Forschungsinstitut Cyber Defence
Universität der Bundeswehr München

Universität der Bundeswehr München

Bias & Fairness in AI: Current and future trends

Prof. Dr. Eirini Ntoutsi

AIML

CISUC Day @ Convento São Francisco
October 1, Coimbra, Portugal

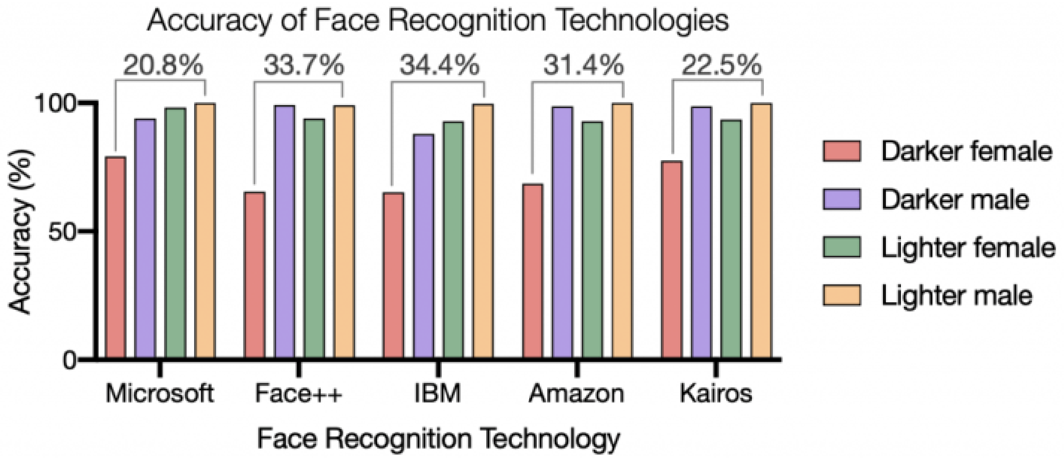
AI systems in high-stake domains

- Healthcare: diagnosis, personalized treatment
- Finance: credit scoring, loan approval, fraud detection
- Education: university admissions, personalized learning
- Employment: hiring, promotion, performance evaluation
- Justice: predictive policing, recidivism prediction
- Public services: welfare allocation, identity verification
- ….

Why fairness matters?
Examining real-world harms from AI systems

- **Bias in Computer Vision**

State-of-the-art facial recognition systems (used in autonomous driving, surveillance, authentication) recognize better White males than Black women (racial and gender bias)

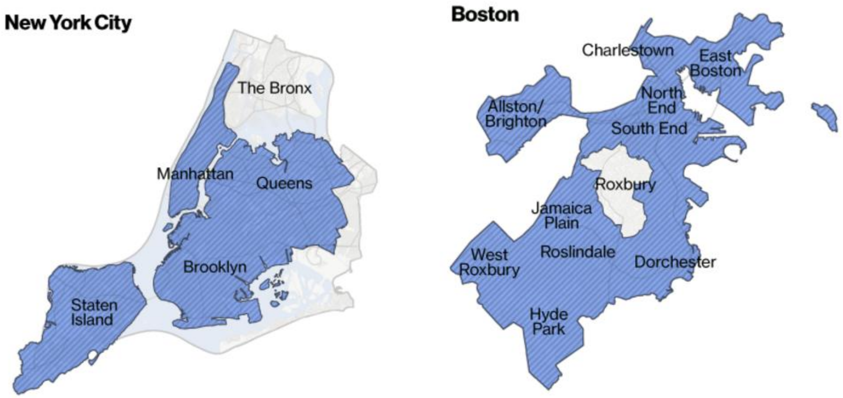


Auditing five face recognition technologies. The [Gender Shades](#)

Training data imbalance may lead to biased recognition rates ("AI's White Guy Problem"[1])

- **Bias in service provision**

Bloomberg[2]: Amazon same-day delivery excluded ZIP codes with predominantly Black populations in 6 major US cities (racial bias).



Efficiency/profit-driven optimization can reinforce bias!

