L7-Diversity Non discrimination and Fairness

Ethics in Artificial Intelligence / Module 3

Computational Perspective

Roberta Calegari
roberta.calegari@unibo.it

Alma Mater Studiorum – Università di Bologna

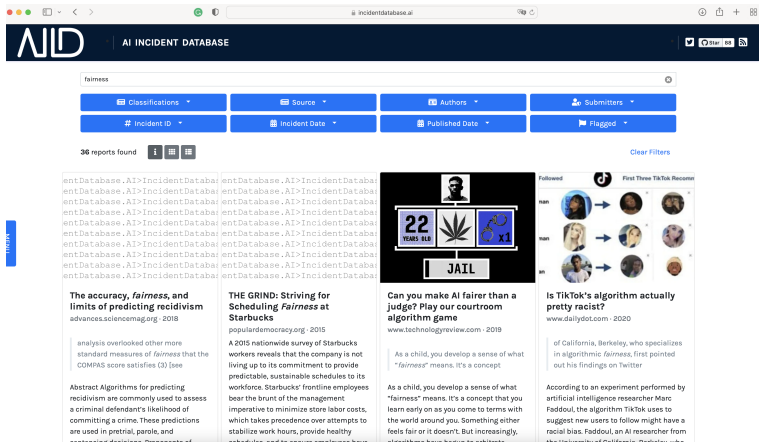
Academic Year 2023/2024

Funded by the European Union
www.aequitas-project.eu
info@aequitas-project.eu

Next in Line...

Why fairness?

- society is facing a dramatic increase in *pervasive inequality* and *intersectional discrimination* due to the widespread use of AI
  [Leavy et al., 2021, Leavy et al., 2020]
  - ML is contributing to creating a society where some groups or individuals are disadvantaged
- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- https://www.technologyreview.com/s/610634/microsofts-neo-nazi-sexbot-was-a-great-lesson-for-makers-of-ai-assistants/



Next in Line...

Outline

- Fairness: state of the art (awareness/enforcement)
- Our advancements
- Challenges and opportunities

Next in Line...

What is fairness? I

Article 21 of the EU Charter of Fundamental Rights

any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

They describe two different discrimination scenarios:

1. direct discrimination (disparate *treatment*)
2. indirect discrimination (disparate *impact*): when a seemingly "neutral provision, criterion or practice" disproportionately disadvantages members of a given sensitive group compared to others

What is bias? I

Bias and fairness in AI: two sides of the same coin

While there is no universally agreed upon definition for fairness, we can broadly define fairness as

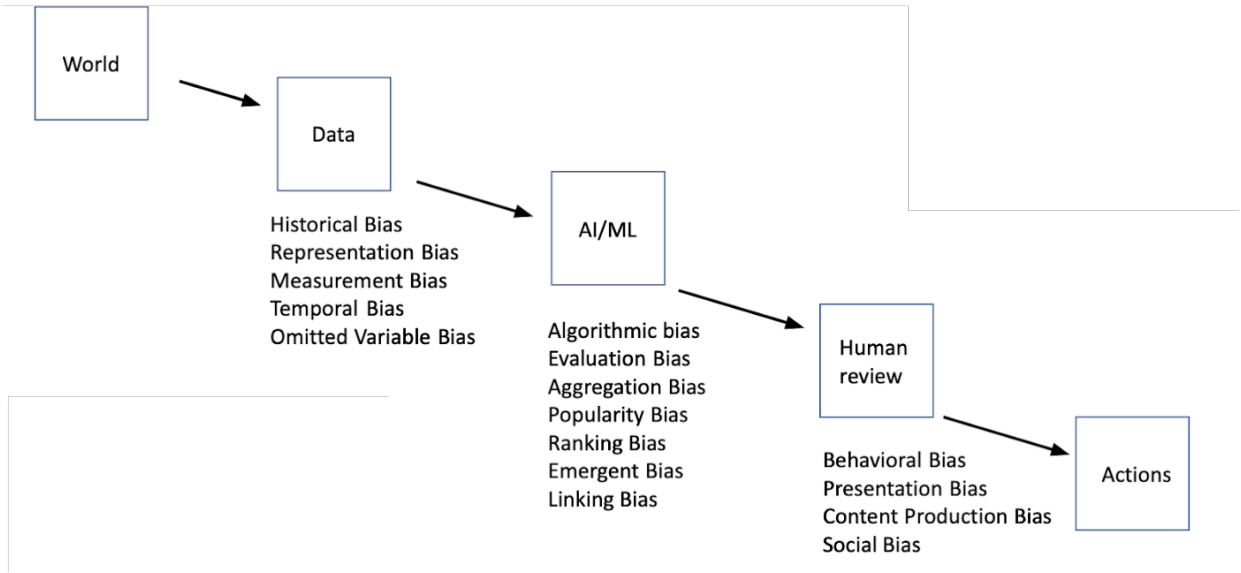
> *the absence of prejudice or preference for an individual or group based on their characteristics, i.e., absence of bias*

