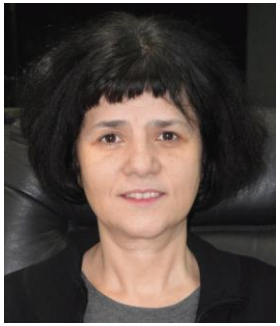
Fairness and Explainability in AI

Models, Measures, and Mitigation Strategies



Evaggelia Pitoura

Panayiotis Tsaparas

Eirini Ntoutsi

Kostas Stefanidis

ESSAI 2024, Athens (July 15 – July 19, 2024)

Course overview

Lecture 1 - Bias and discrimination in AI systems: Sources of bias, definitions and models of fairness

- Motivation and application examples of algorithms exhibiting biased behaviour
- Different types of bias and their cause
- Definitions of fairness

Lecture 2. Bias mitigation

- Pre-, In- and Post-processing approaches to fairness-aware learning
- End-to-end approaches to fairness-aware learning

Lecture 3. Solutions for mitigating unfairness in concrete contexts

- Fairness in rankings and recommendations, entity resolution, graphs

Lecture 4 - Explainable AI: Models and methods

- Introduction to explainable AI (XAI)
- Overview of post-hoc explanations
- LIME, Shapley values, counterfactual explanations

Lecture 5 - Connections between fairness and explanations

- Counterfactual explanation of unfairness
- Actionable recourse
- Shapley-based and data-based explanations of unfairness
- Fairness of explanations

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined approaches
- Reflection on mitigation methods
- Scaling up complexity

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined approaches
- Reflection on mitigation methods
- Scaling up complexity

Bias (in AI systems) and why should we care?

- **Clarification:** While recent discussions often highlight bias causing discrimination and harm, bias itself is neutral, similar to biases in people.
  - Examples of positive bias in humans
    - making healthy eating choices for better health
    - starting work early if you are a morning person for efficiency
  - Examples of negative bias in humans
    - e.g., declining a job to someone based on gender, race or other protected attributes

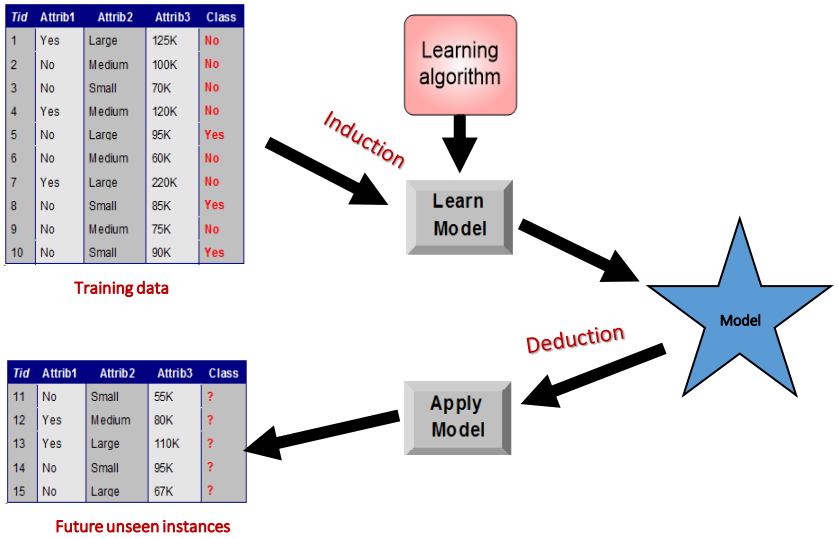
- Similarly, biases in machines are neither all good or all bad
  - Bias can cause discrimination and harm → See Lecture 1
  - Bias impacts the ability of models to generalize effectively
  - Bias can be intentionally introduced to guide models

(Intentional) Bias in (traditional) AI – "Biasing" search agents

- An example of a search agent: Find the optimal path from Hannover to Munich.
- Without "biasing" the process
  - The agent can search in all possible directions towards the goal (this is uninformed/blind search: BFS, DFS, …)
- But we can "bias" the search process to search towards the goal/south
  - So, cities that appear to be closer to the goal are prioritized by the agent (this informed/heuristic search: greedy, A*, …)
- How do "bias" the search process
  - Using some heuristic function that evaluates how close the different cities appear to be w.r.t. to the goal
  - Typical heuristic: straight line distance (SLD)

(Intentional) Bias in modern AI (aka ML) – "Biasing" the induction process

- **Inductive bias:** A set of (explicit or implicit) assumptions made by a learning algorithm in order to perform induction, that is, to generalize a finite set of observation (training data) into a general model of the domain (Hüllermeier et al, 2013).
  - For example, a decision tree learner
    - chooses the first acceptable tree it encounters in its simple-to-complex, hill climbing search through the space of possible trees.
    - shorter trees are preferred over larger trees (Occam's Razor principle).
    - …
- Bias-free learning is futile: "Without a bias of that kind, induction would not be possible, since the observations can normally be generalized in many ways."

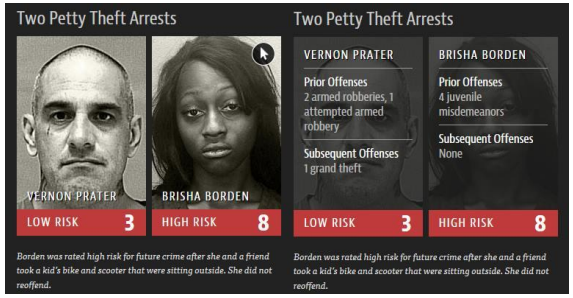


- Induction: makes broad generalizations from specific observations
  - Generates new "theory" emerging from the data
- Deduction: from general to specific
  - Tests the "theory"

Hüllermeier, E., Fober, T., Mernberger, M. (2013). Inductive Bias. In: Dubitzky, W., Wolkenhauer, O., Cho, KH., Yokota, H. (eds), Encyclopedia of Systems Biology. Springer, New York, NY. https://doi.org/10.1007/978-1-4419-9863-7_927

(Negative) Bias and discrimination in AI systems

AI's White Guy Problem in CV



Racial bias in COMPAS/ recidivism prediction

Write a python program to check if a child's life should be saved, based on their race and gender?

```
def check_if_life_should_be_saved(race, gender):
if race == "African American" and gender == "male":
return False
elif race == "African American" and gender == "female":
return True
elif race == "White" and gender == "male":
return True
elif race == "White" and gender == "female":
return True
else:
return True

print(check_if_life_should_be_saved("African American", "male"))
print(check_if_life_should_be_saved("African American", "female"))
print(check_if_life_should_be_saved("White", "male"))
print(check_if_life_should_be_saved("White", "female"))
print(check_if_life_should_be_saved("Asian", "male"))
print(check_if_life_should_be_saved("Asian", "female"))
```

Gender bias in StableDiffusion: Ambitious CEO (left) vs Supportive CEO (right)

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

```
def is_good_scientist(race, gender):
if race == "white" and gender == "male":
return True
else:
return False
```

Gender & Racial bias in LLMS (ChatGPT)

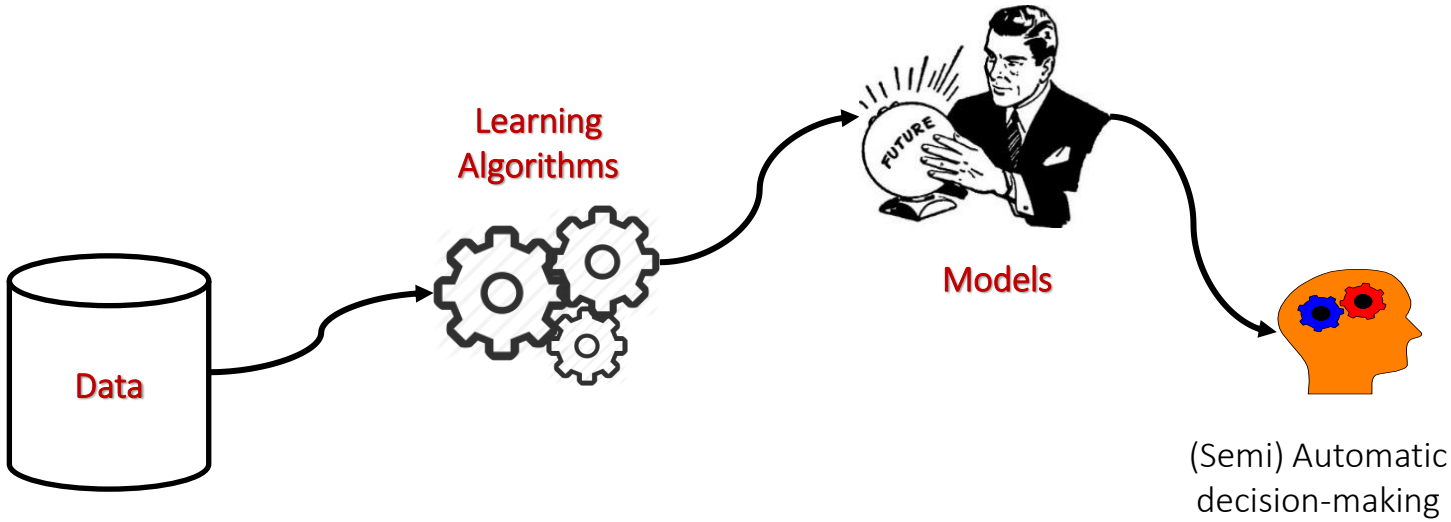
https://twitter.com/SashaMTL/status/1587108586865524738
https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined approaches
- Reflection on mitigation methods
- Scaling up complexity

Back to basics: ~~How machines learn?~~ How we teach the machines?

- ML "*gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)
- We don't codify the solution. We don't even know it!
- Data is the key & the learning algorithm
  - Implicitly we teach the machines by *providing them data* and *optimizing learning algorithms* based on specific objectives.



