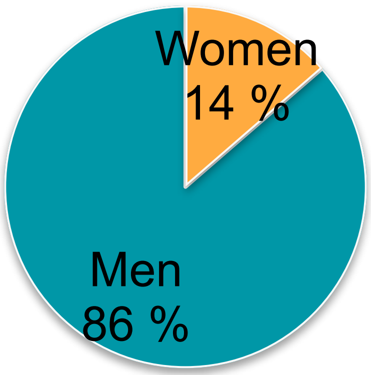
Current Statistics



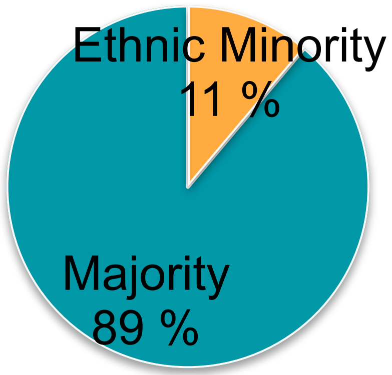
All Employees



All Employees



Managers



Managers

aivancity PARIS-CACHAN

Scenario: Promotion Tool

The Data Company Collects

Previously used criteria

- Previous promotions
- Hours spent at work (including extra hours)
- Plays Football (team sports considered to be informative of leadership and social skills)
- Participation at team-building and social events
- Opinion of co-workers
- Taking professional courses

Additional criteria considered for the model

- Presenting at the conferences
- Having lunch in company's cafe with colleagues (based on face recognition from the security camera operating in the cafe).
- Proposing innovative ideas at the meetings
- Unstructured content from the Instagram account
- Number of high social status friends on Social Media

How to build fair models?

Some rules of thumb

More tools for handling bias in the next lecture!

- Clearly formulate what you want to predict
- Think if data/features are suitable for this goal
- Think about vulnerable groups and individuals that can be affected.
- Can the attributes used, the way they are measured or the way data is collected include historical or underrepresentation bias? Always keep metadata for the future reference.
- In general it is better to keep the group attribute in the data for measuring and mitigating bias. Removing it from the data only opens door to proxy discrimination.

In general we assume person's ability or merit to be independent of the demographic features such as gender or race. That is why good and fair prediction has to be independent of those attributes, even if there is accidental or historical correlations in the data.

However, in medical scenarios, for example breast cancer prediction, not using gender would have an adverse result both on fairness and accuracy.

aivancity

PARIS-CACHAN

advancing education
in artificial intelligence