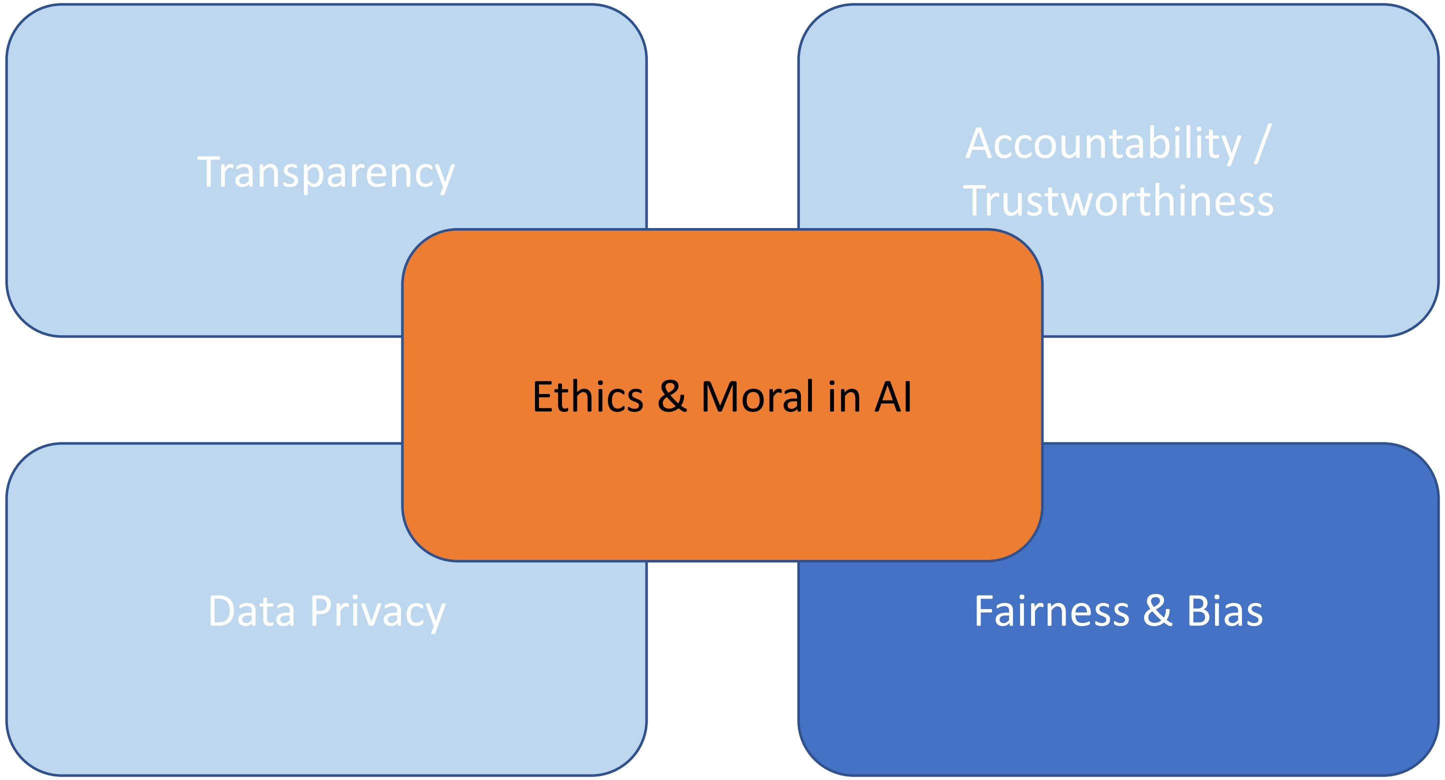
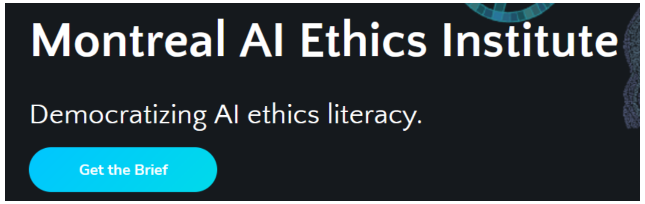
Software Engineering for AI-Enabled Systems

Prof. Dr.-Ing. Norbert Siegmund
Software Systems


Transparency
Accountability / Trustworthiness
Ethics & Moral in AI
Data Privacy
Fairness & Bias

Resources

https://montrealethics.ai/



Timnit Gebru https://twitter.com/timnitgebru

Student-run AI ethics journal: https://ojs.stanford.edu/ojs/index.php/grace/announcement

https://twitter.com/WomeninAIEthics

Ethics course at https://ethics.fast.ai/

Rachel Thomas (https://rachel.fast.ai/)

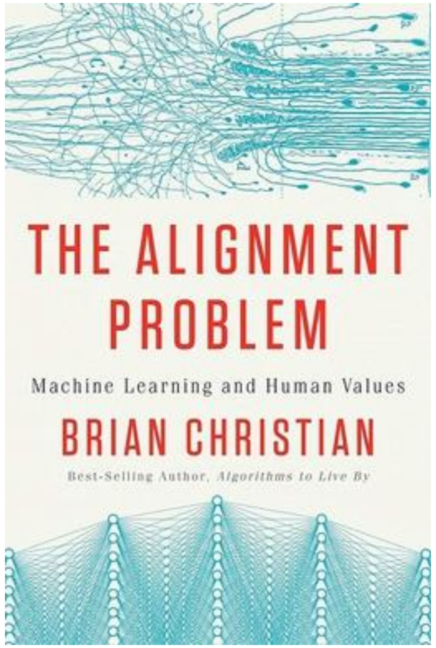
Photo by Gabriela Hasbun

Fairmlbook.org

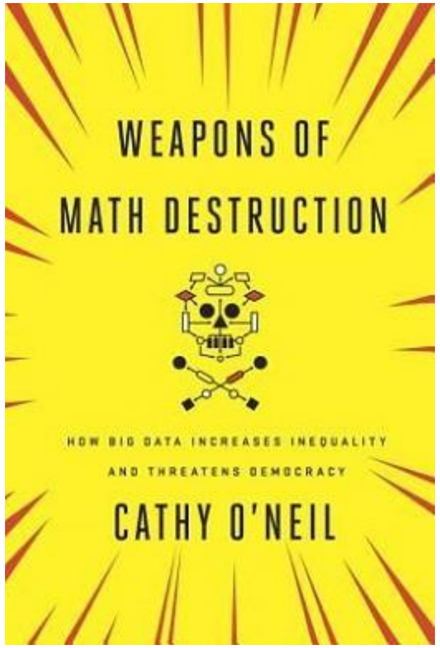
Dealing with Bias&Fairness in AI/ML/DS Systems: Tutorial
https://www.youtube.com/watch?v=N67pE1AF5cM

https://afraenkel.github.io/fairness-book/

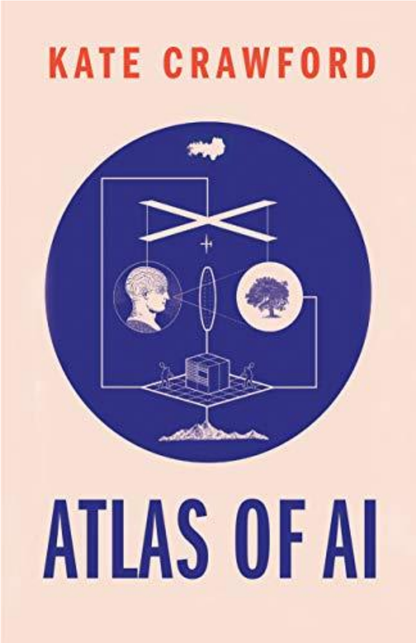
Books



“A jaw-dropping exploration of everything that goes wrong when we build AI systems and the movement to fix them.”



O'Neil, a mathematician, analyses how the use of big data and algorithms in a variety of fields, including insurance, advertising, education, and policing, can lead to decisions that harm the poor, reinforce racism, and amplify inequality.



“The hidden costs of artificial intelligence, from natural resources and labor to privacy, equality, and freedom.”

Topic I:
Ethics & Bias in AI

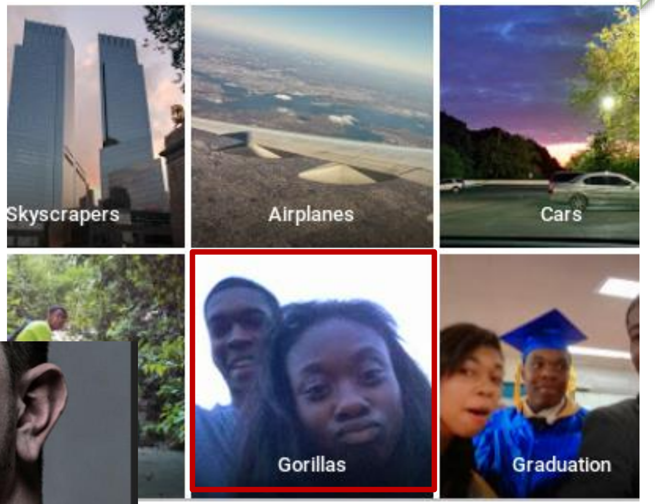
TL;DR:

- Defining ethics, fairness, bias
- Detecting and countering bias in the whole AI system life cycle
- Ethical dilemmas

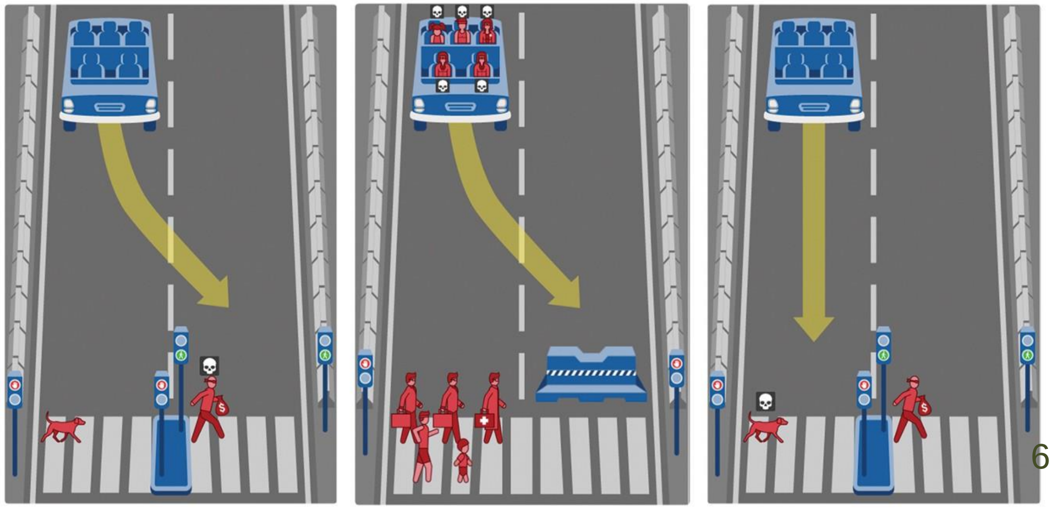
AI and Ethics (Bias&Fairness): A Big Problem

Microsoft's chat bot:
In 24 hours

Google Photo's labeling system



How would even humans decide?