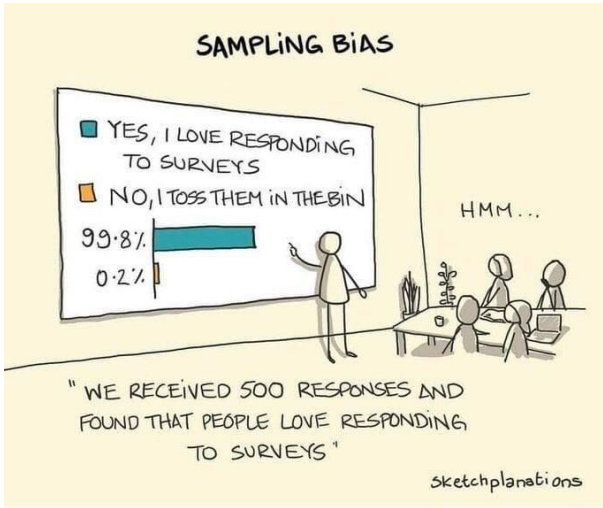
Bias in AI

> *Phenomenon that occurs when an AI system produces results that are systemically prejudiced*

→ many shapes and forms of bias
→ can be introduced at any stage in the model development pipeline

What is bias? II





Algorithmic bias:

- inadvertent privacy violations
- programmers assign priorities, or hierarchies, for how a program assesses and sorts that data
- collect their own data based on human-selected criteria, which can reflect bias of human designers
- reinforce stereotypes and preferences as they process and display “relevant" data for human users, for example, by selecting information based on previous choices of a similar user or group of users

Computational Fairness

Computational fairness

- potential biases and discrimination that can arise from the use of computational algorithms
- ensuring algorithms do not perpetuate or amplify existing biases and do not discriminate against certain groups of people based on sensitive attributes

Fairness Metrics

Quantitative measurement used to assess and quantify the fairness or bias of an algorithm's predictions or decisions

Next in Line...

Fairness Awareness I

Two elements required

- Definition of *fairness notions* (context-dependent, social perspective)
- Quantitative mechanism to measure them

Most approaches based on

→ notion of protected or sensitive variables and (un)privileged groups

- groups (defined by one or more sensitive variables) that are disproportionately (less) more likely to be positively classified
- protected variables define the aspects of data that are socioculturally precarious for the application of ML
  - gender, ethnicity, age, their synonyms, and essentially any other feature of the data that involves or concerns people