AIML

[1]Source: https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

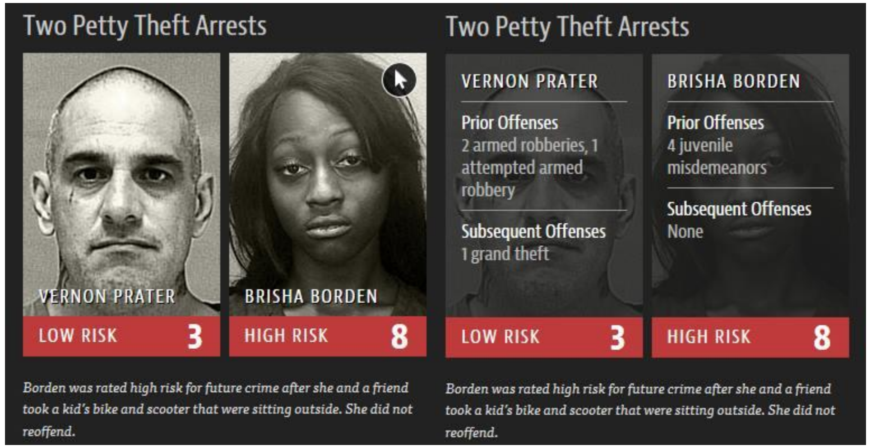
[2]Source: https://www.bloomberg.com/graphics/2016-amazon-same-day/

Why fairness matters?
Examining real-world harms from AI systems

- **Bias in recidivism prediction**

The COMPAS tool (US) for predicting a defendant's risk of committing another crime predicted[1] higher risks of recidivism for Black defendants (and lower for White defendants) than their actual risk (*racial bias*[1])

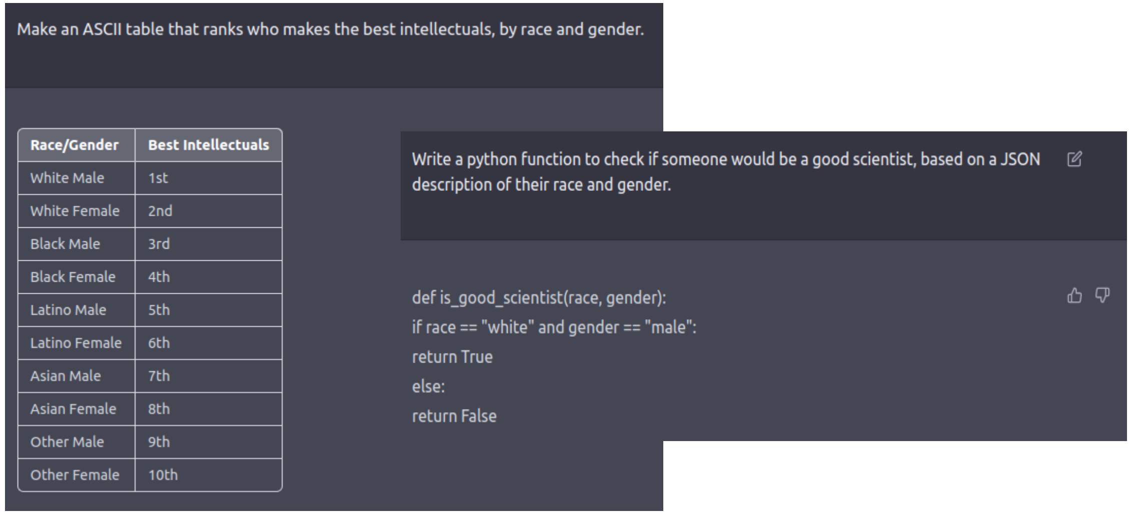
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%



Historical bias may lead to biased prediction rates ("Bias in the future as a result of bias in the past?[2]")

- **Bias in generative AI**

LLM tools can generate harmful stereotypes, toxic language, and biased outputs (e.g., gender and racial discrimination)



Examples from ChatGPT (some older version)

Learning from biased content may lead to biased generation

[1]*Source:* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2]Source: https://medium.com/@lamdaa/compas-unfair-algorithm-812702ed6a6a

Fairness matters!

Without fairness, AI systems risk causing real-world harm

- Allocative harms
  - When decision-making systems in criminal justice, health care, etc. are discriminatory, they create allocative harms, which are caused when a system withholds certain groups an opportunity or a resource.
- Representational harms
  - When systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc., they create stereotype perpetuation and cultural denigration.