Bias in AI systems can have severe personal consequences

The Naive Way of Looking at AI

The Industry Struggles

Los Angeles Times

TECHNOLOGY

A worker objected to Google's Israel military contract. Google told her to move to Brazil

More than 500 Google workers are backing a colleague who has accused the tech giant of retaliation over her objections to a corporate contract with the Israeli military. (Associated Press)

BY SUHAUNA HUSSAIN | STAFF WRITER

MARCH 15, 2022 6 AM PT

More than 500 Google workers have rallied behind a colleague who alleges she is being pushed out of her job because of her activism within the company, the latest flare-up between the tech giant and employees who speak out against its business practices and workplace conditions.

The workers have signed a petition accusing Google leadership of "unjustly retaliating" against Ariel Koren, a product marketing manager at Google for Education, for voicing criticism of Project Nimbus, a 1.2-billion contract Google and Amazon Web Services entered into with the Israeli military and government.

LATEST TECHNOLOGY >

COMPANY TOWN

Apple, Netflix, TikTok strike back against Russian state media content

March 2, 2022

TECHNOLOGY

Musk's SpaceX satellite dishes arrive in Ukraine, drawing minister's thanks

Feb. 28, 2022

TECHNOLOGY

How protesters in Russia and Ukraine are avoiding internet censorship — and jail

Feb. 25, 2022

TECHNOLOGY

Putin targets lots of Americans with disinformation. One example? Anti-vaccine groups

Feb. 25, 2022

ENTERTAINMENT & ARTS

Gaming has led the metaverse, but NFTs pose new ethical challenges. The DICE Summit discussions

Feb. 24, 2022

Guillaume Chaslot @gchaslot · Feb 25

YouTube's AI recommendations promoted Putin's channel Russia Today more than 100,000,000,000 times. I worked on these recos.

The AI community can tell @SusanWojcicki to stop using our work to actively promote dictators

> **Nando de Freitas** 🏳️‍🌈 @NandoDF · Feb 24
>
> What role can the AI community play in a world where bullies attack peaceful democratic countries 🇺🇦 and threaten the world? I'm really curious to hear from everyone.

45 · 1,598 · 3,788

Rachel Thomas @math_rachel · Apr 23, 2021

As many as 900 post office employees in the UK were wrongly prosecuted, convicted, and in some cases jailed, after bad data from an unreliable computer system falsely suggested cash shortfalls.

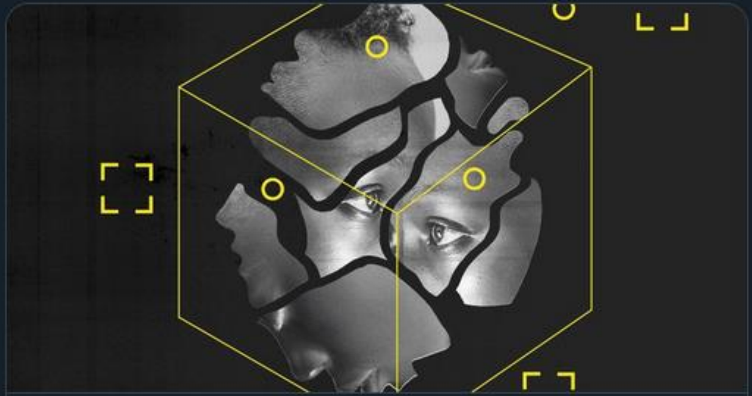
theguardian.com/uk-news/2021/a...

> Campaigners believe that as many as 900 operators, often known as subpostmasters, may have been prosecuted and convicted between 2000 and 2014 after the Horizon IT system installed by the Post Office and supplied by Fujitsu falsely suggested there were cash shortfalls.
>
> In his damning written judgment, Lord Justice Holroyde, sitting with Mr Justice Picken and Mrs Justice Farbey, said the Post Office, which brought the prosecutions itself, "knew that there were serious issues about the reliability of ALT rizon".

5 · 53 · 104

Colin Madland 🇨🇦❤️🇺🇦 @colinmadland · Sep 19, 2020

In case you're wondering, this goes far deeper than who gets to be seen in a zoom call or Twitter, and it's not new.



wired.com

The Best Algorithms Still Struggle to Recognize Black Faces

US government tests find even top-performing facial recognition systems misidentify black people at rates 5 to 10 times higher than th...

24 4,085 14.1K

Colin Madland 🇨🇦❤️🇺🇦 @colinmadland · Sep 19, 2020

Innocent people are in jail because of these same algorithms.

nytimes.com

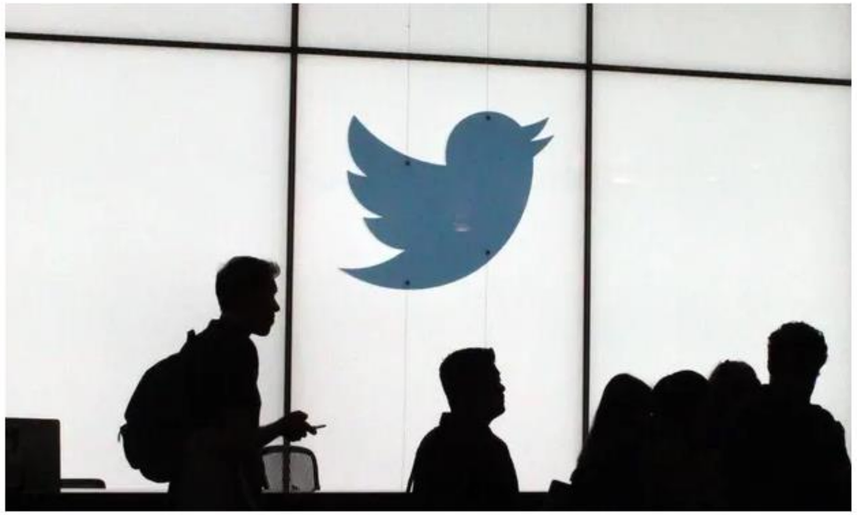
Wrongfully Accused by an Algorithm (Published 2020)

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did no...

22 3,613 14.5K

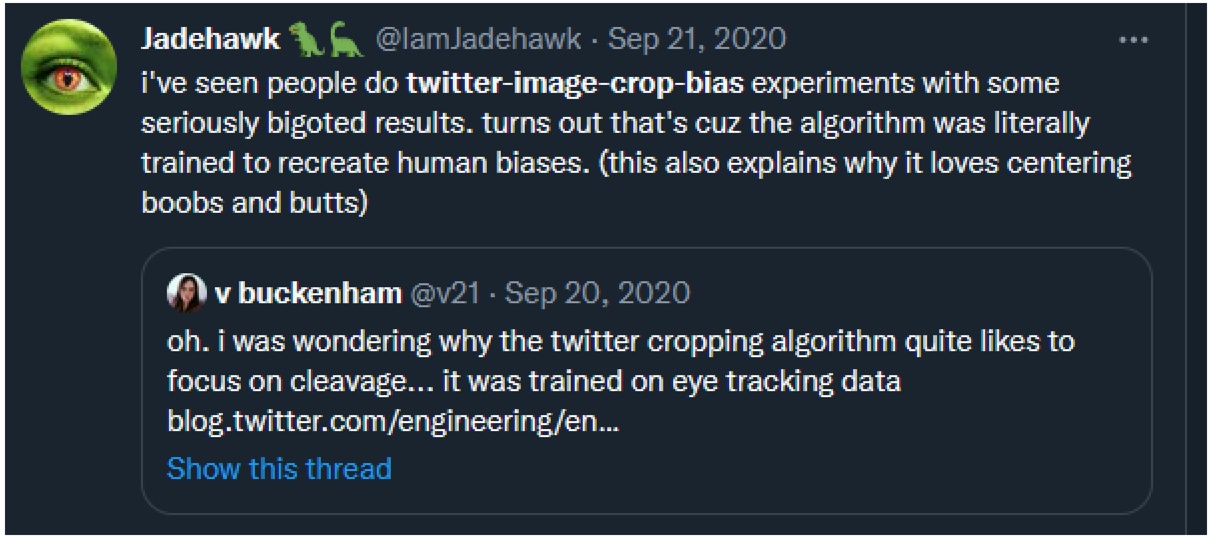
Twitter apologises for 'racist' image-cropping algorithm

Users highlight examples of feature automatically focusing on white faces over black ones



Twitter users began to spot flaws in the feature over the weekend. Photograph: Glenn Chapman/AFP/Getty Images

Twitter has apologised for a “racist” image cropping algorithm, after users discovered the feature was automatically focusing on white faces over black ones.



Jadehawk 🦕🦕 @IamJadehawk · Sep 21, 2020

i've seen people do **twitter-image-crop-bias** experiments with some seriously bigoted results. turns out that's cuz the algorithm was literally trained to recreate human biases. (this also explains why it loves centering boobs and butts)

> **v buckenham** @v21 · Sep 20, 2020
>
> oh. i was wondering why the twitter cropping algorithm quite likes to focus on cleavage... it was trained on eye tracking data blog.twitter.com/engineering/en...
>
> Show this thread

Topic II:
Ethics & Moral

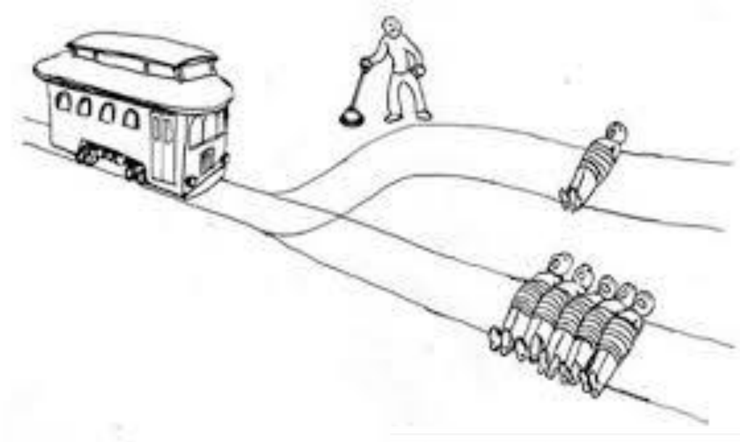
Ethics Overview

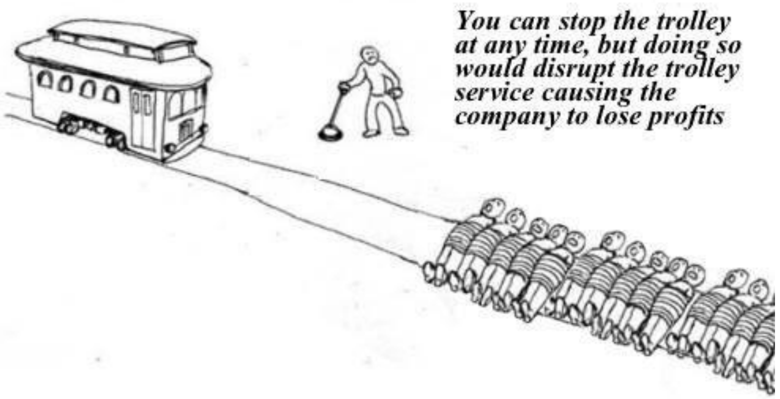
Ethics is concerned with what is good for individuals and society. From that we can infer moral principles that affect how people make decisions and lead their lives.