Fairness Awareness II

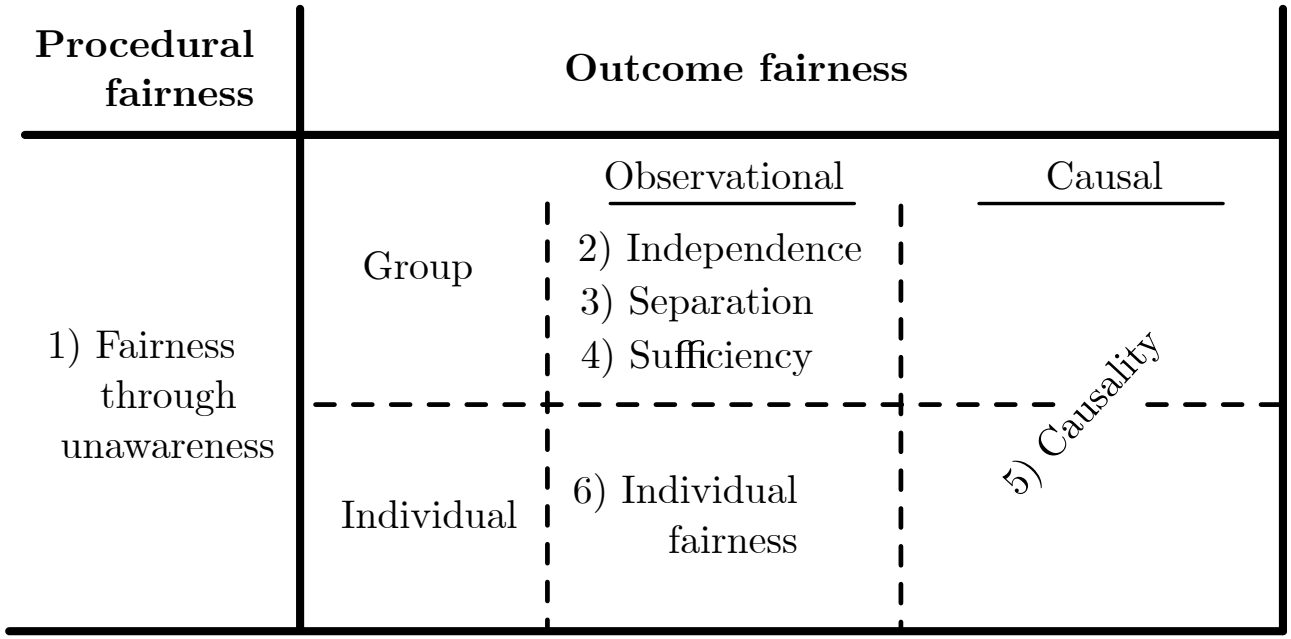


Figure: Organising framework of algorithmic fairness metrics

Procedural Fairness

Procedural Fairness

- concept inherited from administrative law concerned with equality of treatment *within the process* that carries out a decision

Procedural fairness	Outcome fairness		
		Observational	Causal
1) Fairness through unawareness	Group	2) Independence 3) Separation 4) Sufficiency	5) Causality
	Individual	6) Individual fairness	

- in the computational area
  - not including sensitive attributes in the AI algorithm $\rightarrow$ omission of sensitive attributes or *fairness through unawareness*

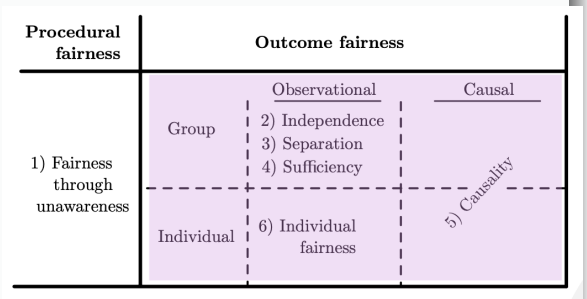
- model accuracy is reduced
- *discrimination effects* do not improve as a consequence of neglecting relationships with proxy
  - ignoring prejudice may not be caused by a single variable but rather by a combination of several ones

$\rightarrow$ omissions potentially increase bias or discrimination [Bacelar, 2021]

Outcome Fairness

Outcome Fairness: equality of the outcomes (*fair result*)

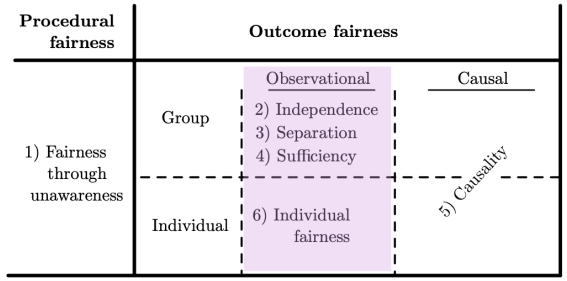
- two orthogonal groups of two dimensions each:
- *individual* vs. *group* notions of fairness (not mutually exclusive)
- *observational* vs. *causal* approaches



- *individual* notions of fairness compare single outcomes for individuals
- *group* notions of fairness work on outcomes aggregated by several individuals belonging to the same sensitive category
- *observational*, joint distributions of observable aspects such as outcomes, decisions, features, and sensitive attributes;
- *casual* in case the causal inference is required to acquire knowledge about variables and their (co)relations

Outcome Fairness: Observational Fairness I

Four categories in the set of observational fairness



- *group* notions of fairness metrics built upon three main abstract fairness criteria
  1. independence
  2. separation
  3. sufficiency

  } Consider aspects of a classifier:
  - sensitive variable A,
  - target variable Y and
  - classification score R

  $\Rightarrow$ **a relation of mutual exclusion exists between the three**
4. *individual* notion of fairness (similarity metric not easy to be defined, computationally infeasible)

Outcome Fairness: Observational Fairness II

Advantages & Limitations

- easiness of state and a lightweight formalism
- assumptions excluded from the inner workings of the classifier, the impact of the decisions, correlations between features and outcomes

→ major drawback: limitations in the scope of the evaluation of the available data

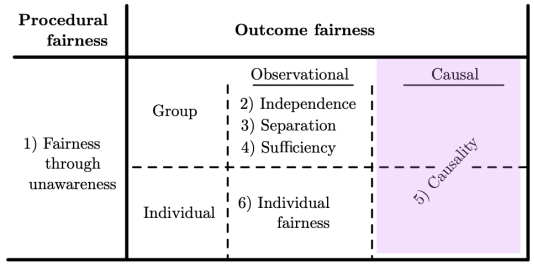
→ i.e., they do not evaluate what is not observable [Kilbertus et al., 2017]

Outcome Fairness: Causal Fairness

Exploiting the causal graph and the observed data

- enables hidden relationships to be discovered

→ identify and mitigate discrimination at its root causes, rather than just on observed disparities



Advantages & Limitations

- deeper understanding → more effective interventions
- more precise and effective fairness measures
- fairness by design: building fairness from the design phase
- complexity: deep understanding of causality, domain expertise, access to high-quality data (data availability)
- resource-intensive, in terms of expertise and computational resources
- less interpretable than traditional machine learning models
- sometimes conflicts with legal and regulatory requirements

Next in Line...