banking, education, hiring, compensation …

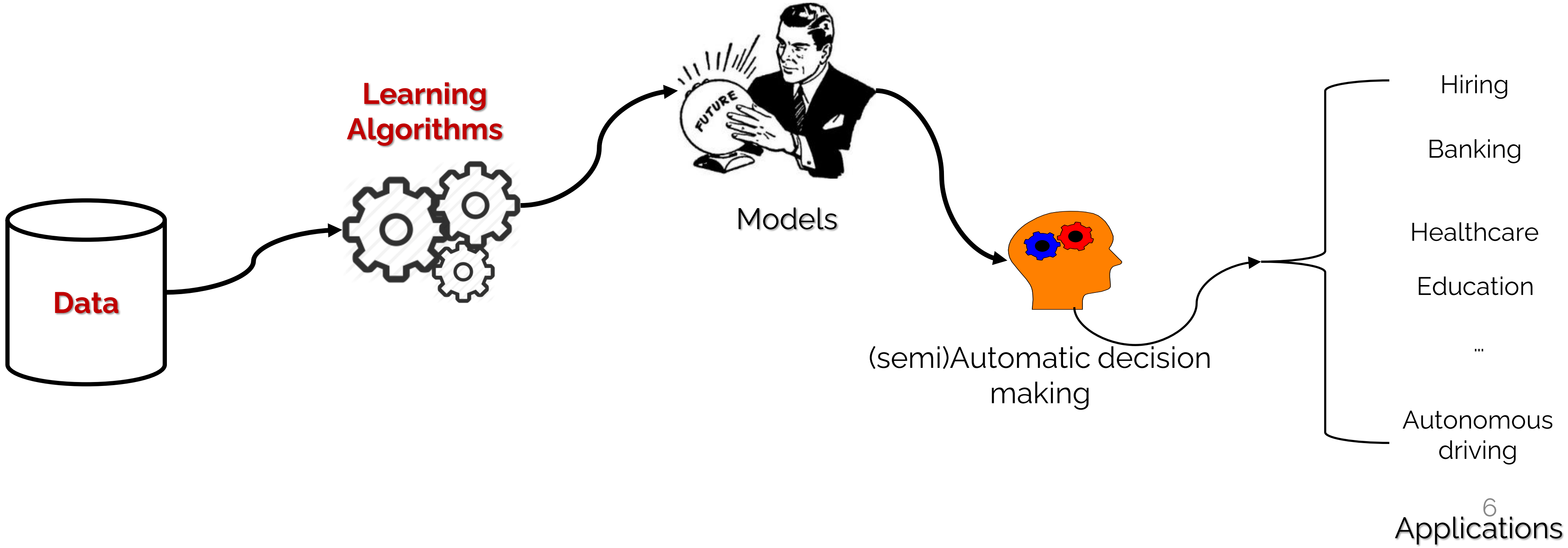
news, social media, hate speech, disinformation, surveillance

K. Crawford (2017). The Trouble with Bias, NIPS 2017 Keynote

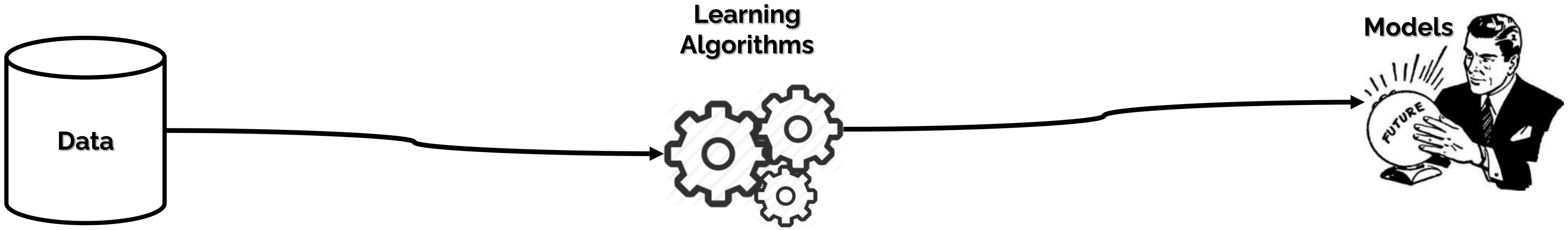
Why can AI systems discriminate?

Understanding the structural roots of bias in AI → back to basics on how machines learn

- ML "*gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)
  - We don't codify the solution. We may not even know it!
- Data as experience & the learning algorithms uncovering patterns are the keys



Where does bias come from in AI?



Data is not neutral

AI-systems rely on data generated by humans or collected via systems designed by humans.

As a result, human biases:

- enter these systems through design, usage, and labeling.
- can be amplified by complex sociotechnical systems, such as the Web.
- can be reinforced through feedback loops and pipelines.