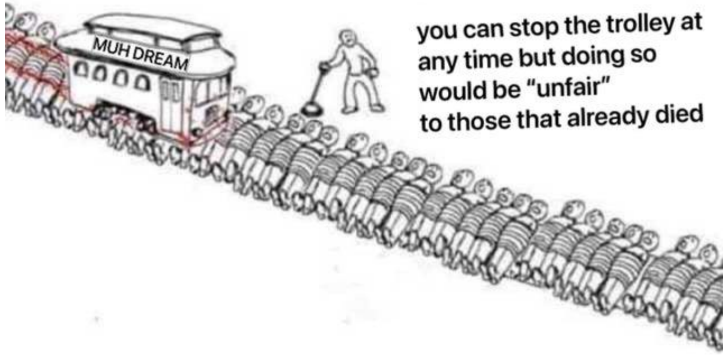
Meta-ethics: Nature of moral judgement; concerns with the origin and meaning of ethical principles
Normative ethics: Studies the criteria of what is right or wrong (why we do things that may appear counterintuitive)
Applied ethics: Investigates the application of ethical theories to controversial topics, such as war, rights of animals and plants, etc.
Normative ethical theories:

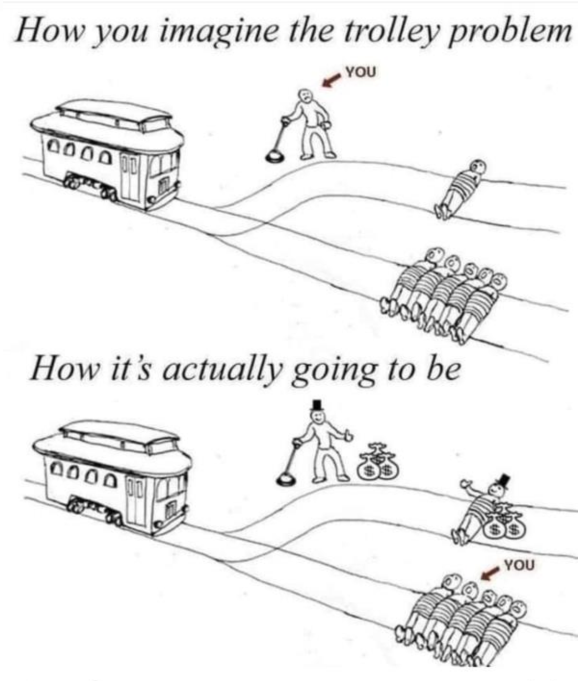
- Utilitarian ethics: Benefit the majority. Cons: Harming minority while benefiting majority; requires outcome prediction
- Deontological ethics: People should be treated with dignity and respect. Cons: Disagreements about principles leading to a decision; making a right choice can lead to bad consequences; possible conflicts in a duty
- Virtue ethics: Determine good virtue and making decisions based on them. Cons: Conflicts in virtues
- Rawls's theory of justice: Primary concern of justice is fairness (thought experiment: "veil of ignorance")
- Others: Ethics of care, Egoism, Religion or divine command theory, Natural Law, Social contract theory, Moral relativism

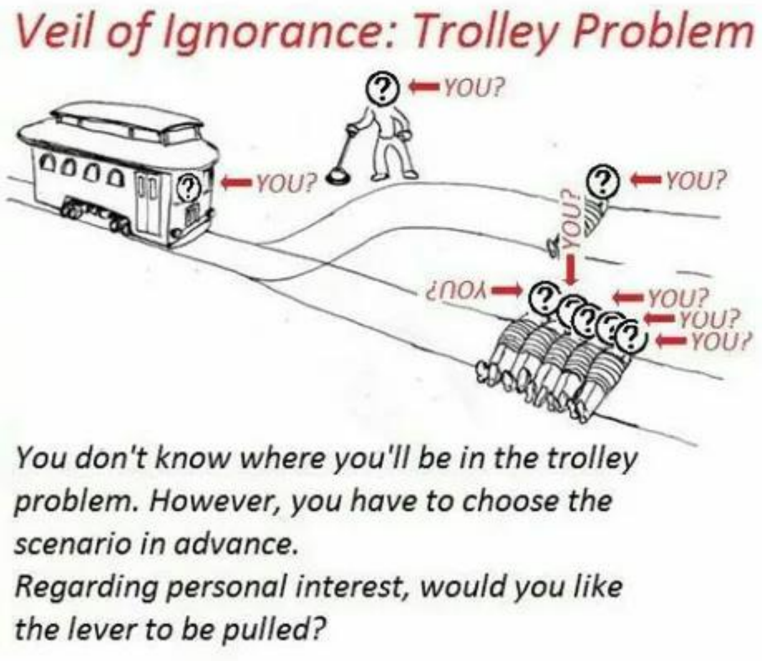
The Trolley Problem



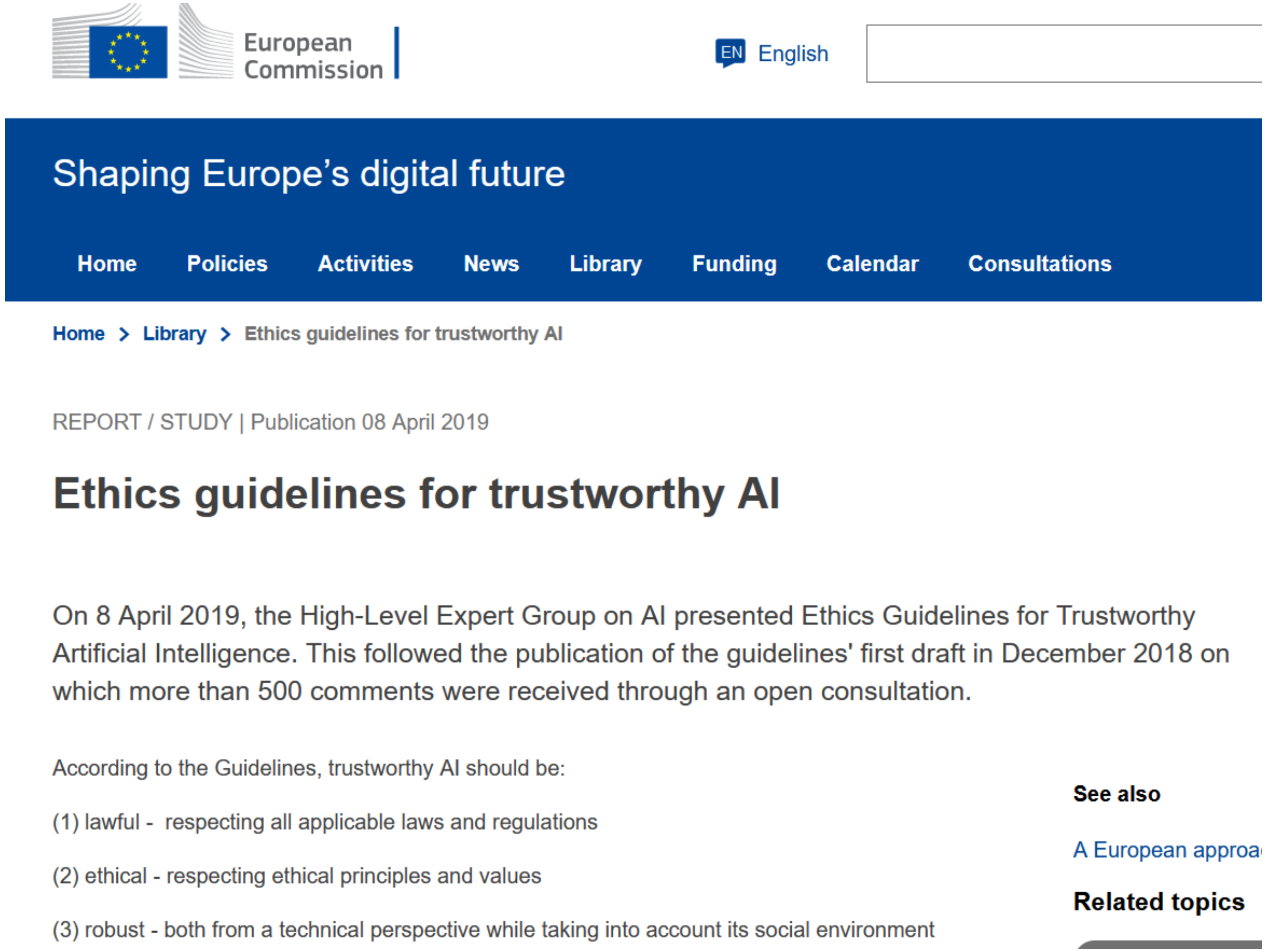








EU Ethics Guidelines


European Commission

EN English

Shaping Europe’s digital future

Home Policies Activities News Library Funding Calendar Consultations

Home > Library > Ethics guidelines for trustworthy AI

REPORT / STUDY | Publication 08 April 2019

Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

See also

A European approa

Related topics

7 key requirements for AI:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

What is the status?