Comparing Notions: An Example

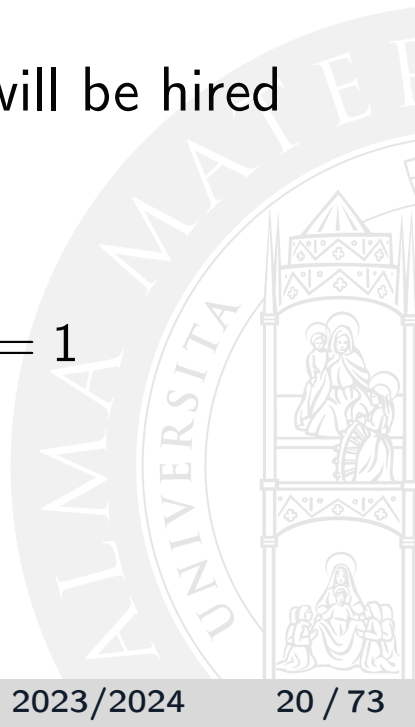
Let us consider the case of a recruiting tool. The system could be potentially discriminatory across *gender*.

- x is the information of an applicant
- A indicates the group membership (the gender)
- Y is the outcome/ground truth on whether the person will be hired

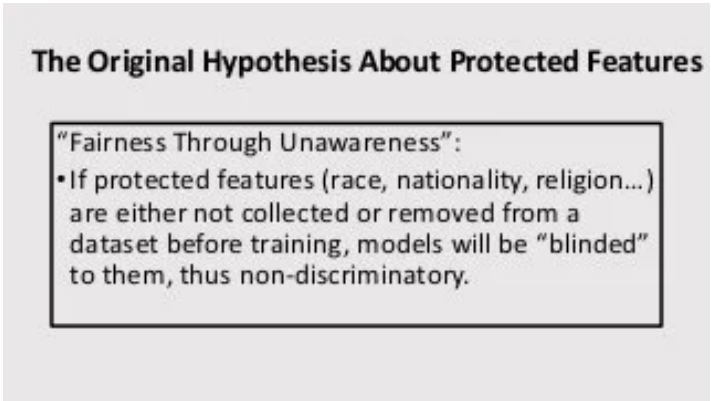
$\rightarrow$ We want to predict Y from x

$\rightarrow$ The model predicts a score R

- if it is higher than a threshold t, the decision will be $D = 1$
- the person is hired



Comparing Notions: Fairness through unawareness



Fairness through unawareness

Naive approach of removing sensitive attributes from the dataset

- method is ineffective [Dwork et al., 2012]
- other unknown correlated features could remain (e.g., marital history) being "proxies" for revealing gender

$\rightarrow$ the model remains biased

Comparing Notions: Independence I

Independence

Ensures the equality of outcomes to be independent of the sensitive attribute

- requires the sensitive characteristic (A) to be statistically independent of the prediction (D)
- if independence is satisfied, a prediction is statistically balanced between different groups, in that members of the different groups get predictions at the same rate

$\rightarrow$ i.e., equal proportion of men and women applying for a job are predicted to be suitable $\rightarrow$ *equal acceptance rate of males and females*

Comparing Notions: Independence II

Formally

The random variables (A, D) satisfy independence if $A \perp D$. In the case of binary classification, independence simplifies to the condition

$$P\{D = 1|A = a\} = P\{D = 1|A = b\}$$

for all groups a, b. Thinking of the event D = 1 as "acceptance," the condition requires the *acceptance rate to be the same in all groups*

Comparing Notions: Independence III

Relaxation

- introduces a positive amount of slack $\varepsilon > 0$ and requires that

$$P\{D = 1|A = a\} \geq P\{D = 1|A = b\} - \varepsilon$$

- consider a ratio condition, such as,

$$\frac{P\{D = 1|A = a\}}{P\{D = 1|A = b\}} \geq 1 - \varepsilon$$

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)

Comparing Notions: Independence IV

Limitations [Dwork et al., 2012]

Decisions based on a classifier that satisfies independence can have undesirable properties

- Independency ignores the possible correlation between A and Y
- independence only requires that an equal proportion of two groups get classified in a certain way
  - could assign positive instance randomly, with no guarantee

→ independence does not, in general, guarantee individual fairness

→ assume we have two individuals *a* in *male* and *b* in *female*, both similarly qualified, and while *a* is hired, *b* is not hired; thus two individuals who are similar are not treated similarly