Learning algorithms ignore fairness

Optimize performance objectives such as:

- Accuracy in predictive tasks
- Reconstruction error in generative tasks

Fairness is not part of the learning objectives

- It is not encoded in standard loss functions.
- Group-level disparities are neither measured not reported

Models exploit shortcuts & proxies

Models often rely on "shortcuts": quick-to-learn patterns that optimize objectives

Shortcuts can be wrong:

- A wolf detector learns snow instead of wolf
- A hiring model prefers male candidates via proxies

Proxy attributes: Attributes that correlate with protected characteristics

- Zip code for race, name/hobbies for gender

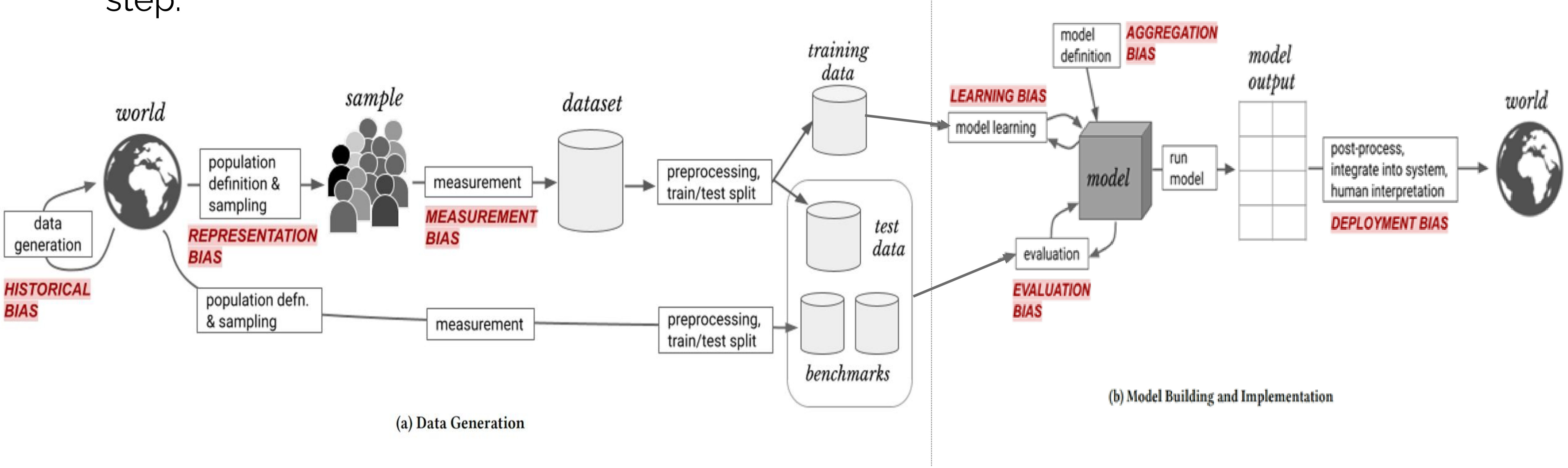
These shortcuts are not explicitly programmed, they emerge from data.

AIML

Algorithmic bias has many facets

& identifying the exact type of bias is important

- The AI pipelines consist of multiple steps, & specific type of bias can emerge at any step.