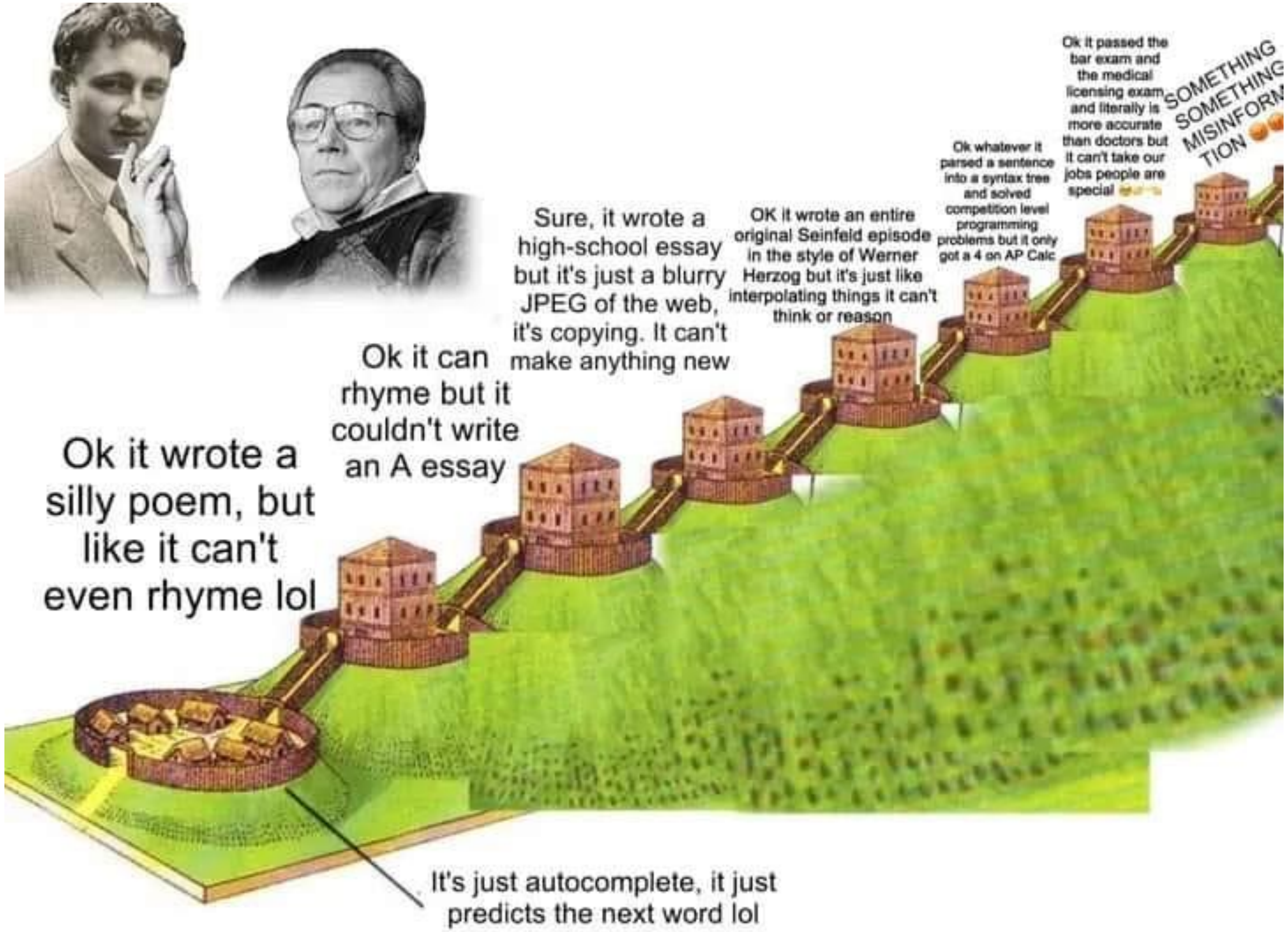
Forschung & Lehre: Herr Professor Metzinger, Sie haben sich in einer EU-Expertengruppe um den fairen Einsatz von Algorithmen bemüht. Wie waren Ihre Erfahrungen?

Thomas Metzinger: Meine Aufgabe war es, in dieser Expertengruppe 800 europäische Universitäten und 37 nationale Rektorenkonferenzen aus 48 europäischen Ländern zu vertreten. Es war enttäuschend, da vor allem die Repräsentanten der Wirtschaft ernsthaftere ethische Ansätze im Keim erstickt haben. In den verabschiedeten Papieren der Gruppe bin ich schlussendlich mit vielen Vorschlägen nicht durchgekommen. Beispielsweise habe ich Professuren für angewandte Ethik gefordert. Jede europäische Universität sollte eine Professur für angewandte Ethik in der Künstlichen Intelligenz bekommen, die Veranstaltungen für Studenten aller Fächer anbietet. Diese interdisziplinären Professuren sollten Forschungsergebnisse zusammenführen, öffentliche Debatten anstoßen, als Fenster von der akademischen in die öffentliche Welt. Dieser Vorschlag wurde in dem Gesetzentwurf von der EU komplett ignoriert.

F&L: Welche Chancen hätte ein derartiges Modell?

Thomas Metzinger: Die großen europäischen Konzerne, die in Zukunftsmärkte hineinwollen, meinen es natürlich nicht ernst mit der Ethik. Für die ist die flankierende Einführung "ethischer Standards" eigentlich nur eine Marketingstrategie, eine Dekoration. Auf Nachfragen, was denn ethisches Verhalten in der Wirtschaft wirklich bedeute, kommen Ausflüchte. Klar ist, dass diese Unternehmen nur freiwillige Selbstverpflichtungen wollen und Pseudo-Debatten inszenieren, um Zeit zu kaufen. Die gesetzliche Regelung scheuen sie, das ist ganz rational, wie der Teufel das Weihwasser. Es kommen dann auch offene Drohungen: Wenn Sie hier anfangen zu regulieren, dann gehen wir eben als Konzerne aus Europa weg. Früher war ich der Meinung, dass wer "schlau" ist, irgendwann in der Forschung landet. Es gibt jedoch viele extrem intelligente und durchaus umgängliche Menschen, die niemals ein politisches Amt oder eine Professur annehmen würden, weil sie nur ihren persönlichen Einfluss erhöhen oder viel Geld verdienen wollen.

Large Language Models: Moving the Goalpost



Topic III:
Bias

What is Bias?

Wikipedia: Bias is a disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair. Biases can be innate or learned. People may develop biases for or against an individual, a group, or a belief.

Stereotyping, prejudice or favoritism towards some things, people, or groups over others.

- Automation bias
- Confirmation bias
- Experimenter’s bias
- Group attribution bias
- Implicit bias
- In-group bias
- Out-group homogeneity bias

Systematic error introduced by a sampling or reporting procedure

- Coverage bias
- Non-response bias
- Participation bias
- Reporting bias
- Sampling bias
- Selection bias

Human Bias in Data Collection

Reporting bias: Sample has other properties, frequencies, and outcomes than whole population; people report only good/bad/interesting/relevant things, so it does not reflect the true frequency in the world

Selection bias: Sample selection is biased towards a certain way

- Coverage bias: Population in sample set does not match population in production
- Sampling bias: No random collection of data from target group (quality of data differs among groups)
- Non-response bias: People from certain groups may opt-out in surveys or feedback mechanisms

Overgeneralization: Data from one group is considered to generalize to others

Unconscious bias from „the world“:

- Labels may be skewed (e.g., by stereotypes)
- Even using mechanisms such as Mechanical Turk may produce such bias

Human Bias in ML Engineering

Automation bias: Favor results / decisions from automation / machines over other sources (despite error rates)

Group attribution bias: Falsely generalize in properties of individuals to the whole group the individuals belong

In-Group bias: ML engineers favors the group they belong to

Out-Group homogeneity bias: ML engineers stereotype individuals of groups they do not belong to or view their characteristics more uniform

Implicit Humas Bias

Assumptions are made based on our own mental model

Confirmation bias: ML engineers unconsciously process data in a way to affirm their own beliefs and hypotheses (in extreme cases, you train and build models until they reach their expectations -> **experimenter's bias**)

Unconscious bias in the procedures:

- Missing feature values may impact more minority groups than the majority
- Example: "Leave of absence" may indicate bad performance, but unfair biases against employees on parental leave

Bias in the World

Real-world data comes from humans who are not free of bias.

- Racisms
- Sexisms
- Stereotypes
- Group-based judgement
- Unfair conditions (working, treatment, interactions)
- Beliefs, misconceptions, etc.

Using a real-world data set means including this bias into your pipeline and when used in a production system **enforcing** this bias even more to the real world if not controlled for.

Bias in the Pipeline

Storing and linking data: misspelling of long, (for some people) uncommon names lead to loss of links for those groups

Preprocessing: Default values; subsumed values (e.g., average) may divert the attributes of a minority individual to an average individual (majority group)

Data exploration: outlier may be cropped away; statistics often only relevant for larger groups; random data picks will hit only common individuals

Identify Bias

Proactively audit for potential sources of bias with a diverse team (representation problem).
Possible red flags:

- Missing feature values: Key characteristics are under-represented -> reporting & selection bias
- Unexpected feature values: Point to possible problems in data collection or inaccuracies that may introduce bias
- Data skew: Any kind of skew that leads to under- or over-representation of groups

Topic IV: Fairness

Group Level Fairness

Individual Fairness

Society Fairness

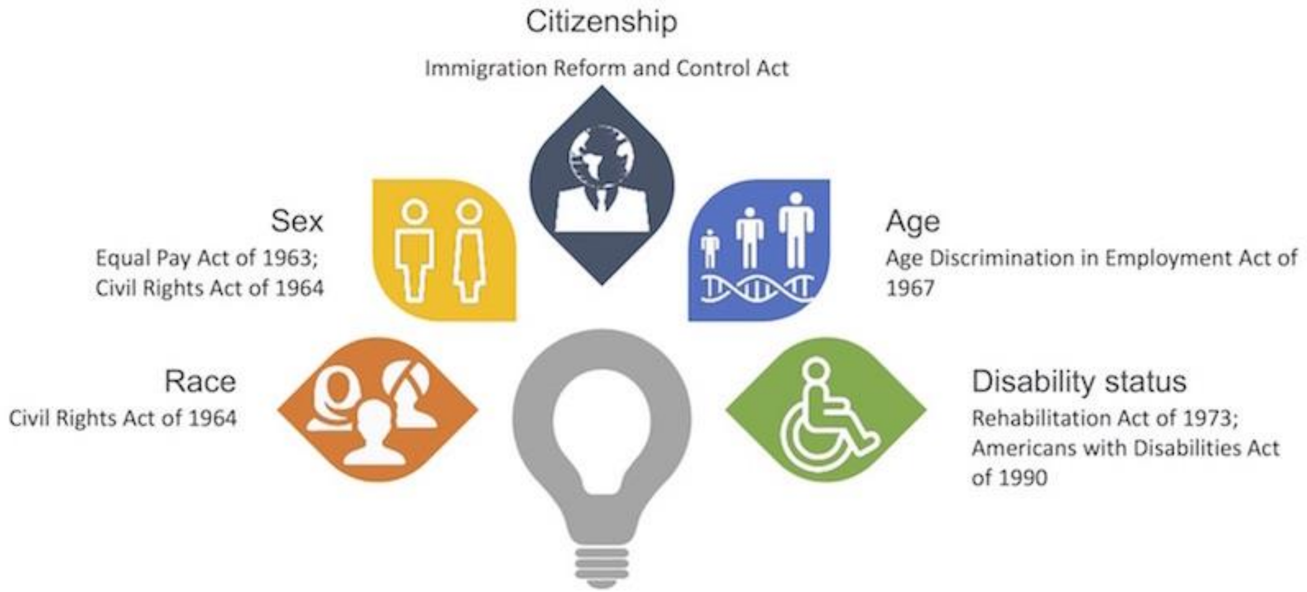
Do outcomes systematically differ between demographic groups?

Am I treated equally as others?

Is the gain of the society maximized?