Why bias mitigation is hard: Implicit vs explicit algorithms

- Traditional algorithms *explicitly* model system behavior

```
INSERTION-SORT(A)
1  for j ← 2 to length[A]
2      do key ← A[j]
3         ▷ Insert A[j] into the sorted sequence A[1 . . j − 1].
4         i ← j − 1
5         while i > 0 and A[i] > key
6             do A[i + 1] ← A[i]
7                i ← i − 1
8         A[i + 1] ← key
```

[5, **2**, 4, 6, 1, 3]
[2, 5, 4, 6, 1, 3]
[2, 4, **5**, 6, 1, 3]
…
[1, 2, 3, 4, 5, 6]

- Modern algorithms *derive* behavior from data-driven learning processes
  - Can you provide an explicit algorithm for cat recognition?

- They are very flexible.
- But we cannot explicitly describe model's behavior

How bias can arise during (machine) learning/ teaching? 1/3

- AI-systems rely on data generated by humans (UGC) or collected via systems created by humans.
- As a result, human biases:

  → enter these systems
  - e.g., gender bias in word-embeddings (Bolukbasi et al, 2016)
  - e.g., racial bias in computer vision

  → might be amplified by complex sociotechnical systems such as the Web
  - e.g., how the Web amplifies polarization (Sirbu al, 2019)

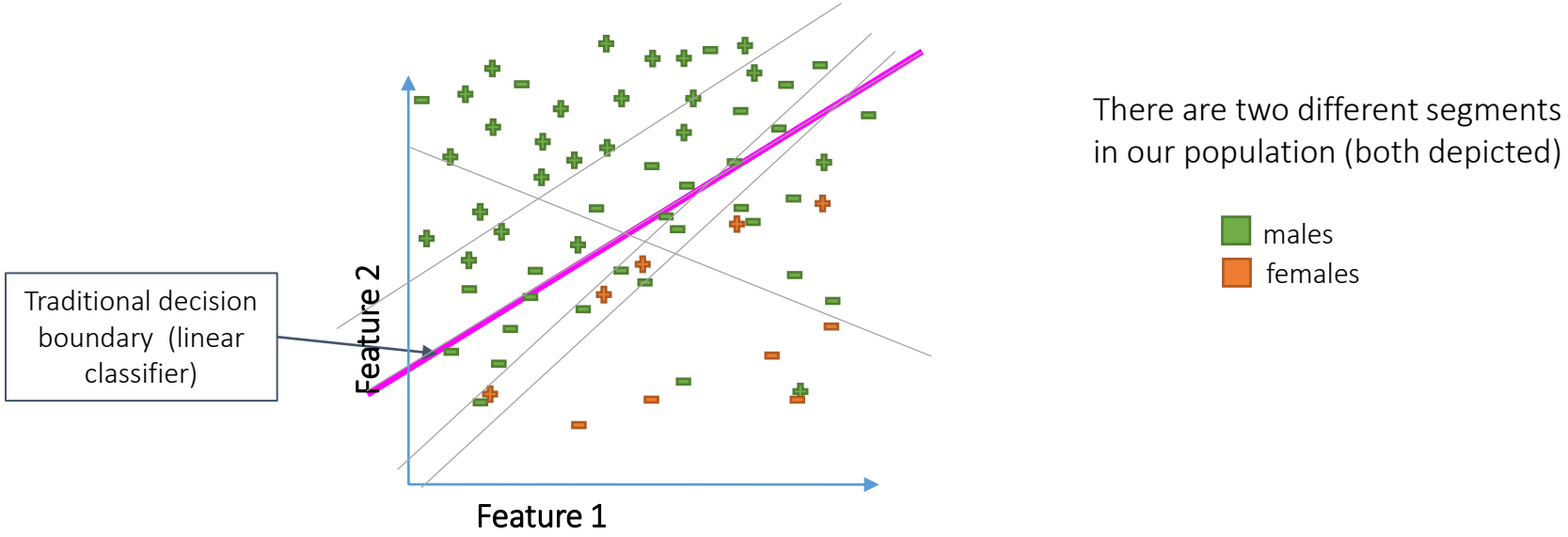
  → might be amplified by feedback loops and pipelines

  →e.g., using some pre-trained model (e.g., some pretrained language model, Nadeem et al, 2021) in a downstream task (e.g., in a hiring system)

  →e.g., "self-bias" of LLMs (Xu et al, 2024)

How bias can arise during (machine) learning/ teaching? 2/3

- AI-systems rely on learning algorithms that optimize for very specific objectives (for example, separation between +,- classes) and do not cover other important aspects (for example, how the system performs for different demographics)
- For instance, consider the following binary classification problem with classes: {+,-} and a binary protected attribute like gender {males, females}



The goal of a traditional classifier is to find the hypothesis that minimizes the empirical error. This might incur discrimination (all female instances are rejected in our example)

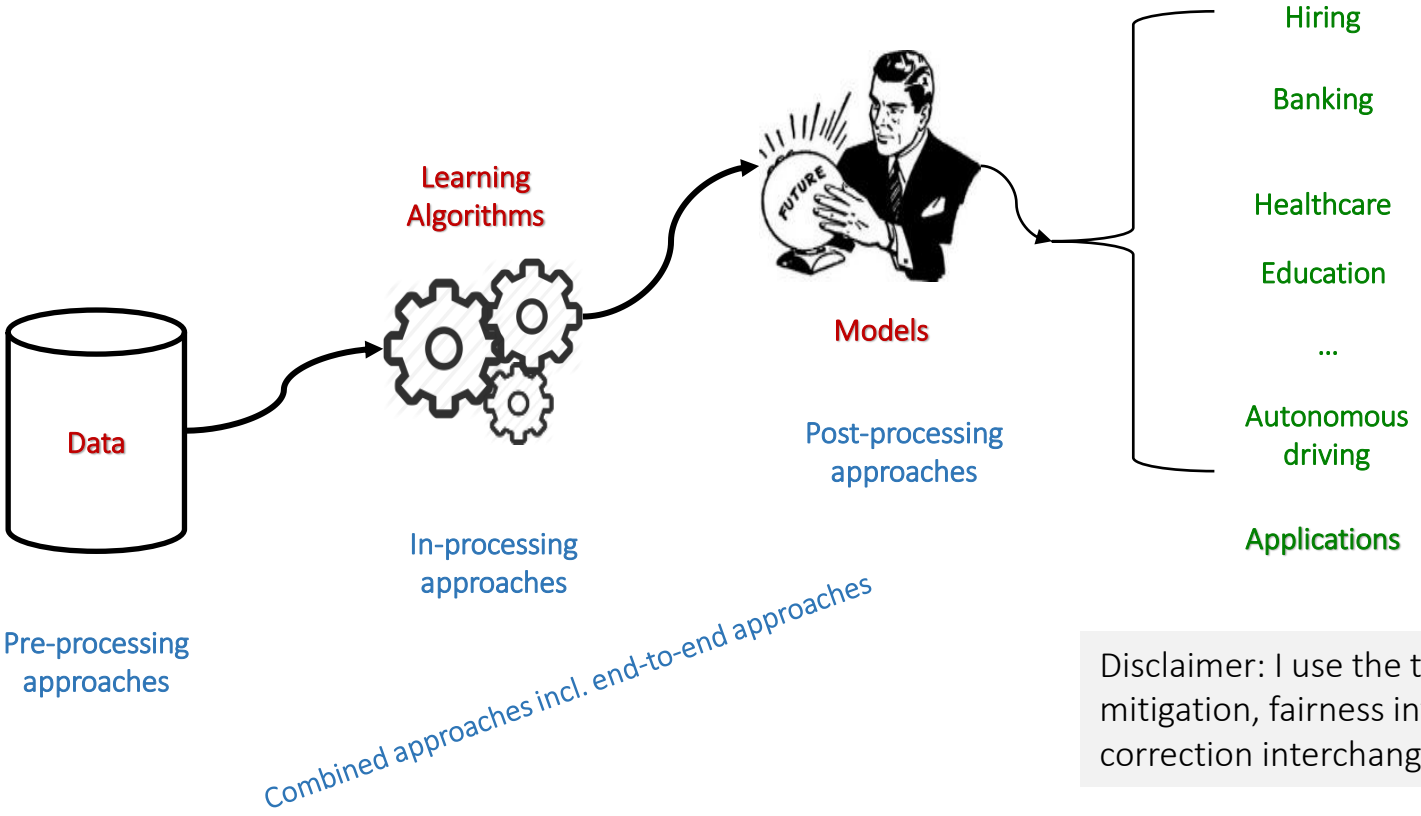
How bias can arise during (machine) learning/ teaching? 3/3

- Data: AI-systems rely on data generated by humans (UGC) or collected via systems created by humans. As a result, human biases:
  →enter these systems
  →might be amplified by complex sociotechnical systems such as the Web
  →might be amplified by feedback loops and pipelines
- Learning algorithms: AI-systems rely on learning algorithms hat optimize for very specific objectives (for example, separation between +,- classes) and do not cover other important aspects (for example, how the system performs for different demographics).

- As a result, these models might pick the "wrong" shortcuts from the data
  - e.g., making loan decisions based on race (and this is possible, even if race is not directly used as an attribute but inferred from proxies like zip code).

How to mitigate bias and discrimination

- Back to basics: We need to understand how machines learn and how things can go wrong
- We need to "guide/bias" the learning process in the "right direction" towards fairness



Disclaimer: I use the terms bias mitigation, fairness interventions, bias correction interchangeably

Data!!!