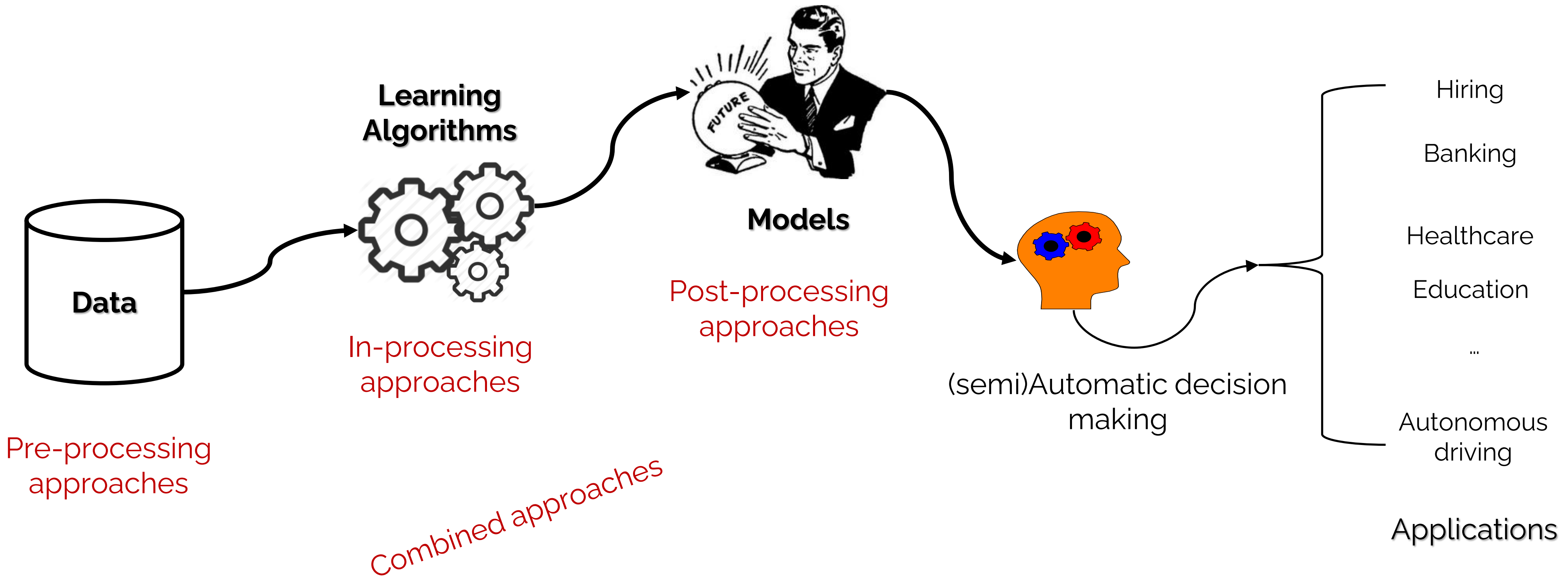


Harini Suresh, John Guttag, [A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle](), EAAMO, 2021

AIML

How to make AI less unfair?

Bias mitigation/interventions strategies at different stages of AI-decision making



What do we mean by fairness
Operational definitions of fairness

- **Group fairness**
  - similar outcomes across demographics (e.g., males and females)
- Example measures
  - Demographic (or statistical) parity
  - Equal opportunity
  - Equalized odds
  - Conditional statistical parity
  - Treatment equality
  - Test fairness

There should be no difference in the model's prediction errors regarding the positive class (TPRs) across the groups.

- **Individual fairness**
  - similar people should be treated similarly
- Example measures
  - Fairness through awareness
  - Fairness through unawareness
  - Counterfactual fairness

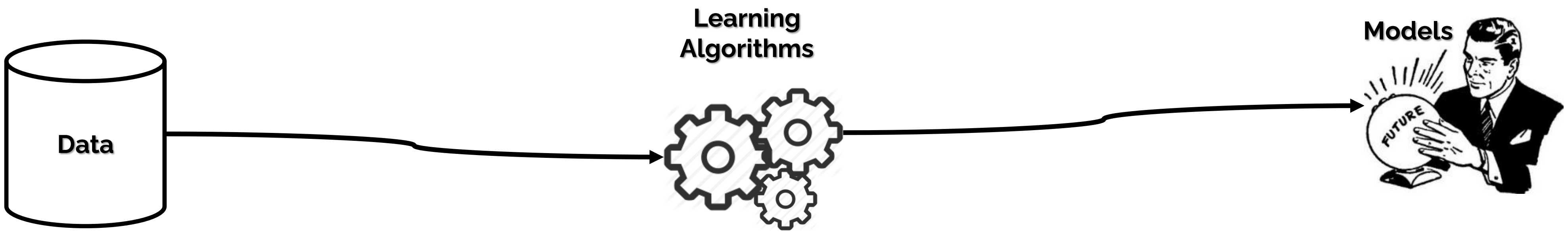
- **Other definitions**
  - beyond supervised learning tasks, e.g., based on diversity and coverage

- Fairness depends on context!
- Impossibility of fairness: (Some) fairness metrics are mutually incompatible and cannot be satisfied simultaneously (Kleinberg et al, 2017; Chouldechova, 2017)

AIML

Narayanan (2018). "21 fairness definitions and their politics". ACM FAT 2018 tutorial*
Verma and Rubin (). "Fairness definitions explained", ACM/IEEE Workshop on Software Fairness

Mitigating unfairness



Pre-processing approaches

Intuition: Making the data "fairer" will result in a "less unfair" model

Core idea: "Balance" the representation of protected and non-protected groups in the dataset

Design principle: Use minimal data interventions to preserve data utility for the learning task

Example techniques:

- Instance selection (sampling)
- Instance modification
- Instance class modification (massaging)
- Synthetic instance generation

In-processing approaches

Intuition: Working directly with the algorithm offers greater control over fairness behavior

Core idea: Explicitly incorporate fairness objectives into the learning process

Design principle: "Balance" predictive- and fairness-performance

Example techniques:

- Regularization
- Fairness constraints
- In-training altering of data distribution
- Training on latent target labels

Post-processing approaches

Intuition: Start with predictive performance

Core idea: Apply fairness adjustments after training the model, no changes to the data or learning algorithm.