Demographic Parity, Equalized Odds, Eq and Predictive Rate Parity

Discrimination & Sensitive / Protected Attributes



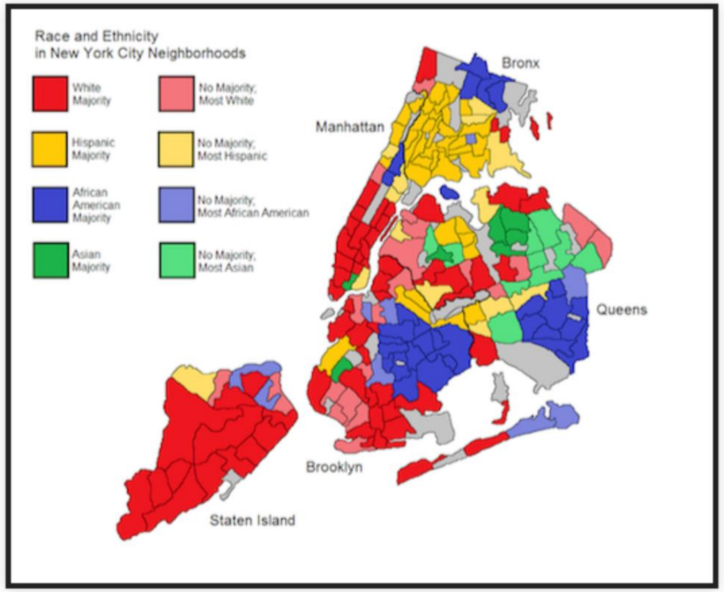
- Population includes various minority groups
    - Ethnic, religious, medical, geographic
    - Marital status
    - Socioeconomical status
- Protected by laws & policies
- How do we monitor & regulate decisions made by ML?

Fairness through Unawareness

Idea: If we are unaware of protected attributes while making decisions, our decisions will be fair. So, remove protected attributes such that the ML model cannot learn discriminating behavior.

Problems:

- **Proxy variables:** Features may correlate with class membership. Neighborhood as a proxy for race. When erasing “race” feature from the data set, the proxy may remain.



Fairness Definitions

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Group fairness / statistical parity / equal acceptance rate / benchmarking

Definitions based on *predicted* and *actual* outcomes. Requires to have a ground truth label to compare predictions with.

Definitions based on predicted probabilities and actual outcome.

Similarity based measures.

Causal reasoning.

Group Measures: Demographic / Statistical Parity / Equal Acceptance

Idea: The probability of a positive outcome of $\hat{Y}$ is independent from the protected attribute A :

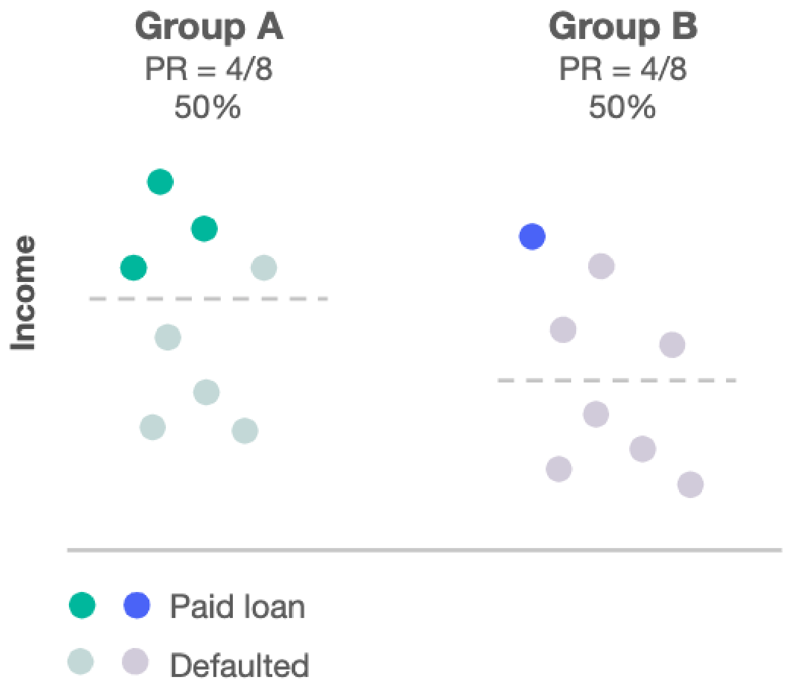
$$p(\hat{Y} = 1 | A = a) = p(\hat{Y} = 1 | A = b) \ \forall a, b, \in A$$

Example:

- Hiring decision should be independent of gender
- Treatment should be independent of age

When to use:

- Change the state of current world to improve it (e.g., minority groups should be better represented)
- Awareness of historical bias affecting the quality of our data



Fairness Measures: Equal Opportunity

False negative error rate balance / equal true positive rate

Idea: Positive outcome should be equal for different groups (i.e., every group should have the same chance to get an opportunity / treatment / etc.) based on the true positive rate (TPR).

$$p(\hat{Y} = 1 | A = a, Y = 1) = p(\hat{Y} = 1 | A = b, Y = 1) \forall a, b, \in A$$

So, we measure whether people who should qualify for an opportunity are equally likely to do so no matter to which group they belong, but with respect to their own group's true positive rate.

Examples:

- Funding stipends
- Admission rates to university

When to use:

- Strong emphasize on accurate positive outcome prediction
- False positives are not costly or severe; label should be objective

Predictive (Rate) Parity / Accuracy Parity

Idea: The probability of a subject with positive predictive value should truly belong to the positive class.

$$\mathrm{p}(\mathrm{Y} = 1|\hat{Y} = 1, \mathrm{A} = \mathrm{a}) = \mathrm{p}(\mathrm{Y} = 1|\hat{Y} = 1, \mathrm{A} = \mathrm{b}\ )\ \forall a, b, \in A$$

So, the chances for an individual to be positively classified are the same no matter what group. In general, similar to equality of opportunity, but more difficult to measure.

Fairness Measure: Equalized Odds

Conditional procedure accuracy equality / disparate mistreatment

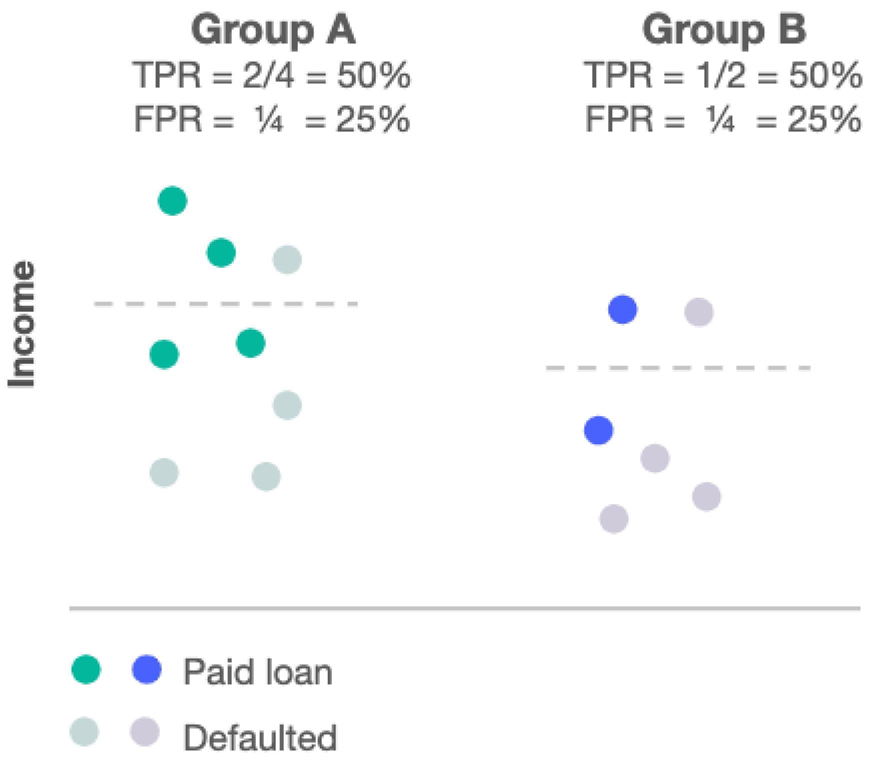
Idea: Not only the true positive outcome should be equal among groups, but also the false positive outcome (i.e., we should be wrong at the same probability when predicting a positive outcome).

$$p(\hat{Y} = 1 | A = a, Y = y) = p(\hat{Y} = 1 | A = b, Y = y), \forall a, b \in A \; and \; y \in \{0,1\}$$

So, **FPR** and **TPR** are the same. Requires to know the ground truth that needs to be collected in an unbiased way.
The consequence may be that we need to reduce the TPR in order to balance it with the FPR. This could result in a loss of profit or render the system non-sensible.

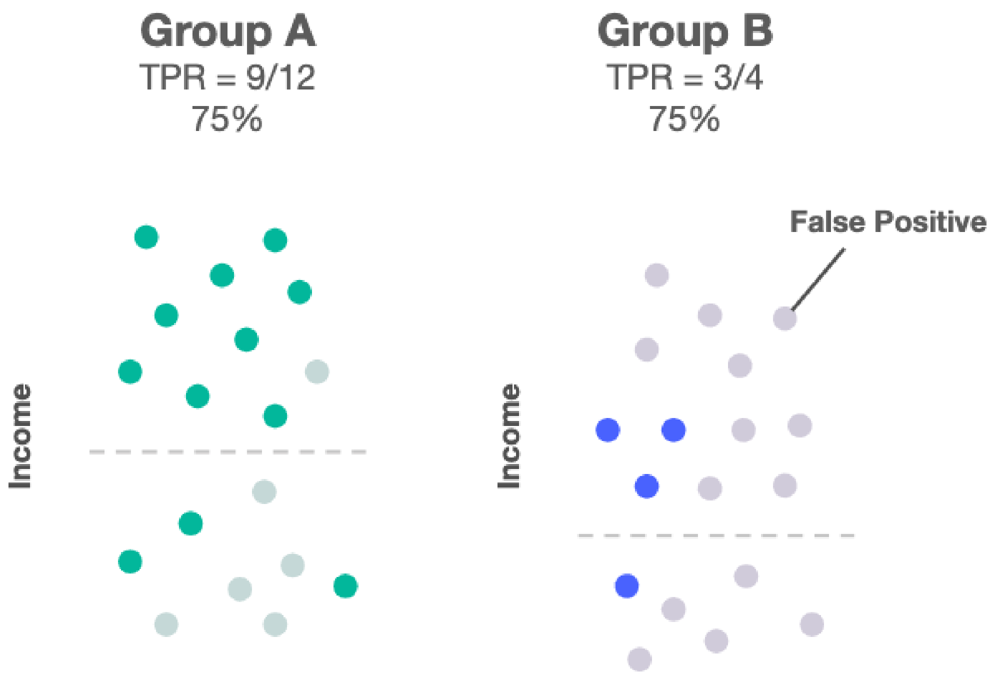
When to use:

- Aim for predicting positive outcome correctly and aim for minimizing costly false positives
- Project goal does not heavily depend on a high recall in FPR