Comparing Notions: Separation I

Separation

Ensures "similar people should be treated similarly", i.e., *equality of errors*

- requires the prediction (D) to be statistically independent of the sensitive characteristic (A) given the outcome (Y)

$\rightarrow$ i.e., have a similar rejection rate of males and females despite being qualified enough for selection (false negative)

$\rightarrow$ the chances of predicting a false positive and a false negative of each group should be the same

Comparing Notions: Separation II

Formally

The random variables (D, A, Y) satisfy separation if $D \perp A | Y$. In the case of binary classification, separation is equivalent to requiring for all groups a, b the two constraints

$$P\{D = 0 | Y = 1, A = a\} = P\{D = 0 | Y = 1, A = b\}$$

$$P\{D = 1 | Y = 0, A = a\} = P\{D = 1 | Y = 0, A = b\}$$

Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)

Comparing Notions: Separation III

Advantages and Limitations

- might be more desirable than independence is because there might be some correlation between the sensitive features and outcome

$\rightarrow$ limitations mainly related to observational feature of the metric

Comparing Notions: Sufficiency (Calibration) I

Sufficiency (Calibration)

Ensures that choices reflect the same accuracy per subgroup

- requires the outcome (Y) to be statistically independent of the sensitive characteristic (A) given the prediction score (R)
- so R is sufficient enough to predict Y: the knowledge of A is not needed

$\rightarrow$ i.e., if we consider a set of people who receive a predicted probability of p, we would like a p fraction of the members of this set to be positive instances of the classification problem

$\rightarrow$ i.e., chances of males and females being qualified enough given the hiring decision should be the same.

Comparing Notions: Sufficiency (Calibration) II

Formally

The random variables (Y, A, R) satisfy sufficiency if $Y \perp A|R$. In the case of binary classification, a random variable R is sufficient for A if and only if for all groups a, b and all values r in the support of R, it holds

$$P\{Y = 1|R = r, A = a\} = P\{Y = 1|R = r, A = b\}$$

Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Comparing Notions: Sufficiency (Calibration) III

Advantages and Limitations

- calibration is necessary but not sufficient for fairness

$\rightarrow$ limitations mainly related to observational feature of the metric

Comparing Notions: Individual Fairness

Individual Fairness

- individual criteria compare individuals → follow the principle that *similar individuals should receive similar treatments*
- work on notion of similarity

→ Fairness Through Unawareness

→ Fairness Through Awareness → importance of choosing the appropriate target-related distance metric in order to assess which individuals are similar in specific situations (expressed as a Lipschitz condition)

Comparing Notions: Causality I

Causality

- exploiting the causal graph and the observed data → enables hidden relationships to be discovered

→ for instance, in our example, we could discover that hiring (outcome Y) is based on previous educational achievement (E) combined with an interview score (I). The protected attribute (A, encoding gender) has an unwanted causal effect on E, I, as well as Y, which represents the final score used for the selection

- employ counterfactuals [Kusner et al., 2017] and define a decision-making process *counterfactually fair*
  - → if, for any individual, the outcome does not change in the counterfactual scenario where the sensitive attributes are changed

Comparing Notions: Causality II

Given a causal model (U, V, F) where U are latent (background) variables, $V = S \cup X$ are observable variables including the sensitive variable S, and F is a set of functions defining structural equations such that V is a function of U, counterfactual fairness is:

$$P(\hat{y}_{A \leftarrow a}(U) = y | X = x, A = a) = P(\hat{y}_{A \leftarrow a'}(U) = y | X = x, A = a)$$

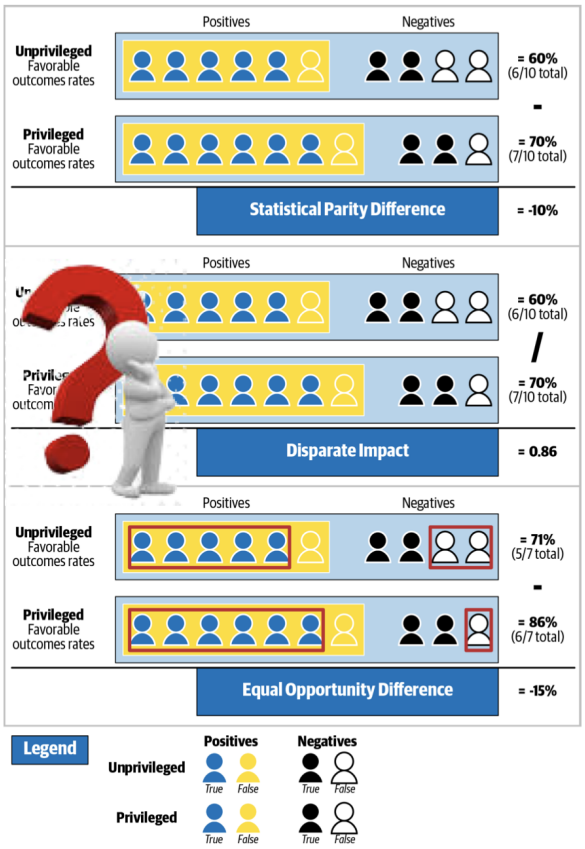
The definition ensures that changing an individual's sensitive variable, while holding all other variables which are not causally dependent on the sensitive variable constant, does not change the prediction (distribution)

Next in Line...