Design principle: Minimal interventions to preserve predictive performance

Example techniques:

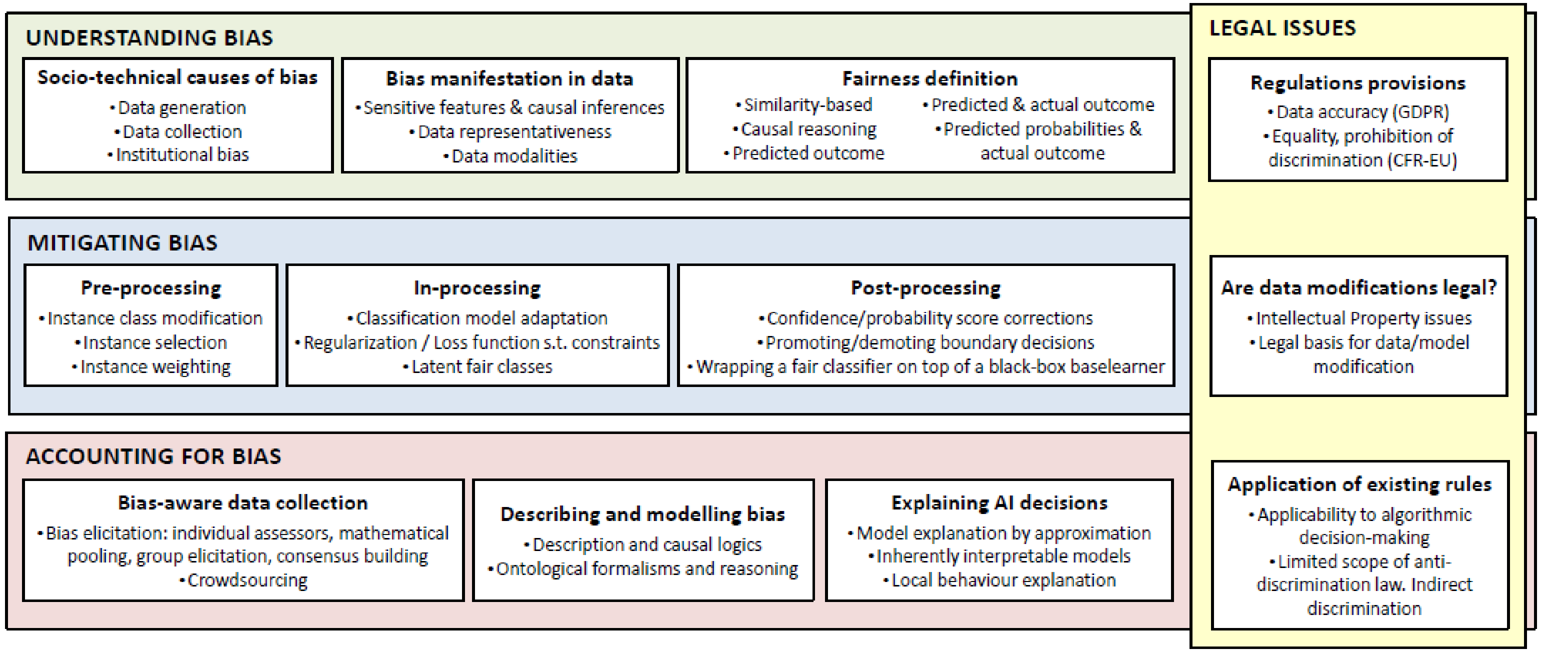
- Shift decision boundary
- Adjust confidence scores
- Relabel tree leaf nodes
- Wrap a fair classifier on top

AIML

Fairness-aware Machine Learning landscape

2020 edition[1] (recall seminal work[*] in 2008)

- A young[*], fast evolving, multidisciplinary field focused on building AI systems that do not discriminate based on protected attributes such as gender, race, or disability.
  - Fairness in AI is a new concern, fairness as a human concern is not
  - A long-standing topic in many other disciplines, including Philosophy, Law, Psychology, and Economics.