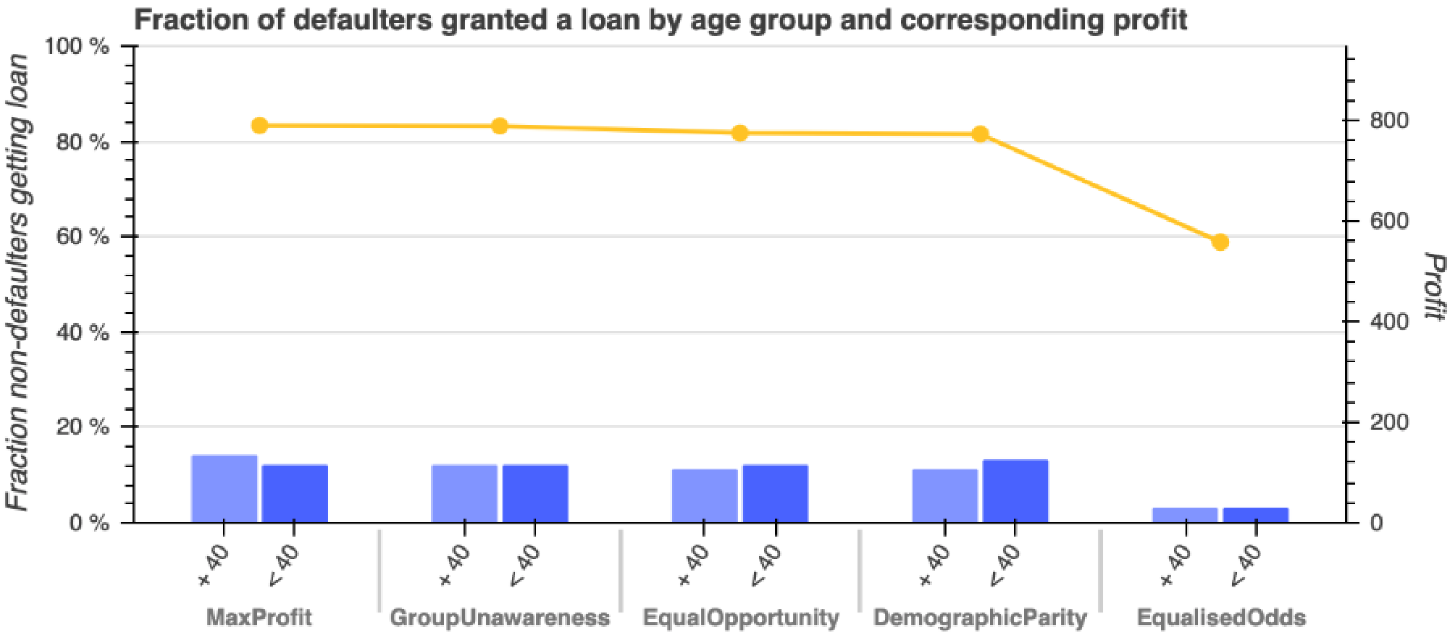


Problems of Group Fairness

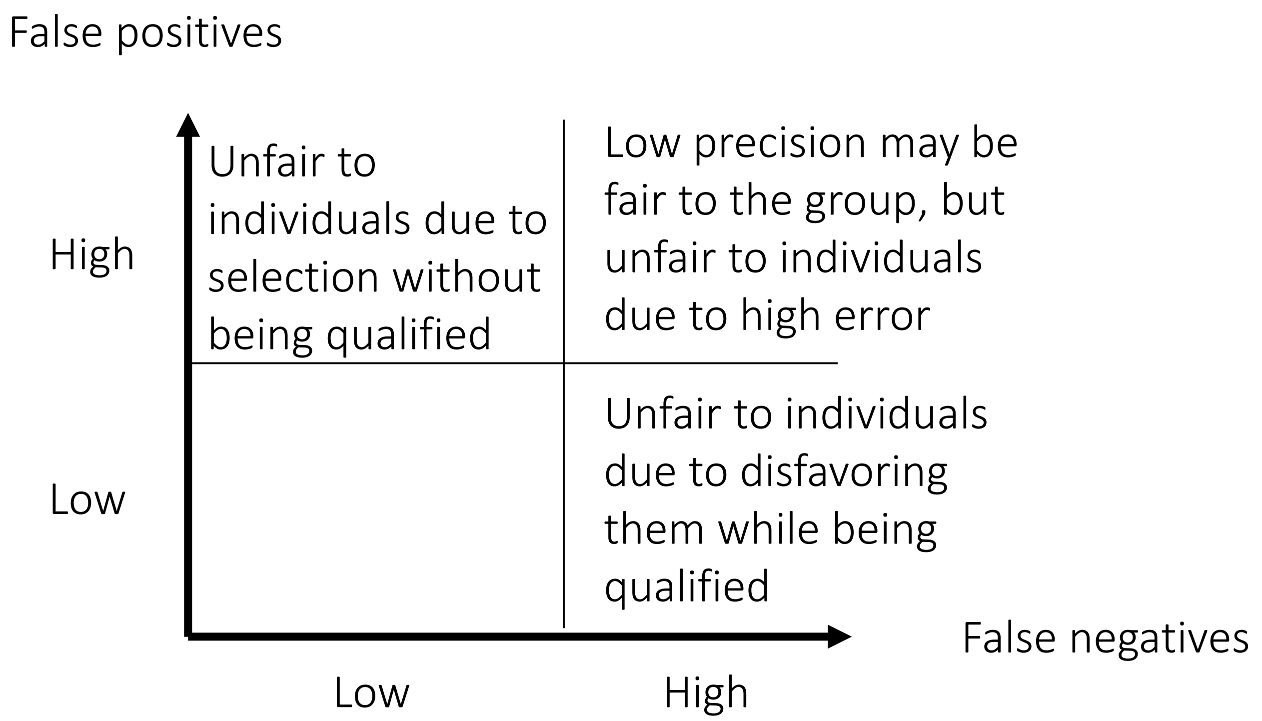
Equal Opportunity: Too many false positives



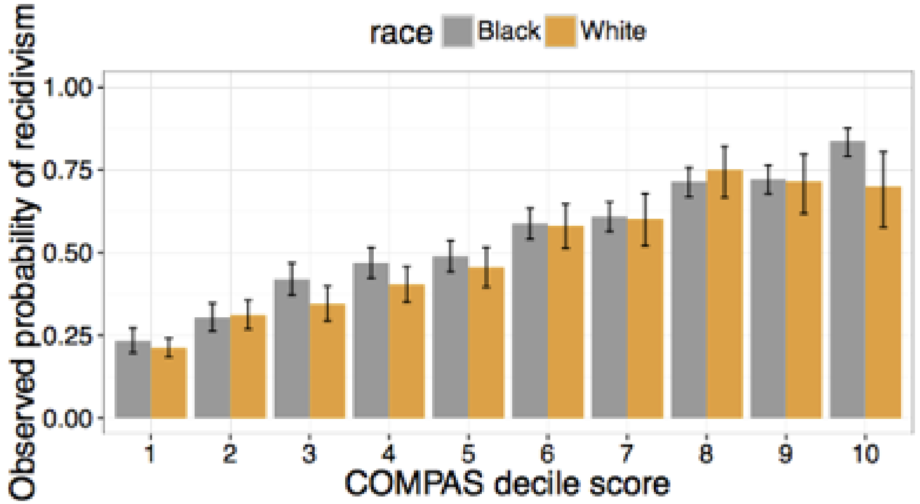
Equalized Odds: Too low profit



Unfairness of Group Level Fairness



AI algorithm predicts likelihood of recidivism (risk of committing crimes in the future), which was used by judges to decide the length and type of sentencing while considering the input-output relationship as a black box.



	Black Defendants Low	Black Defendants High		White Defendants Low	White Defendants High
Survived	990	805	Survived	1139	349
Recidivated	532	1369	Recidivated	461	505
FP rate: 44.85			FP rate: 23.45		
FN rate: 27.99			FN rate: 47.72		

COMPAS accuracy for white defendants: 67%; for black defendants 64%. Demographic parity / fairness satisfied. What is the problem?

The algorithm makes up for detaining releasable Black defendants by wrongly releasing white defendants.

Base rates differ, so no trade-off-free fairness is possible (see next).

Prediction Fails Differently for Black Defendants

	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%

Group Fairness: Impossibility theorem

Demographic parity


Machine Bias

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

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Equalized odds
(same probability for FPR & TPR)

$$\frac{FP}{FP+TN} \quad \frac{TP}{TP+FN}$$

Predictive rate parity
(same probability for PPV)

$$\frac{TP}{TP+FP}.$$

if an instrument satisfies **predictive parity** … but the prevalence differs between groups, the instrument cannot achieve **equal false positive [rates]** and **[equal] false negative rates** across those groups.

A Matter of Perspective

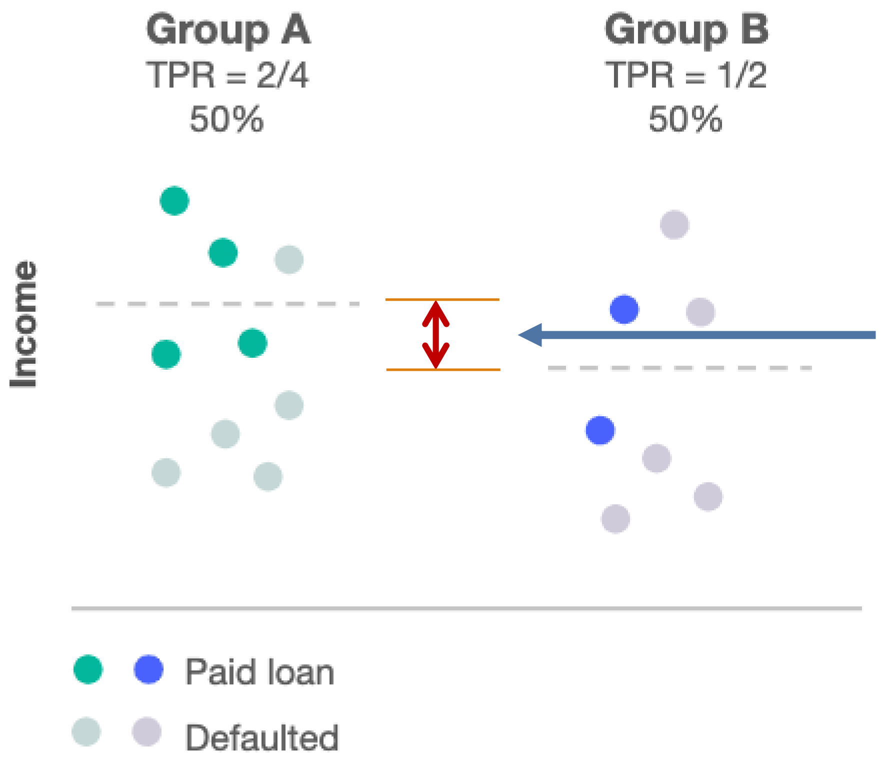
No “correct” fairness measure!