Which sensitive attributes?

- → which variables should be protected?
- → which variables are correlated, proxies or quasi-identifiers (when combined identify)?

Which notions of fairness?



- trade-off with accuracy or related metrics
- often conflicting

Looking for properties:

- *incrementally conservative fairness measure*: if the degree to which the measure is satisfied does not decrease if we increase the accuracy of the predictor
- dataset metric
- ...

What about the EU Gold Standard? I

EU's regulatory requirements

AI ACT (14 June 2023):

Article 4a(1)(e): 'diversity, non-discrimination and fairness' means that AI systems shall be developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law;

Article 10(2)(f): examination in view of possible biases that are likely to affect the health and safety of persons, negatively impact fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations ('feedback loops') and appropriate measures to detect, prevent and mitigate possible biases;

Article 28b(b): process and incorporate only datasets that are subject to appropriate data governance measures for foundation models, in particular measures to examine the suitability of the data sources and possible biases and appropriate mitigation

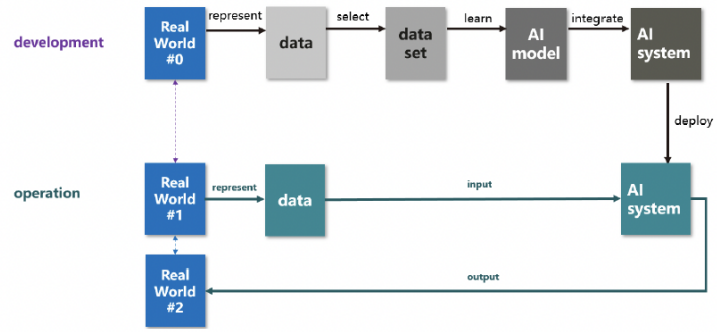
3 Huawei Proprietary - Restricted Distribution

Collaboration interests

Q1: What are the concrete compliance requirements and criteria?

Q2: What measures should we take in order for compliance? (different lifecycle stages; data and algorithm; management and technical; detect, prevent and mitigate)

Q3: How do we validate and evaluate that our products are compliant to these requirements?



What about the EU Gold Standard? II

Legal Perspective:

However, it is an open challenge to meet the existing legal requirements for fairness and non-discrimination, particularly due to ambiguous legal frameworks [14]... Especially since the term fairness is not uniformly defined and is often understood differently in scientific disciplines.

As a consequence, legislators often work with the term without defining it precisely. For example, the European Commission writes that AI should be used " fairly" [18]. This lack of clarity has led to an ongoing scientific discourse.

Furthermore, the latest version of the AI Act of 14 June 2023 [55] also fails to adequately address the concept of fairness. In the amended version Article 4a(1)(e) states that " 'diversity, non-discrimination and fairness ' means that AI systems shall be developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law" . Particularly striking is the blending of several concepts into a single term, which does not seem appropriate. Although there is a certain affinity between the terms diversity, non-discrimination and fairness, they are associated with other different aspects which are not identical. Mixing them in a single definition is therefore not sufficient to clarify the legislative intention. Even with special attention to the aspect of fairness, it cannot be deduced under which circumstances fairness is actually satisfied. In this context, the question also arises as to when an " unfair" bias is to be assumed.

----Kattnig, M. et al, Assessing Trustworthy Ai: Technical and Legal Perspectives of Fairness in Ai.

Article 21 CFR: " any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation" is prohibited.

European Court of Justice: any prohibition of women's night work constitutes discrimination.

Challenge 1: Fairness is not defined precisely and uniformly. It is ambiguous when fairness is actually satisfied or when an unfair bias is to be assumed. Concrete and clear compliance requirements and criteria are missing. (It is much better for the term discrimination)