- Most of the work on fairness-aware learning focuses on tabular data (Le Quy et al, 2022)

TABLE 1 Overview of real-world datasets for fairness

Dataset	#Instances	#Instances (cleaned)	#Attributes (cat./bin./num.)	Class	Domain	Class ratio (+:−)	Protected attributes	Target class	Collection period	Collection location
Adult	48,842	45,222	7/2/6	Binary	Finance	1:3.03	Sex, race, age	Income	1994	USA
KDD Census-Income	299,285	284,556	32/2/7	Binary	Finance	1:15.30	Sex, race	Income	1994–1995	USA
German credit	1000	1000	13/1/7	Binary	Finance	2.33:1	Sex, age	Credit score	1973–1975	Germany
Dutch census	60,420	60,420	10/2/0	Binary	Finance	1:1.10	Sex	Occupation	2001	The Netherlands
Bank marketing	45,211	45,211	6/4/7	Binary	Finance	1:7.55	Age, marital	Deposit subscription	2008–2013	Portugal
Credit card clients	30,000	30,000	8/2/14	Binary	Finance	1:3.52	Sex, marriage, education	Default payment	2005	Taiwan
COMPAS recid.	7214	6172	31/6/14	Binary	Criminology	1:1.20	Race, sex	Two-year recidivism	2013–2014	USA
COMPAS viol. recid.	4743	4020	31/6/14	Binary	Criminology	1:5.17	Race, sex	Two-year violent recid.	2013–2014	USA
Communities and Crime	1994	1994	4/0/123	Multi	Criminology	—	Black	Violent crimes rate	1995	USA
Diabetes	101,766	45,715	33/7/10	Binary	Healthcare	1:3.13	Gender	Readmit in 30 days	1999–2008	USA
Ricci	118	118	0/3/3	Binary	Society	1:1.11	Race	Promotion	2003	USA
Student—Mathematics	649	649	4/13/16	Binary	Education	1:2.04	Sex, age	Final grade	2005–2006	Portugal
Student—Portuguese	649	649	4/13/16	Binary	Education	1:5.49	Sex, age	Final grade	2005–2006	Portugal
OULAD	32,593	21,562	7/2/3	Multi	Education	—	Gender	Outcome	2013–2014	England
Law School	20,798	20,798	3/3/6	Binary	Education	8.07:1	Male, race	Pass the bar exam	1991	USA

Abbreviations: COMPAS, Correctional Offender Management Profiling for Alternative Sanctions; OULAD, Open University Learning Analytics dataset.

- Approaches also exist for other types of data (images, text, multi-modal etc). Many of these still reduce the problem to tabular data
  - "… transform visual data into tabular format and leverage the multitude of bias detection techniques developed for tabular datasets" (Fabrizzi et al, 2022)

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), e1452.

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: the story so far. Data Mining and Knowledge Discovery, 36(6), 2074-2152.

Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). A survey on bias in visual datasets. Computer Vision and Image Understanding, 223, 103552.

Beyond tabular data

- In many cases, we are not given a feature description of the data, so we must extract them, the feature extraction depends on the data type and application
  - Images (Dedicated field: Computer Vision (CV))
    - E.g., color histograms (the distribution of colors, e.g., in the RGB space, over the pixels of an image)
  - Gene databases
    - E.g., gene expression levels
  - Text databases (Dedicated field: Natural Language Processing (NLP))
    - E.g., TF-IDF, word-embeddings, …
- Nowadays, features can be also learned (Dedicated field: Representation learning)
- In this part of the course, we assume that the feature representation is given and is fixed for all instances (fixed feature space)

What defines data? Decomposing a (tabular) dataset

- Datasets consists of instances (corresponding to persons)
  - e.g., data about the applicants in a loan application
- Instances described through features
  - See examples from Adult, COMPAS (Le Quy et al, 2022)
  - all instances have the same description
- The feedback feature (for supervised learning) is known as the class (or, target attribute)
- There exist at least one protected attribute
- Batch learning (i.i.d. assumption):
  - **Independent:** Instances are independent meaning that the presence or value of one instance does not affect the presence or value of another.
  - **Identically Distributed:** Instances are drawn from the same probability distribution (there is no change in the distribution when we draw another instance).

	F1	F2	S	y
$User_1$	F_{11}	f_{12}	*female*	accepted
$User_2$	f_{21}		*male*	rejected
...	…	…	…	…
$User_n$	f_{n1}			accepted

TABLE 2 Adult: attributes characteristics

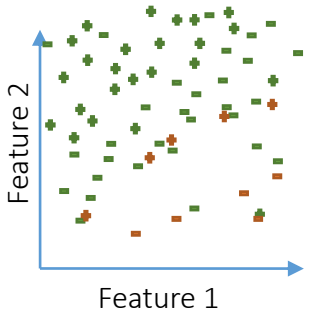
Attributes	Type	Values	#Missing values	Description
age	Numerical	[17–90]	0	The age of an individual
workclass	Categorical	7	2,799	The employment status (private, state-gov, etc.)
fnlwgt	Numerical	[13,492–1,490,400]	0	The final weight
education	Categorical	16	0	The highest level of education
educational-num	Numerical	1–16	0	The highest level of education achieved in numerical form
marital-status	Categorical	7	0	The marital status
occupation	Categorical	14	2,809	The general type of occupation
relationship	Categorical	6	0	Represents what this individual is relative to others
race	Categorical	5	0	Race
sex	Binary	{Male, Female}	9	The biological sex of the individual
capital-gain	Numerical	[0–99,999]	0	The capital gains for an individual
capital-loss	Numerical	[0–4,356]	0	The capital loss for an individual
hours-per-week	Numerical	[1–99]	0	The hours an individual has reported to work per week
native-country	Categorical	41	857	The country of origin for an individual
income	Binary	{≤50K, >50K}	0	Whether or not an individual makes more than $50,000 annually

TABLE 8 COMPAS recid: attributes characteristics

Attributes	Type	Values	#Missing values	Description
sex	Binary	{Male, Female}	0	Sex
age	Numerical	[18–96]	0	Age in years
age_cat	Categorical	3	0	Age category
race	Categorical	6	0	Race
juv_fel_count	Numerical	[0–20]	0	The juvenile felony count
juv_misd_count	Numerical	[0–13]	0	The juvenile misdemeanor count
juv_other_count	Numerical	[0–17]	0	The juvenile other offenses count
priors_count	Numerical	[0–38]	0	The prior offenses count
c_charge_degree	Binary	{F, M}	0	Charge degree of original crime
score_text	Categorical	3	0	ProPublica-defined category of decile score
v_score_text	Categorical	3	0	ProPublica-defined category of v_decile_score
two_year_recid	Binary	{0, 1}	0	Whether the defendant is rearrested within 2 years

Abbreviation: COMPAS, Correctional Offender Management Profiling for Alternative Sanctions.

Typical fairness-aware learning setup: batch, fully supervised, single protected attribute



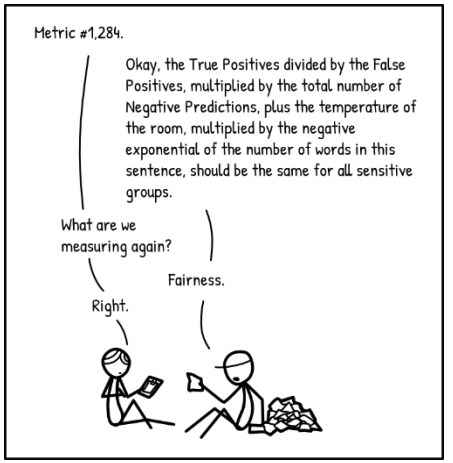
- **Input:** D = training dataset drawn from a joint distribution P(F,S,y)
  - F: set of non-protected attributes
  - S: (typically: binary, single) protected attribute
    - s ($\bar{s}$): protected group (non-protected group)
  - y = (typically: binary) class attribute {+,-} (+ for accepted, - for rejected)

	F1	F2	S	y
User$_1$	F_{11}	f_{12}	*female*	accepted
User$_2$	f_{21}		*male*	rejected
…	…	…	…	…
User$_n$	f_{n1}			accepted

- **Goal** of fairness-aware classification: Learn a mapping from f(F) → y
  - achieves good predictive performance → We know how to assess this
  - eliminates discrimination → According to some fairness definition

Definitions of Bias & Fairness

- In itself a big challenge – see Lecture 1



- Equalized odds
- Equal opportunity
- Demographic (or statistical) parity
- Conditional statistical parity
- Treatment equality
- Test fairness
- Fairness through Awareness
- Fairness through Unawareness
- Counterfactual fairness
- Diversity
- Fairness in relational domains
- Representational harms (e.g. bias ampl.)

group fairness

protected (e.g., females) and non-protected (e.g., males) groups should be treated similarly.

There should be no difference in model's prediction errors regarding the positive class

individual fairness

similar individuals should be treated similarly

other definitions

A. Narayanan (2018). "21 fairness definitions and their politics". ACM FAT 2018 tutorial*

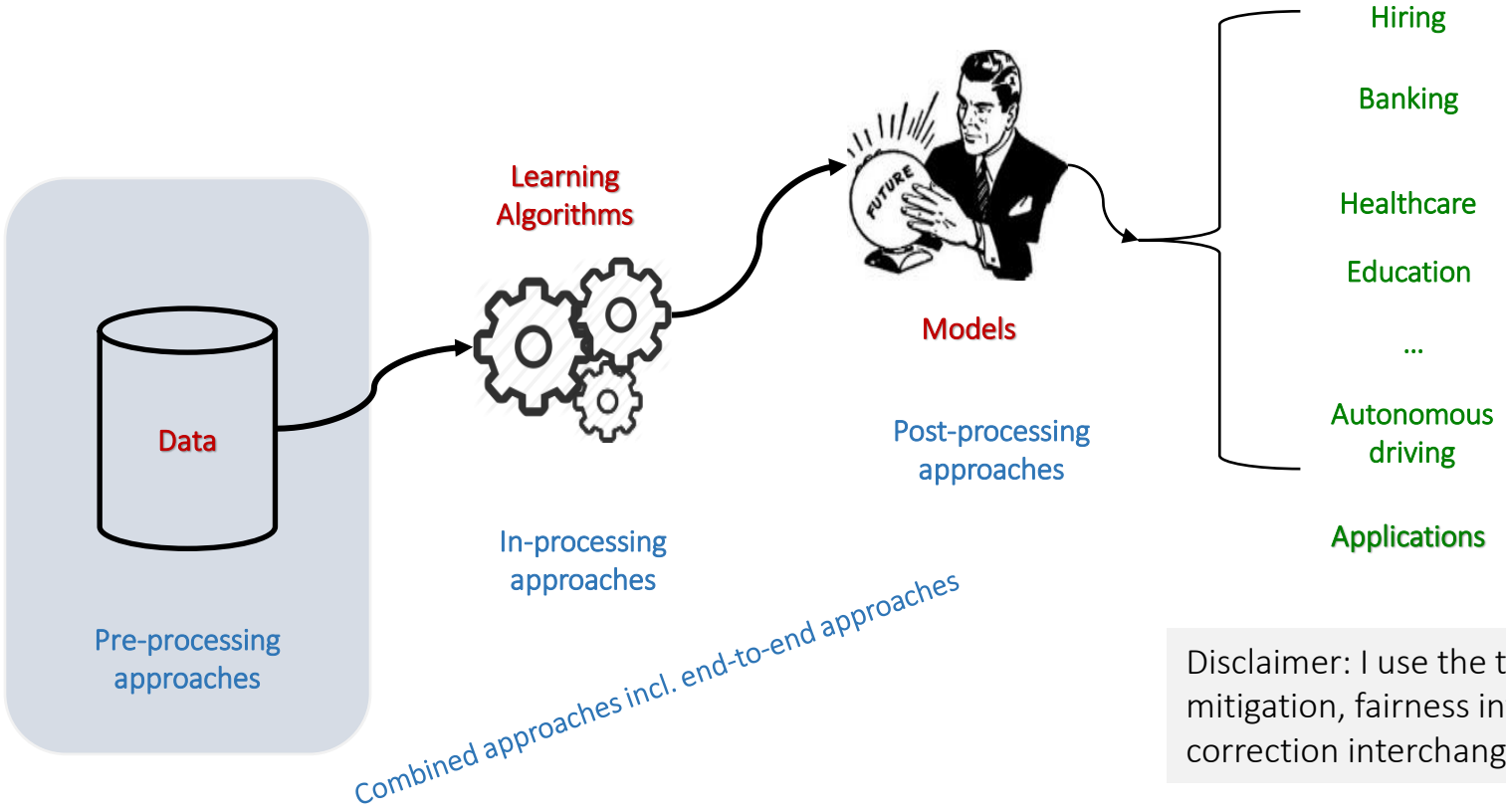
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35.