Those labeled high risk, how many recidivated?
Predictive rate parity: Because high risk individuals should be classified as high risk.
COMPAS achieves that!

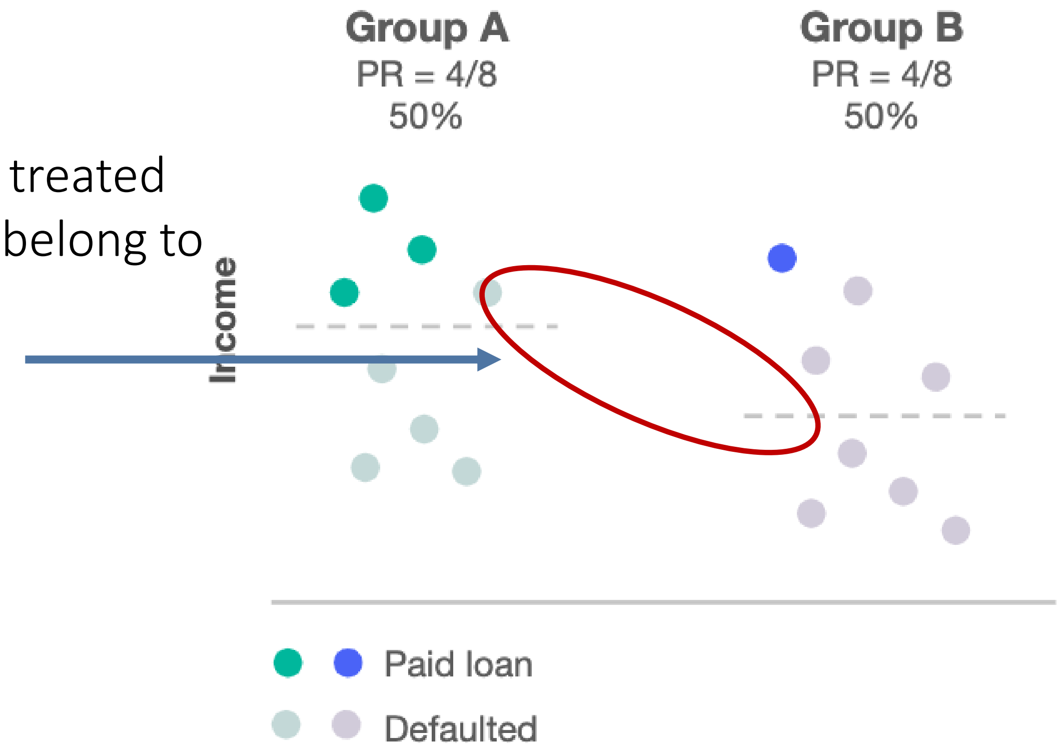
What is the probability I‘ll be incorrectly classified as high risk?
Equal opportunity: False positive rate should be fair.

Is it demographically fair?
Demographic parity: Protected attribute should not affect prediction.
COMPAS fails this for individual positive / negative error rates.

Fairness of Groups vs. Fairness of the Individual

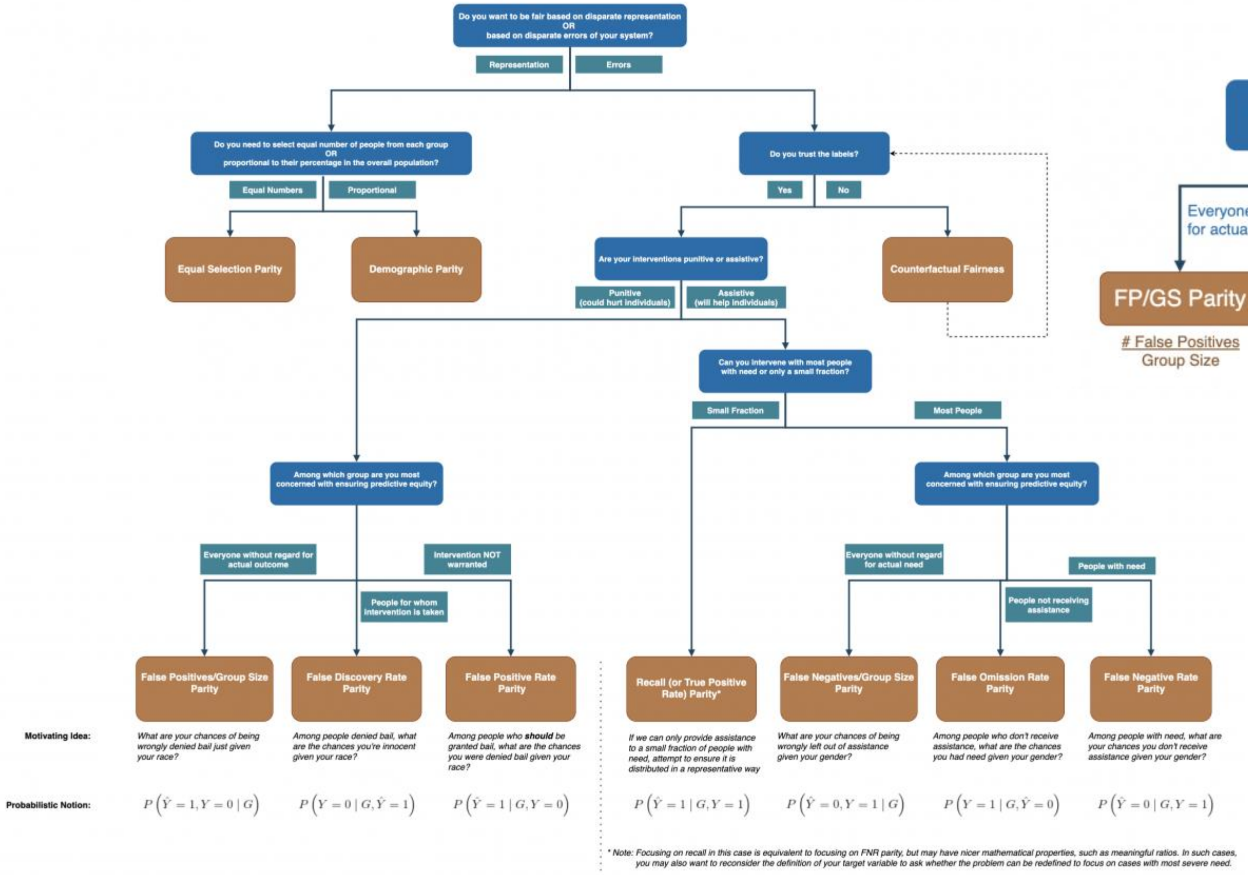


Why is an individual treated differently just they belong to a certain group?

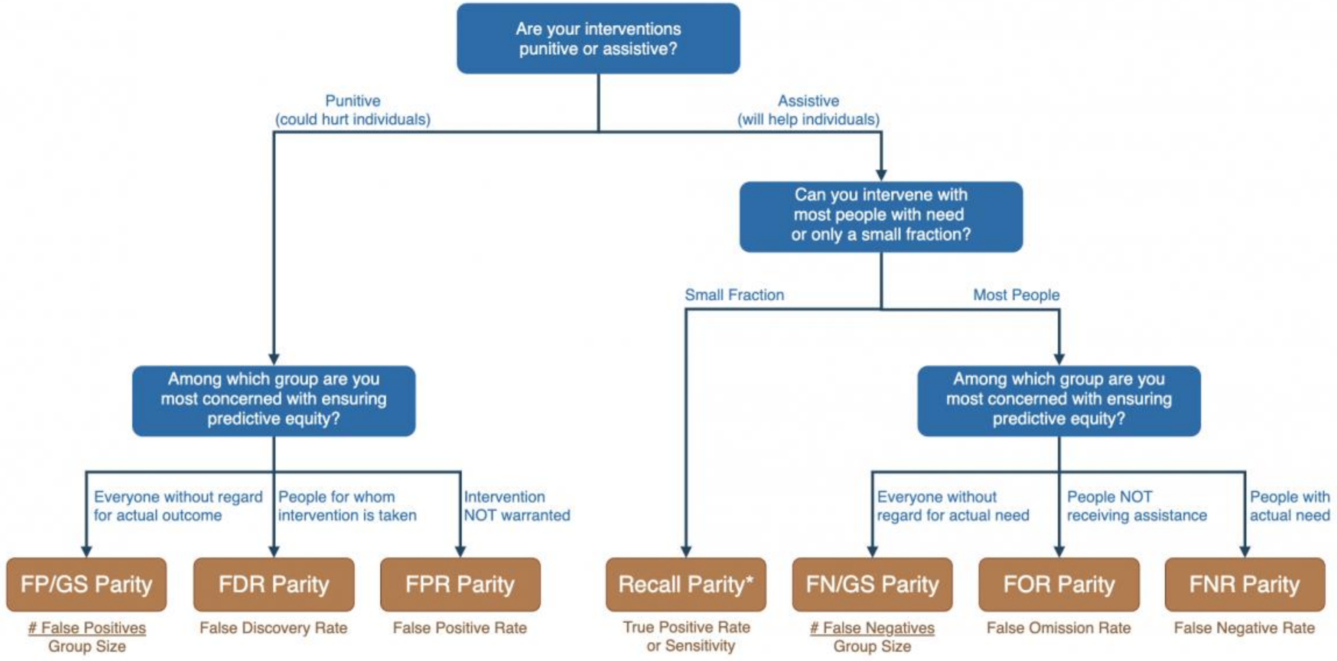


https://research.google.com/bigpicture/attacking-discrimination-in-ml/

FAIRNESS TREE



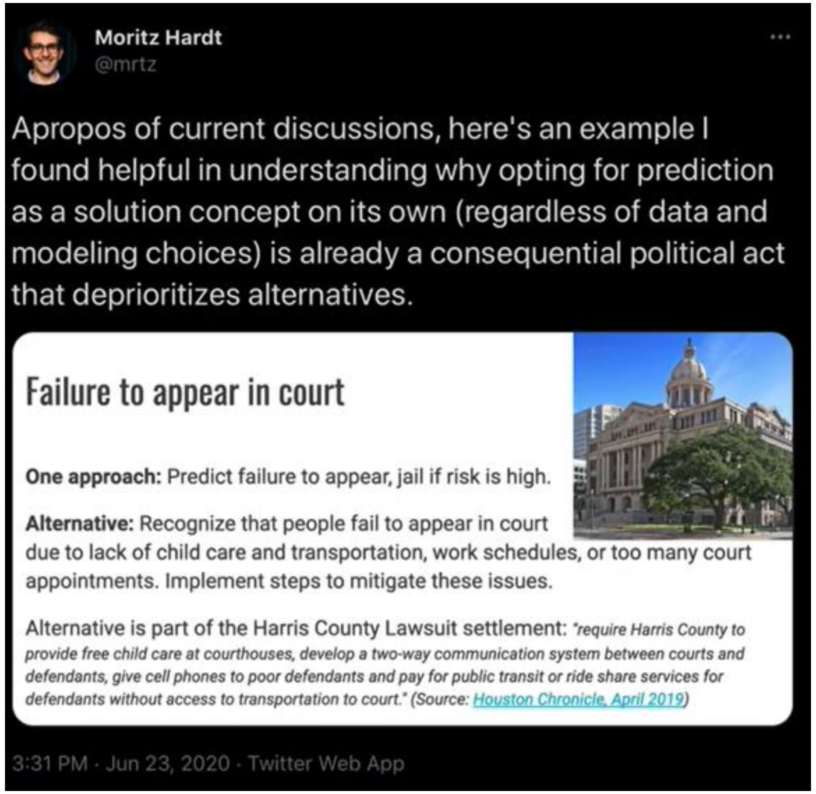
FAIRNESS TREE (Zoomed in)