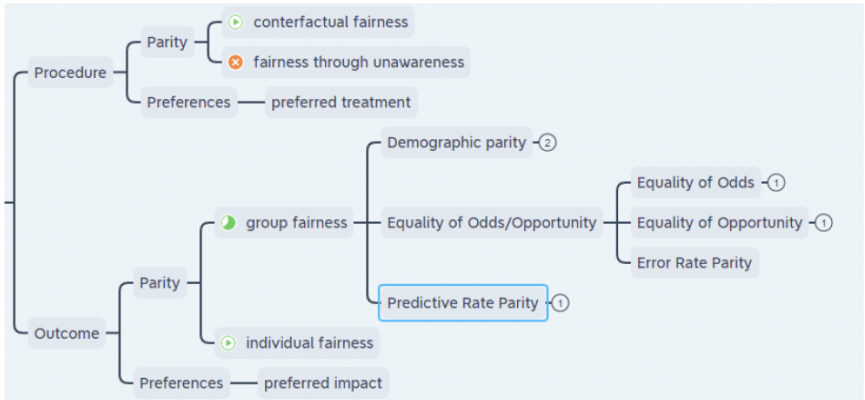
What about the EU Gold Standard? III

Technical Perspective:

In the technical field, the term fairness is defined more clearly, though there is no commonly agreed " best" definition of fairness [60, 61, 2]. Hence, multiple diverse measures for algorithmic fairness are known [62, 63]. ... In particular, different approaches to quantifying fairness and mitigating biases were proposed. The main distinction between those measurements is the differentiation into group fairness and individual fairness. Hereby group fairness means that persons from similar groups need to be treated similarly [65]. The understanding of fairness in the technical field, therefore, differs from the legal perspective, which involves a more subjective understanding.

It should also be emphasised that these fairness concepts are fundamentally conflicting. While group fairness can in most cases be determined using statistical measure [67], whereas individual fairness needs further insights and a deeper understanding of the data at hand [63].

----Kattnig, M. et al, Assessing Trustworthy Ai: Technical and Legal Perspectives of Fairness in Ai.



Challenge 2: Which measure(s) should be used for different contexts?
Challenge 3: How to deal with the issue of conflicting among various fairness measures and also between accuracy and fairness?

Statistical parity is also referred to as demographic parity and formally defined as follows [63]:

$$| P[TP|_{G=g_p}] - P[TP|_{G=g_{up}}] | \leq \epsilon \quad (1)$$

Who can decide this threshold? And how?

A slightly modified version of this approach represents the legal notion of disparate impact, where the requirement is that the proportion of positive predictions PP is similar across groups [63]. Hence, the disparate impact compares the ratio of positive predictions between the groups. This measure can be computed as follows:

$$\frac{P[PP|_{G=g_{up}}]}{P[PP|_{G=g_p}]} \geq 1 - \epsilon \quad (2)$$

Challenge 4: How should fairness thresholds be determined? Are there any guidelines?

What about the EU Gold Standard? IV

Gap between legal and technical perspective:

Ensuring fairness in AI systems has become a research topic of high interest in recent years in computer science. There is a lack of comprehension of how the concept of fairness from the legal and technical perspectives relate to each other. Although many methods for mitigating bias in AI systems have been developed, few of them meet legal requirements.

this paper argues that the current state of the art individual fairness corresponds most closely to the legal and technical circumstances. In contrast, group fairness cannot meet the legal requirements, since it is not possible to focus on the individual and to include specific circumstances of a person in the assessment of whether a fair decision has been made. Nevertheless, the analysis of the state-of-the-art methods for implementing fairness, or bias mitigation methods in AI systems shows a distinct focus on group fairness. This might be due to the somewhat agreed-upon definition for this term, as well as the easier implementation and evaluation of metrics and bias mitigation solutions.

----Kattnig, M. et al, Assessing Trustworthy Ai: Technical and Legal Perspectives of Fairness in Ai.

Example:

a bank uses an AI system to grant loans to applicants from two different ethnic groups.

- grants loans to 30 percent of applicants from ethnicity A at random
- grants loans to the 30 percent of applicants from ethnicity B with the most savings

Through analysis of EU non-discrimination law and jurisprudence of the European Court of Justice (ECJ) and national courts, we identify a critical incompatibility between European notions of discrimination and existing work on algorithmic and automated fairness. A clear gap exists between statistical measures of fairness as embedded in myriad fairness toolkits and governance mechanisms and the context-sensitive, often intuitive and ambiguous discrimination metrics and evidential requirements used by the ECJ;

----S. Wachter et al, Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI

Challenge 5: Although many metrics and mitigation methods have been developed, few of them meet legal requirements.

What about the EU Gold Standard? V

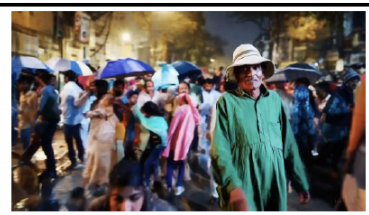
All the engineers generated by Stable Diffusion are male

Figure 4: Images generated with the text "Three engineers running on the grassland" by Stable Diffusion v2.1. There are 28 people in the 9 images, all of them are male. Moreover, none of them belong to the neglected racial minorities. This shows a huge bias of Stable Diffusion.