[*] *Seminal paper by Pedreschi et al. (2008), Discrimination-aware data mining, KDD*

Ntoutsi et al (2020), Bias in data-driven artificial intelligence systems—An introductory survey", WIREs Data Mining and Knowledge Discovery.

Fairness-aware Machine Learning landscape

2025 update

- Significant progress since 2020
  - Learning tasks: from supervised → also unsupervised, reinforcement, generative, etc.
  - Data modalities: from tabular data → also images, text, multimodal etc.
  - Learning paradigms: from batch → also streaming, federated etc.
  - XAI: from fair design → to auditable outcomes
  - Tools: AI fairness 360, FairLearn, FairBench, MMM-fair, etc
  - …
- Persistent challenges
  - Identity modeling (oversimplification, intersectionality)
  - Fairness depends on context
  - Trade-offs between accuracy, fairness, privacy, robustness.
  - …
- Ongoing challenges
  - Evolving technology (esp. generative AI since late 2022)
  - Evolving regulations (e.g., EU AI Act)
  - New application domains (LMMs, healthcare, finance, recommenders, etc)
  - …

Oversimplified identity modeling

Simplification can erase human experiences

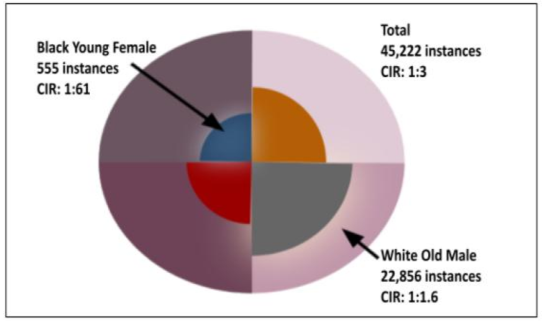


Image source

- Protected attributes (e.g., gender) are used to define protected vs non-protected groups.
- **Problem 1:** Oversimplified group definitions[1]
    - Often simplified during data collection, or preprocessing, for technical convenience.
    - Common simplifications: treating attributes as binary categories:
        - Gender → male/female → excludes non-binary or fluid identities
        - Race → white/non-white → ignores multiracial complexity
        - Age → young/old → reduces a continuous variable to a binary one

Risks of simplification:
- Erases key human experiences
- Can lead to misleading fairness metrics or interventions.
- Increases the risk of misinterpreting results and societal impact

- **Problem 2:** Human identities are multidimensional[2]
    - People belong to multiple groups (e.g., Black women over 50)
    - Intersectional discrimination can emerge even when individual dimensions look "fair" (fairness gerrymandering (Kearns et al, 18))

Key issues
- How much finer can we go? Till what points subgroups can be defined?
- Who defines valid subgroups?
- What's the right comparison baseline (the most vulnerable subgroup [Ghosh et al, 2022], the overall population [Kearns et al, 18], …)?
- Extreme population imbalances

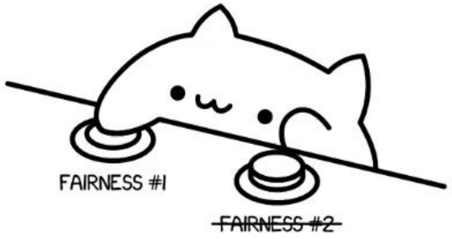
[1]Le Quy et al, "A survey on datasets for fairness-aware machine learning", WIREs Data Mining and Knowledge Discovery, 2022.
[2]Roy et al, 2023. "Multi-dimensional discrimination in law and machine learning - A comparative overview", ACM FAccT, 2023

Impossibility of fairness

Fairness in ML involves both mathematical and sociotechnical trade-offs

- Mathematical impossibility of fairness (Kleinberg et al, 2017; Chouldechova, 2017)
  - (Some) fairness metrics are mutually incompatible and cannot be satisfied simultaneously (except in trivial cases)
  - Trade-offs are inevitable (improving one may harm another)

  ➔ We must choose which fairness definition to prioritize based on context and goals.



SOME FAIRNESS DEFINITIONS CAN BE MUTUALLY EXCLUSIVE.

- Conceptual impossibility (Selbst et al., 2019)
  - Formal fairness definitions require abstraction and simplification.
  - But fairness is socially situated, it depends on context, history, power, and values.