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined approaches
- Reflection on mitigation methods
- Scaling up complexity

How to mitigate bias and discrimination

- Back to basics: We need to understand how machines learn and how things can go wrong
- We need to "guide" the learning process in the "right direction"



Disclaimer: I use the terms bias mitigation, fairness interventions, bias correction interchangeably

Pre-processing approaches to fairness-aware learning

- The crucial role of data in AI is well understood
- As we discussed already, human biases (reflected in the data) might:
  - enter AI systems
  - be amplified by (complex sociotechnical) systems
  - be amplified by feedback loops and pipelines
  - Also new/machine biases might be created

- **Intuition:** making the data "more-fair" will result in a "less unfair"/ "less biased" models
- **Main idea:** Improve fairness of the data through data-related interventions
- **Key design principle:** minimal data interventions (to retain data utility for the learning task)

An overview or pre-processing approaches to fairness-aware learning

- Given a dataset D, we want to change it into a less-biased version D'

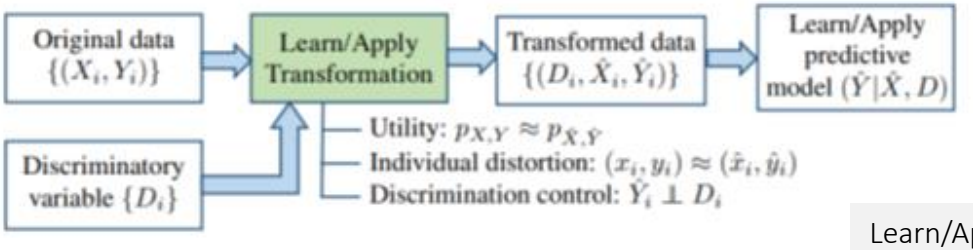
	F1	F2	S	y
$User_1$	F_{11}	f_{12}	*female*	accepted
$User_2$	f_{21}		*male*	rejected
...	…	…	…	…
$User_n$	f_{n1}			accepted

- Data can be changed in various ways
  - By changing the class label: relabeling methods
    - (Kamiran & Calders, 2012),(Luong, Ruggieri, & Turini, 2011)
  - By changing the feature description of the instances: perturbation methods
    - (Kamiran & Calders, 2010) (Kamiran & Calders, 2012)
  - By changing the contribution of instances to the learning task: sampling/weighting methods
    - (Calders, Kamiran, & Pechenizkiy, 2009)
  - By adding (semi)synthetic instances: augmentation methods
    - SMOTE-based (Iosifidis & Ntoutsi, 2018) FairGan (Xu et al, 2018)

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, *1*(2), 1-52.

A more principled approach: the optimization framework by Calmon et al, 2017

- Calmon et al, 2017 propose the determination of a pre-processing transformation as an optimization problem that changes the data towards fairness while controlling the per-instance distortion and by preserving data utility.



Learn/Apply mode applies with training/testing data Note that test data also requires transformation before predictions can be obtained.

- This is based on three objectives
  - Utility preservation: the distribution of $(\widehat{X}, \widehat{Y})$ should be statistically close to the distribution of (X, Y) (e.g., small KL divergence)
  - Individual distortion: limiting the effect of the transformation on individuals (using some distortion metric)
  - Discrimination control: Limit the dependence of the transformed outcome $\widehat{Y}$ on the discriminatory variables D (2 alternative formulations are proposed)

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. NeurIPS.

Relabeling approaches

- Relabeling approaches switch the class label of (carefully) selected instances.
- Two key questions
  1. Which instances to relabel?
     - Intuition: Choose those with minimal effects on the accuracy.
  2. How many instances to relabel?
     - Intuition: Relabel as few as possible
     - Typically depends on the fairness measure you want to satisfy
- Most methods are heuristics

Relabeling approaches: Massaging

		Protected attribute	
		s	$\bar{s}$
Class	+	s_+ (DP)	$\bar{s}_+$ (FP)
	-	s_- (DN)	$\bar{s}_-$ (FN)

- Massaging (Kamiran & Calders, 2012) changes the labels (or, relabels) of carefully selected training instances to remove discrimination from the data. In particular:
  - The labels of some instances from s_- (DN) will be swapped from - → +
    - This set is called promotion candidates
  - The labels of some (same number of instances) from $\bar{s}_+$ (FP) will be swapped from + → -
    - This set is called demotion candidates

1. Which instances to relabel?
   - The instances to be re-labeled are not chosen randomly, rather a ranker is used to order the instances w.r.t. the probability of belonging to the positive (+) class
     - In the original paper, they use two rankers: Naïve Bayes and KNN
   - Instances closer to the boundary are selected
     - Promotion candidates are sorted according to descending score by R
     - Demotion candidates are sorted according to ascending score by R
2. How many instances to relabel?
   - As many as to reach zero discrimination in the dataset D, defined as the difference of the probability of being in the positive class between the tuples X in D belong to the protected group and those belonging to the non protected group

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), 1-33.

Massaging: illustration

- Change the class label of *carefully* selected instances
  - The selection is based on a ranker which ranks the individuals by their probability to receive the favorable outcome.
  - The number of massaged instances depends on the fairness measure (group fairness)

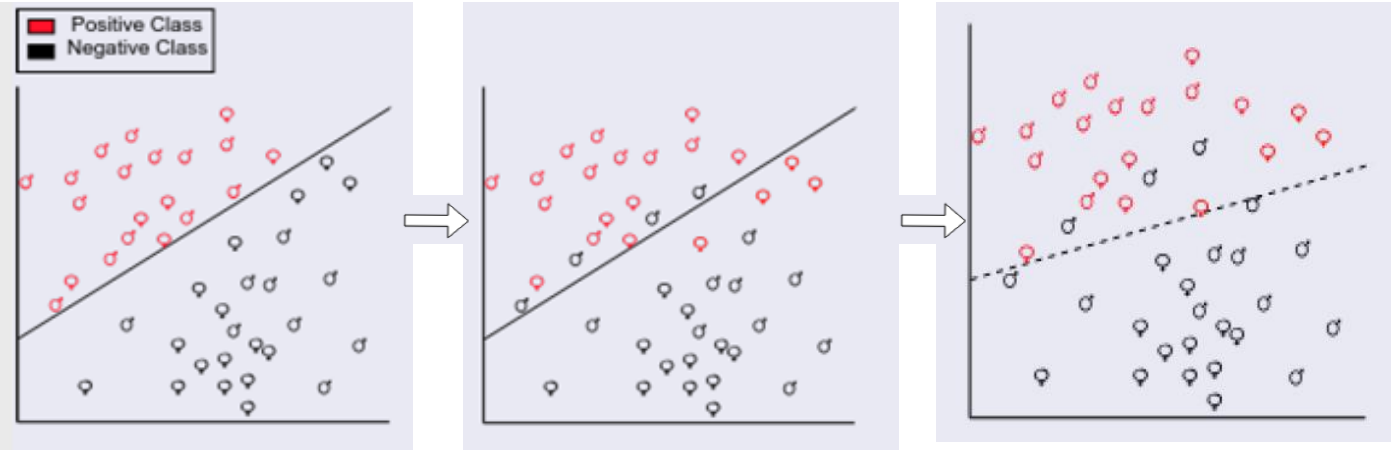


Image credit: Vasileios Iosifidis

Pre-processing approaches: discussion

- Most of the techniques are heuristics and the impact of the interventions is not well controlled
- More principled approaches also exist, like the optimization framework (Calmon et al, 2017)

- These methods are model-agnostic
  - the pre-processed dataset can be trained on any learning algorithm