Disagreement is not special to ML. Alternatives Approaches

Examples from real world: Custom controls, credit loans, insurance rates

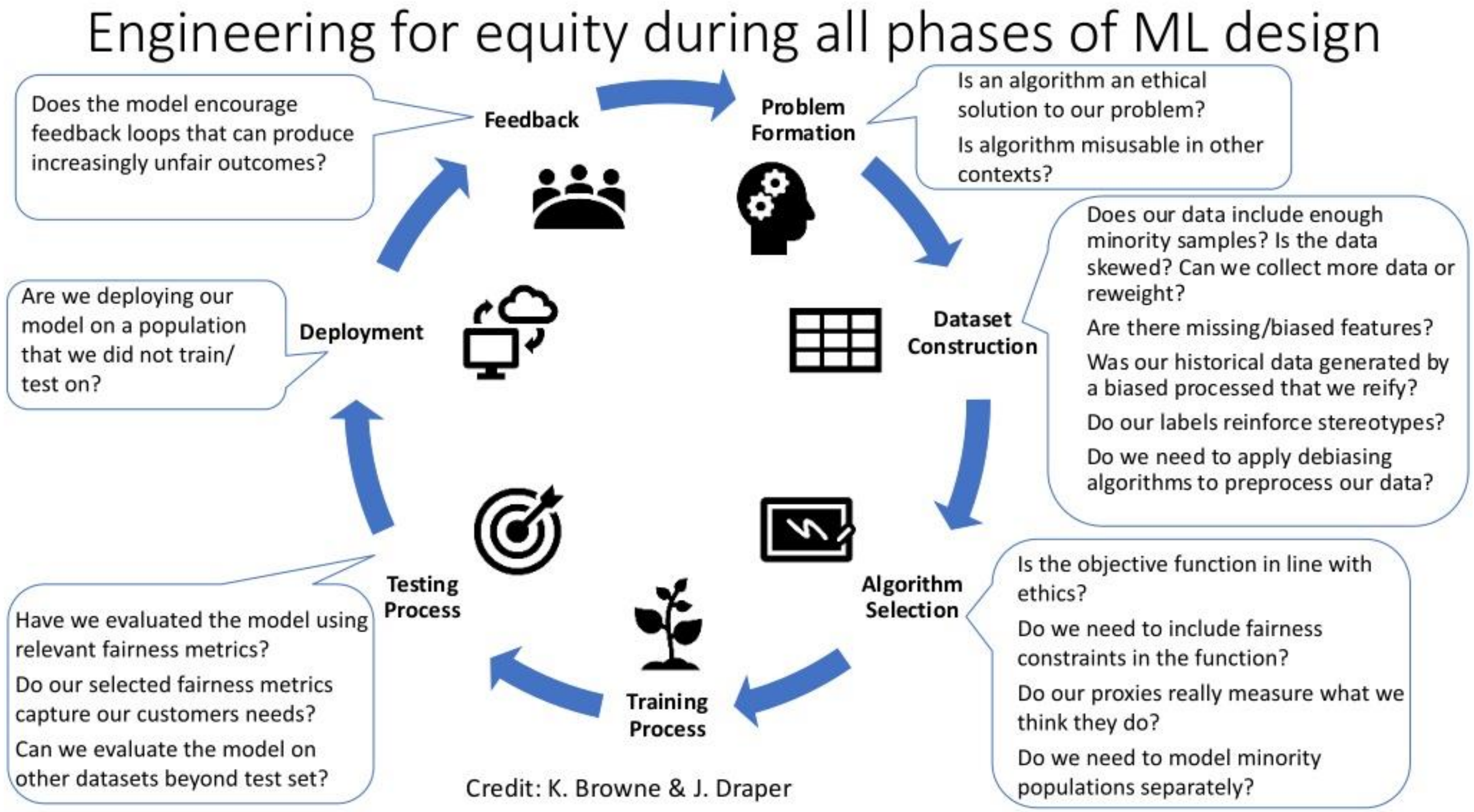
Opting for prediction might already limiting alternatives. In other words, deciding to predict a certain outcome at all may already cause bias and unfairness.



Equality

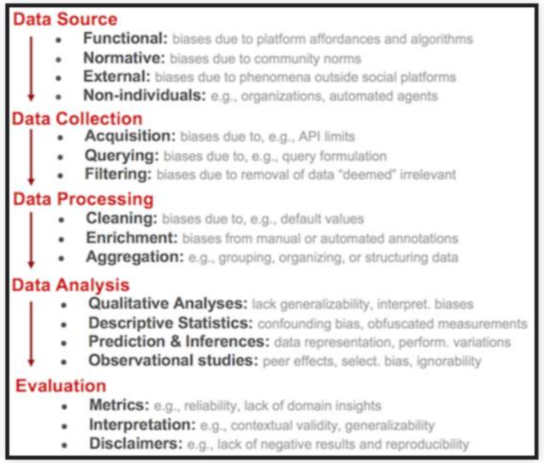
The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Fairness-aware Machine Learning



Bennett et al., WSDM Tutorial (2019)

Counter Bias



Population bias: Check demographics in the target population

Under-&over-representation: Ensure sufficient amount of data for all groups and avoid over-representation

Data augmentation: Synthesize data for minority groups

Fairness evaluation: Collect more data for groups with highest error rates

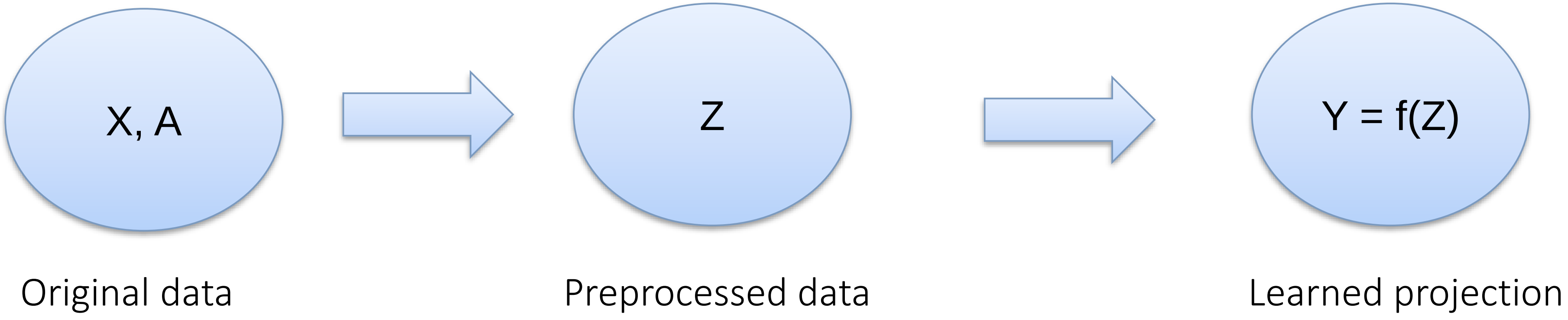
Bias at any stage of the ML pipeline: Be aware and do counter measures

Document data sets to log: Purpose, provenance, creation, composition, and distribution of data

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Preprocessing: Removing sensitive attribute

Idea: Preprocess data X in way that any information correlated with a sensitive attribute A is removed while maintaining as much information from the data as possible.



Pros: Can be used for any ML task; does not require to adapt the learning algorithm; testing does not require the access of sensitive attributes

Cons: Optimizes only statistical parity or individual fairness (Y label not available, which is needed for group level fairness)

Preprocessing: Relabelling, Reweighing, …

Idea: Find the causes of bias and try to solve them

Approaches:

- Relabelling: Assess representation bias of the labels (e.g., labels come only from one group) and relabel them again with a wider representation
- Reweighing: Increase the weight of minority groups with respect to protected attributes
- Data collection: Obtain further data samples to achieve parity of data samples on protected attributes

Bias Mitigation in Algorithms & Post-Processing Bias Mitigation

Adversarial debiasing: Model with two goals: (i) maximizing prediction accuracy, and (ii) reduce adversary's ability to determine / predict a protected / sensitive attribute from the prediction

Prejudice remover: Add a regularization term to the objective functions of the ML model to penalize discrimination of protected attributes

Fairness measures: Apply the different fairness metrics and test for bias

Bias Mitigation in Algorithms & Post-Processing Bias Mitigation

Adversarial debiasing: Model with two goals: (i) maximizing prediction accuracy, and (ii) reduce adversary's ability to determine / predict a protected / sensitive attribute from the prediction

Prejudice remover: Add a regularization term to the objective functions of the ML model to penalize discrimination of protected attributes

Fairness measures: Apply the different fairness metrics and test for bias

Open Questions

At the same time, debating the merits of these technologies on the basis of their likely accuracy for different groups may distract from a more fundamental question: should we ever deploy such systems, even if they perform equally well for everyone? We may want to regulate the police's access to such tools, even if the tools are perfectly accurate. Our civil rights—freedom or movement and association—are equally threatened by these technologies when they fail and when they work well.

https://fairmlbook.org/pdf/fairmlbook.pdf

Business Versus Ethics

The close link between business and science is not only revealed by the fact that all of the major AI conferences are sponsored by industry partners. The link between business and science is also well illustrated by the AI Index 2018 (Shoham et al. 2018). Statistics show that, for example, the number of corporate-affiliated AI papers has grown significantly in recent years. Furthermore, there is a huge growth in the number of active AI startups, each supported by huge amounts of annual funding from Venture Capital firms. Tens of thousands of AI-related patents are registered each year. Different industries are incorporating AI applications in a broad variety of fields, ranging from manufacturing, supply-chain management, and service development, to marketing and risk assessment. All in all, the global AI market comprises more than 7 billion dollars (Wiggers 2019).

Ethics in Practice

Do ethical guidelines bring about a change in individual decision-making regardless of the larger social context? In a recent controlled study, researchers critically reviewed the idea that ethical guidelines serve as a basis for ethical decision-making for software engineers (McNamara et al. 2018). In brief, their main finding was that the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behavior of professionals from the tech community. In the survey, 63 software engineering students and 105 professional

The Ethics of AI Ethics: An Evaluation of Guidelines

Thilo Hagendorff

Minds and Machines **30**, 99–120 (2020) | Cite this article