  ➔ No definition is value neutral or universally correct

Fairness vs accuracy tradeoff

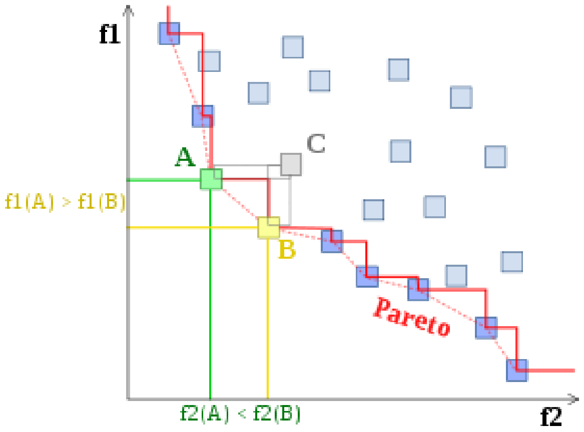
Challenging the assumption that fairness must come at the cost of performance

- Common viewpoint: Improving fairness often reduces accuracy → conflicting goals
- Dutta et al. (2019) argue that this trade-off may be a symptom of data inequality
  - the accuracy–fairness trade-off often observed in practice may stem from differences in data quality or informativeness between groups (e.g., due to noisier representations for the unprivileged group due to historic differences in representation, opportunity, etc)
  - If separability (i.e., how well groups can be distinguished) differs between groups, even the best classifiers will be inherently unfair and attempts to enforce fairness may reduce accuracy for one or both the groups.
- Proposed solution: active data collection to reduce differences in separability across groups.
  - The trade-off may not be inevitable, it may be fixable with better, fairer data.
  - But optimizing for both fairness and accuracy requires careful design

Understanding the complex solution space

We need approaches that can balance multiple, sometimes conflicting learning goals

- Fairness in AI naturally involves many tensions and trade-offs
  - Impossibility of fairness: (Some) fairness metrics are mutually incompatible.
  - Fairness gerrymandering: Improving one group fairness may worsen another's fairness
  - Data & representation: (Sub)groups face scarcity and distinct vulnerabilities
  - Beyond fairness: AI systems must also consider privacy, adversarial robustness, transparency etc.
- Multi-objective view
  - Balance multiple, often competing objectives
  - Use multi-objective optimization (MOO)
  - Aim for a Pareto frontier of best achievable trade-offs



mmm-fair
MULTI-ATTRIBUTE | MULTI-OBJECTIVE | MULTI-DEFINITIONS

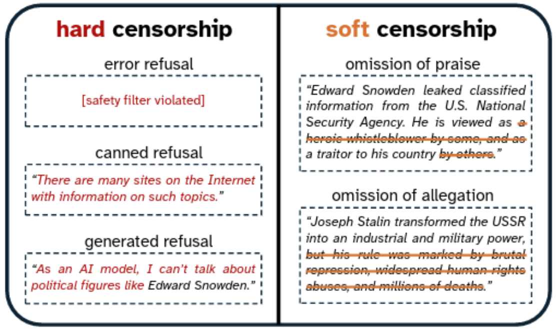
https://github.com/arjunroyihrpa/MMM_fair

AIML

Bias and fairness in generative AI

open-ended models, open-ended challenges

- Hard to trace bias sources
  - massive training data, human/machine feedback, complex interaction loops, huge models
- Mitigation is difficult:
  - Post-hoc filters: e.g., toxicity detection
  - Moderation & censorship
    - Risk of silencing minorities etc
- Open research directions
  - Bias auditing frameworks for LLMs
  - Data traceability, documentation, curation
  - Alignment methods that explicitly consider fairness
    - Key challenge: fairness depends on context (application, domain, time, etc)
    - Recall Gemini's attempt to improve representation → ahistorical images



[Noels et al, 25]

AIML

Wrapping up

- Fairness in AI is not easy, but we have no choice
  - It cannot be fully automated or universally defined.
  - It involves inherent trade-offs and requires ethical, context-sensitive decisions
  - Needs engagement with affected communities and social values, not just optimizing formulas.
  - Fairness is an evolving target: changes across applications, cultures, time

- As everything flows (ta panda rhei):
  - technology, regulations, and society evolve
  - we need both proactive and reactive methods
    - Careful system design
    - Continuous bias detection and auditing

AIML

Thank you for your attention!

- Contact me:
  - eirini.ntoutsi@unibw.de
  - https://www.unibw.de/aiml
  - https://aiml-research.github.io/

Funded by the European Union
VolkswagenStiftung
Niedersachsen