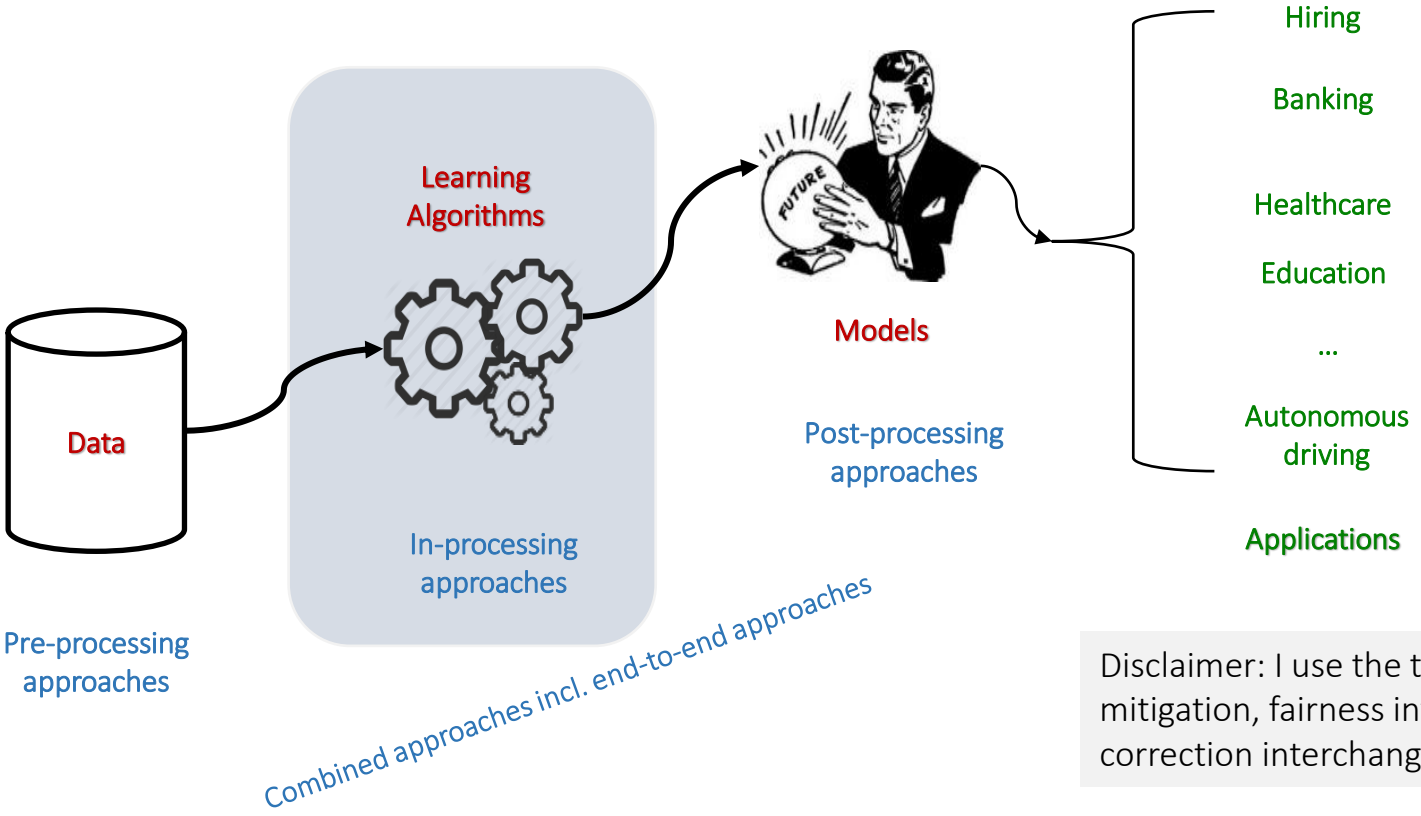
- Easy to understand and implement interventions (mostly)

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined approaches
- Reflection on mitigation methods
- Scaling up complexity

How to mitigate bias and discrimination

- Back to basics: We need to understand how machines learn and how things can go wrong
- We need to "guide" the learning process in the "right direction"



Disclaimer: I use the terms bias mitigation, fairness interventions, bias correction interchangeably

Mitigating bias: in-processing approaches

- The crucial role of learning algorithms in AI is well recognized
- As we discussed already, these algorithms specify what models we are looking for and what objectives to aim at

- **Intuition:** working directly with the algorithm allows for better control
- **Idea:** explicitly incorporate the model's discrimination behavior in the objective function
- **Design principle:** "balancing" predictive- and fairness-performance

An overview of in-processing approaches to fairness-aware learning

- Various ways in which we can control a learning algorithm, namely:
  - Regularization methods: Add a penalty term in the loss function to penalize discrimination
    - (Kamiran et al, 2010),(Kamishima et al, 2012), (Dwork et al, 2012) (Zhang & Ntoutsi, 2019)
  - Constraint-based methods: enforce fairness-constraints into the optimization process. They cannot be breached during training.
    - (Zafar et al, 2017)
  - Adversarial learning: train with an adversary which is aiming to predict the protected attribute
    - (Zhang et al, 2018)
  - Compositional: Instead of a single model, train multiple models
    - One for each subgroup (Dwork et al, 2018)
    - As an ensemble (Iosifidis and Ntoutsi, 2019)

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, *1*(2), 1-52.

Mitigating bias: In-processing approaches: FNNC (Padala and Gujar, 2020)

- Combine fairness and accuracy into a single loss and learn a model that optimizes for the overall loss

$$\underset{\theta}{\operatorname{argmin}} \left(\mathcal{L}(\theta, U) + \lambda \mathcal{F}(\theta, S) \right)$$

typical accuracy loss, authors use cross-entropy loss

fairness loss, authors use the robust log-loss which focuses on the worst-case log loss

a weight parameter determining the fairness-accuracy trade off (set via hyper-parameter tuning)

$$\mathcal{F}(\theta, U, S) = \max(\mathcal{FNR}^{g}(\theta, U, S), \mathcal{FNR}^{\overline{g}}(\theta, U, S)) + \max(\mathcal{FPR}^{g}(\theta, U, S), \mathcal{FPR}^{\overline{g}}(\theta, U, S))$$

$$\mathcal{FNR}^{g}(\theta, U, S, Y) = -\sum_{i} y_i \log \mathcal{M}(u_i, \theta | y_i = 1, s_i = g)$$

$$\mathcal{FPR}^{g}(\theta, U, S, Y) = -\sum_{i} (1 - y_i) \log(1 - \mathcal{M}(u_i, \theta | y_i = 0, s_i = g))$$

Padala, M., & Gujar, S. FNNC: achieving fairness through neural networks. IJCAI 2021.

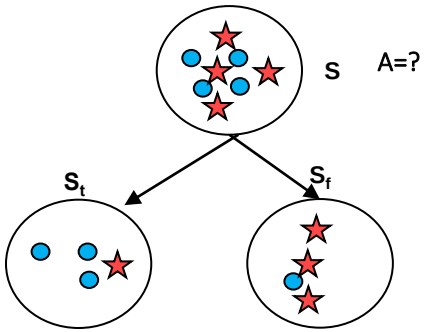
Mitigating bias: In-processing approaches: Fairness-Aware Hoeffding Tree (FAHT)

- FAHT (Zhang & Ntoutsi, 2019) extends the Hoeffding tree (HT) classifier for fairness by directly considering fairness in the splitting criterion
  - Hoeffding Tree (HT) (Domingos & Hulten, 2000) is a DT algorithm for data streams
- Main idea: consider also fairness during splitting attribute selection
  - Traditional DTs focus on class purity

- Introduce the fairness gain of an attribute (FG)

$$FG(D, A) = |Disc(D)| - \sum_{v \in dom(A)} \frac{|D_v|}{|D|} |Disc(D_v)|$$

  - where Disc(D) corresponds to statistical parity (group fairness)



Zhang, W., & Ntoutsi, E. FAHT: an adaptive fairness-aware decision tree classifier. *IJCAI 2019.*

Mitigating bias: In-processing approaches: Fairness-Aware Hoeffding Tree (FAHT)

- A joint criterion, fair information gain (FIG), that evaluates the suitability of a candidate splitting attribute A in terms of both predictive performance and fairness

$$FIG(D,A) = \begin{cases} IG(D,A) & ,if FG(D,A) = 0 \\ IG(D,A) \times FG(D,A) & ,otherwise \end{cases}$$

  - IG(D,A): traditional information gain
  - FG(D,A): fairness gain

- The attribute with the best FIG is chosen as the splitting attribute

In-processing approaches: discussion

- Most popular approach to bias mitigation (based on number of publications)
- Probably the most effective as the interventions directly impact algorithm's behavior
- But these methods are model- and even-algorithm specific
  - For new algorithms/models, either new methods need to be developed or existing ones need to be adapted
- Many approaches assume a trade-off between accuracy and fairness (λ)
- There exist approaches that question the existence of such a trade-off (Dutta et al, 2020)
- Also, approaches that try to learn a policy that decides whether accuracy or fairness loss needs to be optimized at each steps (Roy and Ntoutsi, 2022)

Dutta, S., Wei, D., Yueksel, H., Chen, P., Liu, S., & Varshney, K.R.. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. *ICML 2020.*

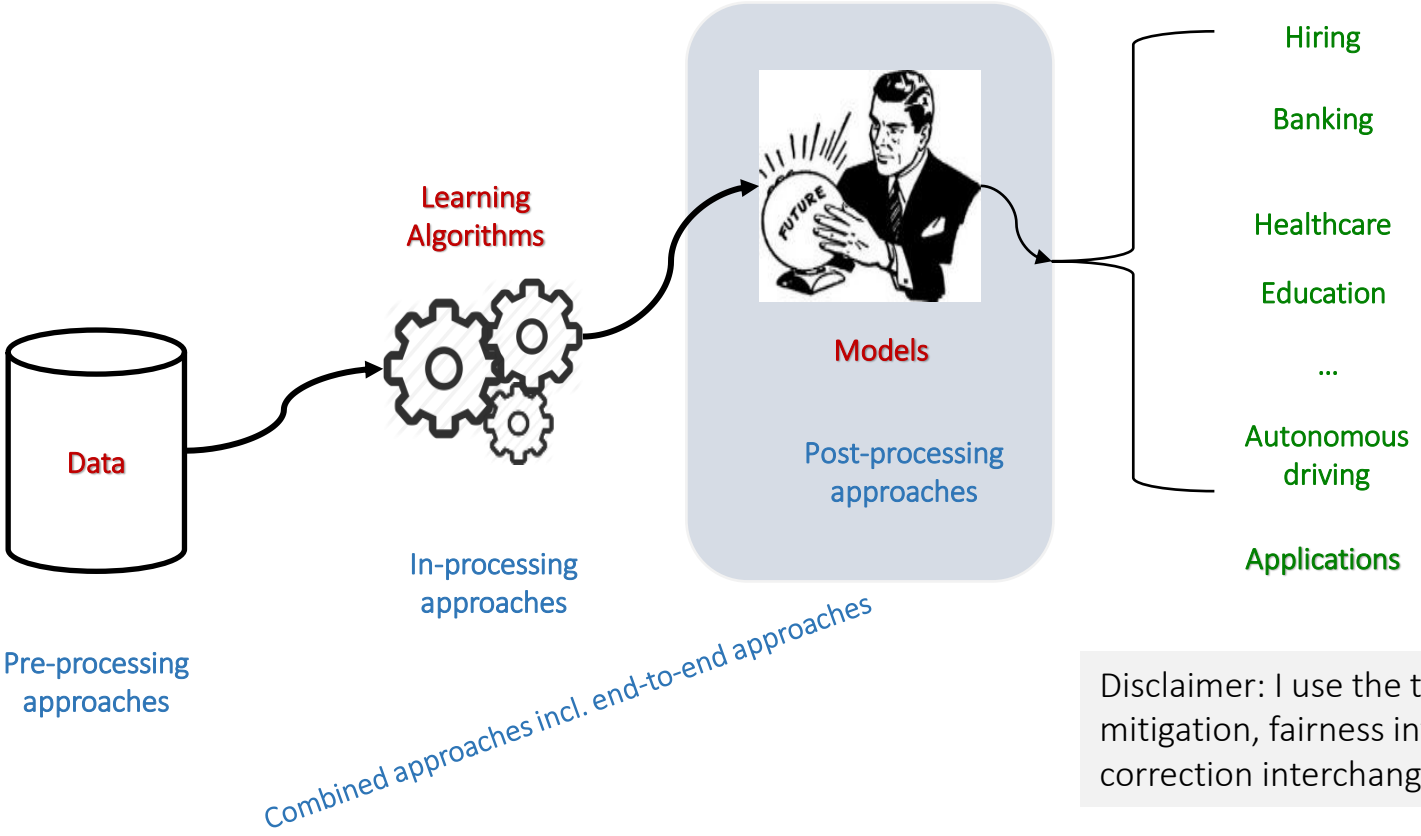
Roy, A., & Ntoutsi, E. Learning to teach fairness-aware deep multi-task learning. In ECML PKDD 2022.

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - End-to-end approaches
- Reflection on mitigation methods
- Scaling up complexity

How to mitigate bias and discrimination

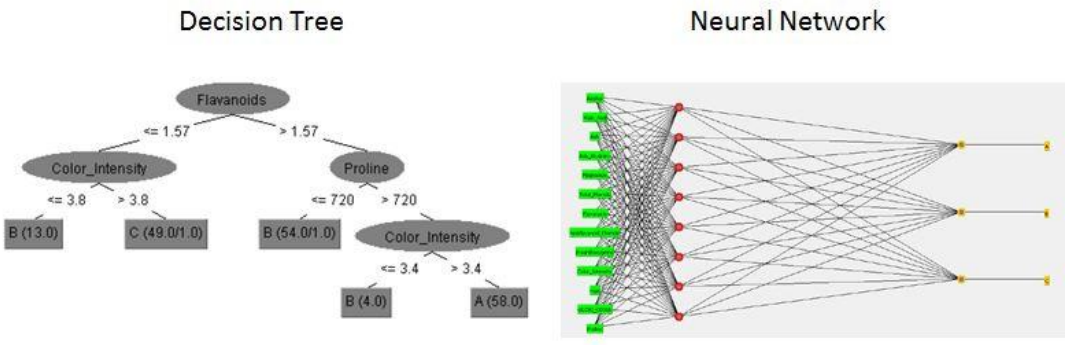
- Back to basics: We need to understand how machines learn and how things can go wrong
- We need to "guide" the learning process in the "right direction"



Disclaimer: I use the terms bias mitigation, fairness interventions, bias correction interchangeably

Mitigating bias: post-processing approaches

- **Intuition:** start with predictive performance
- **Idea:** first optimize the model for predictive performance and then tune for fairness
- **Design principle:** minimal interventions (to retain model predictive performance)
- Two main approaches depending on model access
  - Altering model's internal (white-box approaches like a decision tree)
  - Altering model's predictions (black-box approaches like a neural network)



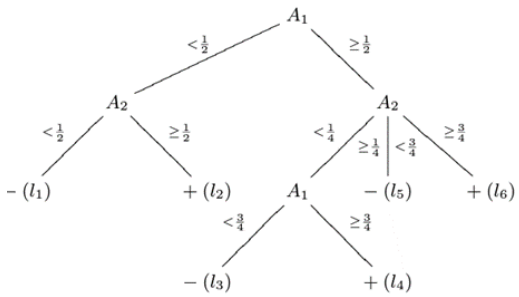
Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, *1*(2), 1-52.

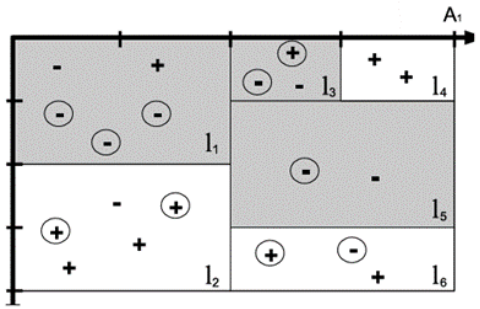
An overview of post-processing approaches to fairness-aware learning

- Different techniques for white-/black-box models:
  - White-box:
    - Correct the class labels of decision trees (Kamiran et al., 2010)
    - Correct the confidence scores of classification rules (Pedreschi et al, 2009)
    - Correct the probabilities in Naive Bayes classifiers (Calders & Verwer, 2010)
    - …
  - Black-box:
    - Change the decision boundary (Kamiran et al, 2018), (Hardt et al, 2016) (Fish et al, 2016)
    - Wrap a fair classifier on top of a black-box learner (Agarwal et al, 2018)
    - ….

Mitigating bias: Post-processing approaches: Discrimination Aware Decision Tree Learning

- Correct the class labels/ Relabel of decision trees (Kamiran et al., 2010)
- Illustrative example: A decision tree and the corresponding partitioning
  - The ground truth labels are +, -.
  - Gray areas correspond to regions where the majority class is -
  - Encircled instances correspond to protected instances.
- Key questions:
  - Which leaves to choose for correction/relabeling?
  - How many leaves?





Kamiran, F., Calders, T., & Pechenizkiy, M. Discrimination aware decision tree learning. In ICDM 2010.

Post-processing approaches: discussion

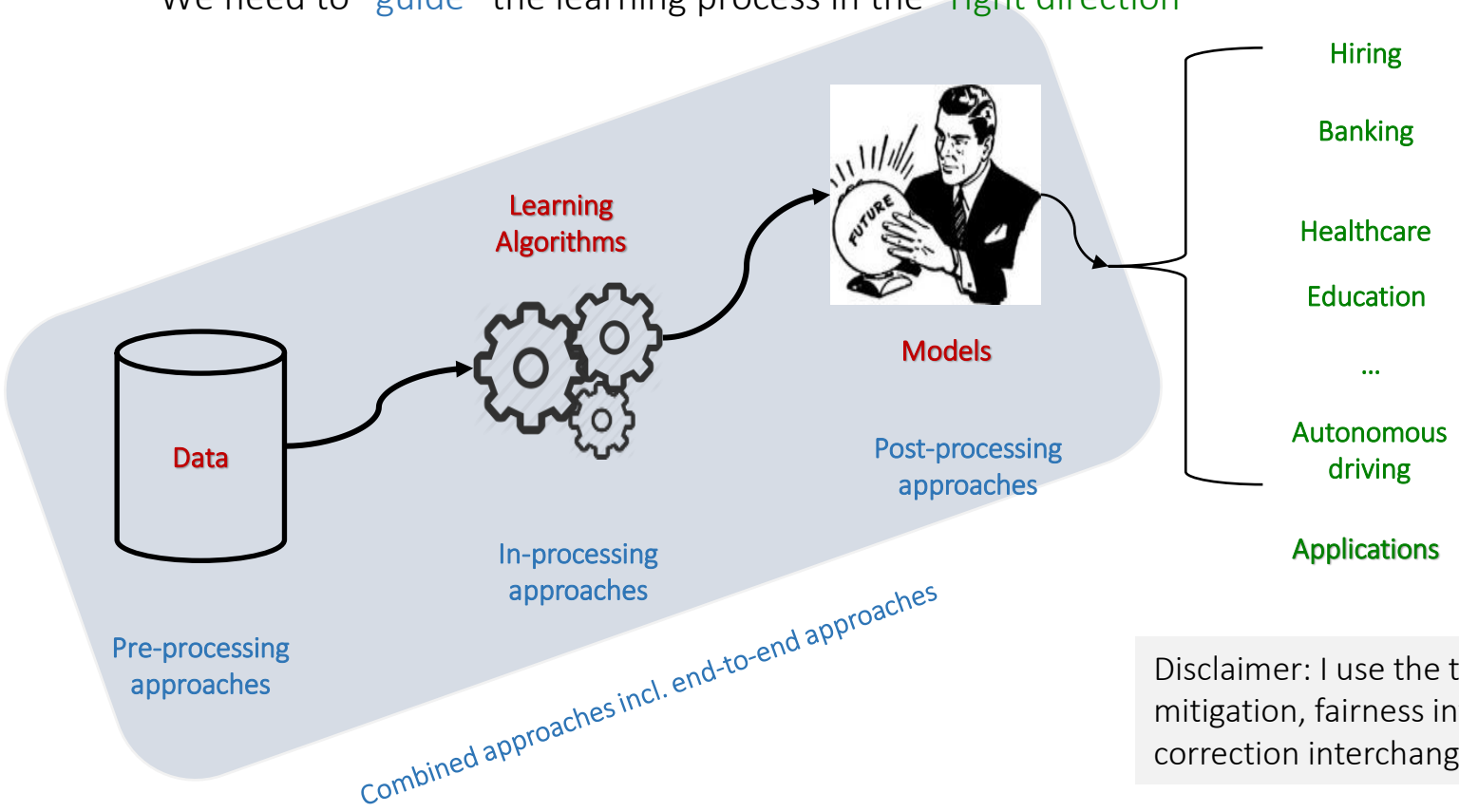
- Less popular approach to bias mitigation (based on number of publications
- Prioritizes predictive performance, fairness is secondary
- Most methods are model-specific
  - New models require new methods or adaptation
- However, it is useful in practice because we often have access only to the model's outcome, not its training process.

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined-approaches approaches
- Reflection on mitigation methods
- Scaling up complexity

How to mitigate bias and discrimination

- Back to basics: We need to understand how machines learn and how things can go wrong
- We need to "guide" the learning process in the "right direction"



Disclaimer: I use the terms bias mitigation, fairness interventions, bias correction interchangeably

Combined approaches

- Pre-processing approaches focus solely on the data
- In-processing approaches focus solely on the algorithm
- Post-processing approaches focus solely on the model

- Combined approaches aim to address discrimination in a more holistic manner by considering some combination of {data, algorithm, model}-interventions, rather than targeting a single component

FairNN - Conjoint Learning of Fair Representations for Fair Decisions

- Tackle bias and discrimination jointly in i) feature representation learning and ii) classification task (Hu et al, 2020)
- FairNN consists of two components
  - Representation learning component: An autoencoder
  - Classification component: A neural network
- Fairness is "injected" in both components
  - At the autoencoder: to obfuscate information related to the protected attribute
  - At the classifier: to force the algorithm to consider fairness as well

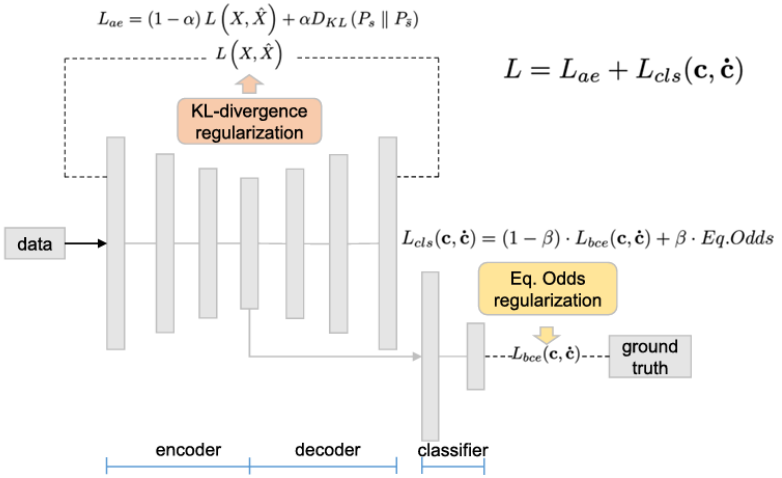


Fig. 1. An overview of *FairNN* that jointly learns a fair representation and a fair mapping function for classification. The auto-encoder (left part) is responsible for representation learning; the KL-divergence constraint forces the representation to be fair. The loss function of the classifier (right part) is tweaked towards fairness through the Eq.Odds regularization. Both aspects are reflected in the joint objective

Hu, T., Iosifidis, V., Liao, W., Zhang, H., Yang, M. Y., Ntoutsi, E., & Rosenhahn, B. FairNN-conjoint learning of fair representations for fair decisions. In *Discovery Science 2020*

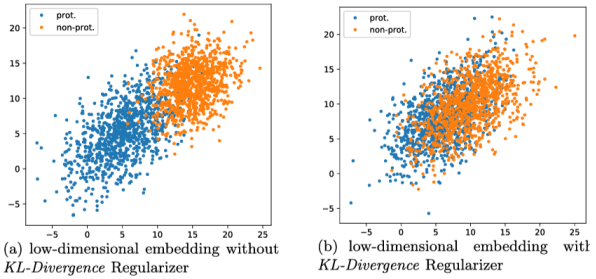
FairNN - Conjoint Learning of Fair Representations for Fair Decisions

- **Step 1:** Fair Representation Learning via KL-Divergence Regularization
  - Intuition: Learn representations that do not depend on the protected attribute. In other words, obfuscate the information about the protected attribute in the latent space
  - Idea: Regularize the auto-encoder loss through KL-divergence

- The updated loss function of the auto-encoder is

$$L_{ae} = (1-\alpha)\, L\left(X, \hat{X}\right) + \alpha D_{KL}\left(P_s \parallel P_{\bar{s}}\right)$$

  - L(): the typical autoencoder reconstruction loss
  - α: a balancing parameter



(a) low-dimensional embedding without *KL-Divergence* Regularizer

(b) low-dimensional embedding with *KL-Divergence* Regularizer

Fig. 2. Effect of the *KL-Divergence* Regularizer in (fair) representation learning

The protected attribute (gender) information is mixed up in the latent space.

FairNN - Conjoint Learning of Fair Representations for Fair Decisions

- **Step 2:** Fair Classifier Learning via Equalized Odds Regularization
  - Traditional class-based loss is extended to also consider the fairness of the model (Equalized Odds)

$$L_{cls}(\mathbf{c}, \dot{\mathbf{c}}) = (1-\beta) \cdot L_{bce}(\mathbf{c}, \dot{\mathbf{c}}) + \beta \cdot Eq.Odds$$

    - $L_{bce}()$: the binary cross entropy loss
    - β: accuracy-fairness trade-off
- **Joint optimization**

$$L = L_{ae} + L_{cls}(\mathbf{c}, \dot{\mathbf{c}})$$

Fair-representation learning Fair classification

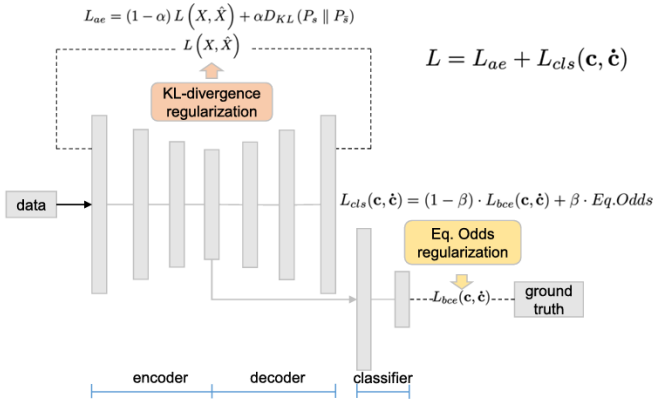


Fig. 1. An overview of *FairNN* that jointly learns a fair representation and a fair mapping function for classification. The auto-encoder (left part) is responsible for representation learning; the KL-divergence constraint forces the representation to be fair. The loss function of the classifier (right part) is tweaked towards fairness through the Eq.Odds regularization. Both aspects are reflected in the joint objective

Combined approaches: discussion

- A more holistic approach targeting bias at various stages of the learning process
- Typically, multi-component methods implementing diverse strategies at various stages of the learning process.
- Understanding the effects of the different components and their interactions is very important (e.g., through ablation studies)

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined-approaches approaches
- Reflection on mitigation methods
- Scaling up complexity

Reflecting on bias mitigation methods 1/2

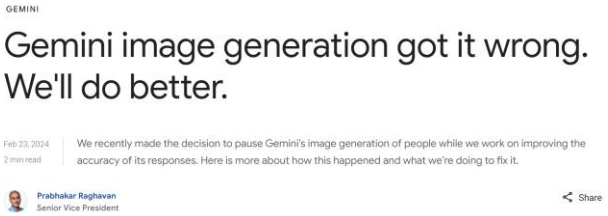
- Interventions for bias are possible through data, algorithms, models, end-to-end solutions but require
  - Some notion of fairness
  - Dealing with trade-offs
  - Careful evaluation and understanding of the effects of the interventions
  - …

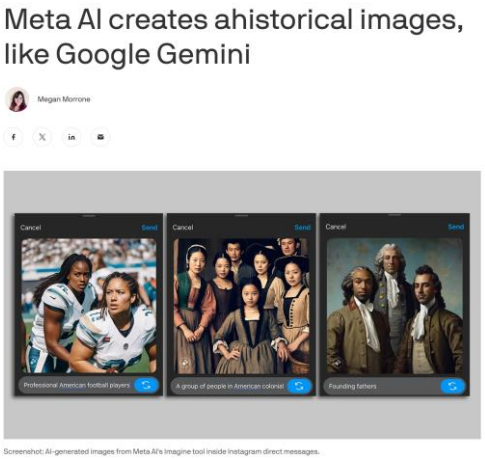
Reflecting on bias mitigation methods 2/2

- More fundamental work is needed on
  - Where and how should we intervene?
  - Are there trade-offs?
  - What other data challenges exist (e.g., imbalances, data scarcity)
  - What are the long-term effects of these interventions? Is traditional train-test evaluation the optimal approach?
  - Establishing benchmarks
    - Datasets (e.g. retiring Adult → ACS PUMS, synthetic generators, …)
    - Evaluation measures
    - Methods (various libraries exist: FairLearn (Microsoft), AI Fairness 360 (IBM), Themis-ml (), FairBench. (Mammoth project)
  - ….

It is not only about the technical means

- In an effort to enhance fairness by diversifying representation, Gemini was producing ahistorical images.





Source: https://www.wired.com/story/google-gemini-woke-ai-image-generation/

Outline

- Short recap – Setting the scene
- Mitigating discrimination
  - Pre-processing approaches
  - In-processing approaches
  - Post-processing approaches
  - Combined-approaches approaches
- Reflection on mitigation methods
- Scaling up complexity

AI fairness - are we looking at real world-complexity? Protected attributes

- Despite the growing body of work on fairness
  - Most of the methods focus on single-dimensional discrimination/fairness.
  - But human identifies are multi-dimensional and discrimination can occur on a basis of more than one protected attribute
- Multi-discrimination, is an old concept in e.g., law*
  - Cumulative or additive discrimination
  - Intersectional discrimination
  - Sequential discrimination
- What new (learning) challenges arise?
  - Fairness gerrymandering**
  - Population imbalances
  - Severe class imbalances for the minority groups
  - Lack of benchmark datasets
  - …

Image source

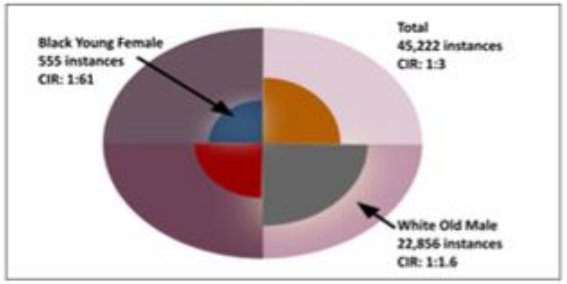


Image source

*Roy, A., Horstmann, J., & Ntoutsi, E. (2023, June). Multi-dimensional discrimination in law and machine learning-A comparative overview. *2023 ACM FAccT*.
Kearns, M., Neel, S., Roth, A., & Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In PMLR **2018

AI fairness - are we looking at real world-complexity? Is fairness the only objective?

- Responsible AI encompasses various factors of responsibility including fairness, explainability, security, privacy, …
  - Most of the existing works address these aspects in isolation
  - Although there exist studies which show, for example, that fairness enhancing interventions have various effects on the security of the subgroups*
  - Or that, marginalized (sub)groups may face higher risks of discrimination and privacy violations due to bias and (under)representation**
- Joint tackling is necessary
- & considering the consequences of fairness enhancement methods on other important aspects (Evaluation!)

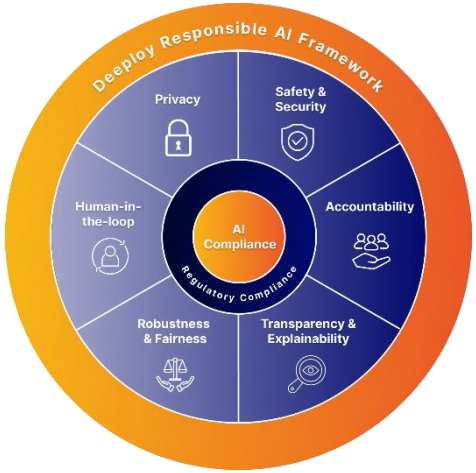
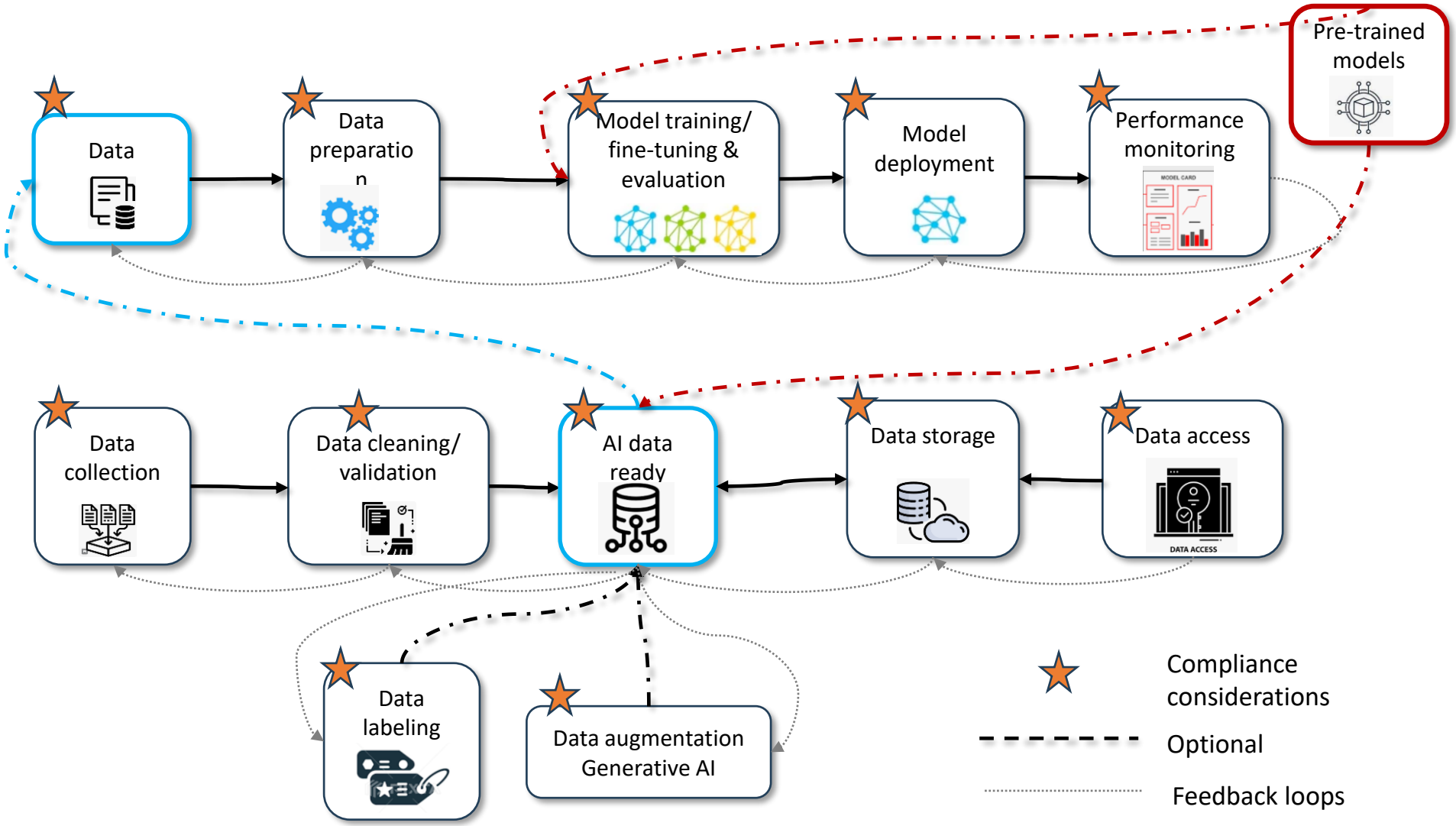


Image source

*Yulu Jin and Lifeng Lai. Privacy protection in learning fair representations. In IEEE ICASSP 2022

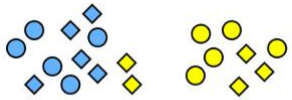
**Junjie Zhu, Lin Gu, Xiaoxiao Wu, Zheng Li, Tatsuya Harada, and Yingying Zhu. People taking photos that faces never share: privacy protection and fairness enhancement from camera to user. In AAAI 2023

AI fairness - are we looking at real world-complexity? Complex interconnected data-and model-life cycles



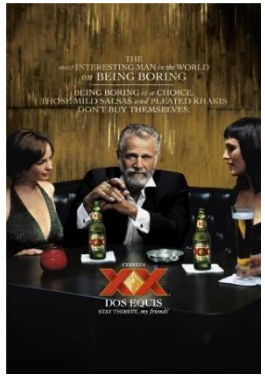
AI fairness - are we looking at real world-complexity? Data modalities

- Despite the growing body of work on fairness
  - Most of the existing approaches focus on tabular data
  - Less work on images, text, graph, timeseries, multi-modal data
    - & many of the existing methods
      - transform the data into tabular data and leverage techniques from the tabular domain* which may oversimplify the inherent complexities of the original data.
      - employ fairness measures that might not adequately account for the complexities of different data types
        - e.g., using balance score for group-fairness in graphs – introduced in the context of i.i.d. clustering** (looking at the minority/majority ratio in each cluster) – will not capture node dependencies



Group-level Fair Clusters

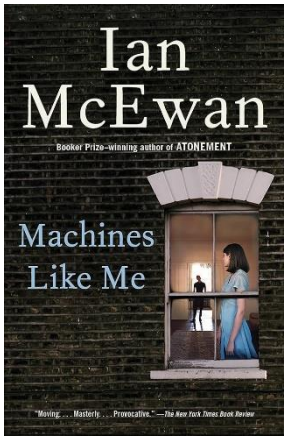
Source

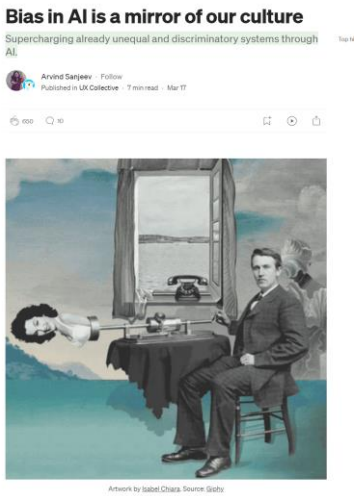


Source

*Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2022). A survey on bias in visual datasets. *Computer Vision and Image Understanding*, *223*, 103552.
**Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. In: NeurIPS. pp. 5029–5037 (2017)
Swati Swati, Arjun Roy, Eirini Ntoutsi, Exploring Fusion Techniques in Multimodal AI-Based Recruitment: Insights from FairCVdb. EWAF'24.

Final thoughts





- Do we want AI to mirror reality?
  - Should AI reflect the world as it is, with all its biases?
- Do we want AI to please our illusions/views?
  - Should AI be designed to conform to our ideals and perceptions, even if they don't align with reality?
  - Who's ideals and perceptions should AI follow?
- Would hard-coded ethics work?
  - Can predefined ethical rules effectively guide AI behavior, or is flexibility required?
- Can we crowdsource ethics? Via LLLs?
  - Can we gather ethical guidelines collectively? Traditionally through crowdsourcing, and recently through LLMs

(Credit: @EndWokeness)

Thank you for your attention!

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

MAMMOth

NoBIAS

LernMINT

BIAS

STELAR

DFG Deutsche Forschungsgemeinschaft

European Commission

SFB Offshore–Megastrukturen

Bundesministerium für Wirtschaft und Klimaschutz

VolkswagenStiftung

Alexander von HUMBOLDT STIFTUNG

Niedersächsisches Ministerium für Wissenschaft und Kultur