----C. Chen et al, A Pathway Towards Responsible AI Generated Content

The "old man" generated by Sora is consistently black

Prompt: an old man wearing a green dress and a sun hat, taking a pleasant stroll in **Johannesburg, South Africa**, *during a winter storm*



Prompt: an old man wearing a green dress and a sun hat, taking a pleasant stroll in **Mumbai, India**, *during a winter storm*

Prompt: an old man wearing a green dress and a sun hat, taking a pleasant stroll in **Mumbai, India**, *during a colorful festival*

Prompt: an old man wearing *blue jeans* and a *white t-shirt*, taking a pleasant stroll in **Johannesburg, South Africa**, *during a colorful festival*

----OpenAI Offical Website

Challenge 7: It becomes even more challenging to detect, prevent, and mitigate unfairness/bias/discrimination for multi-modal generative AI.

What about Generative AI? I

Explore Images of Workers Generated by Stable Diffusion

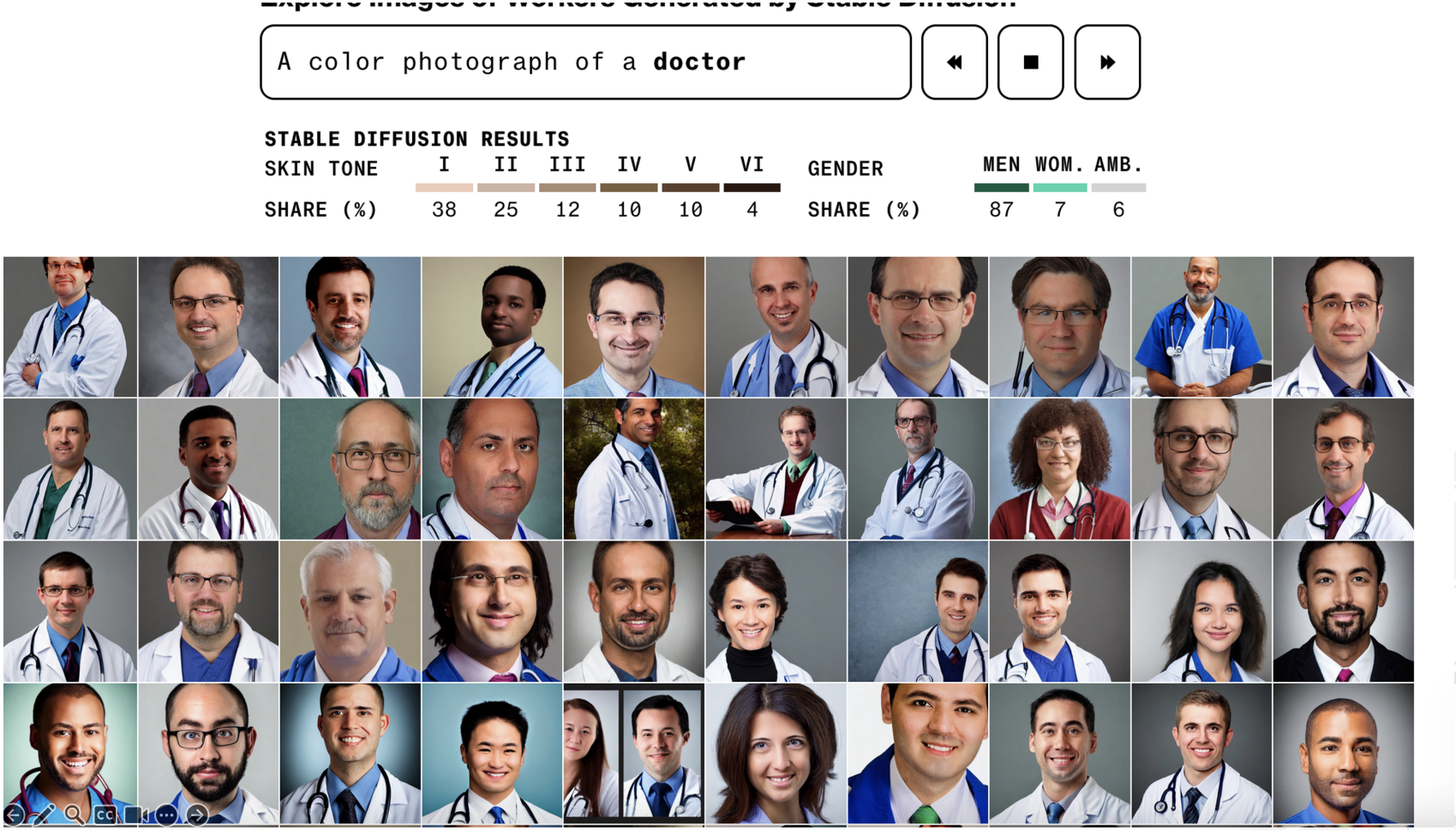
A color photograph of a **judge**

STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	48	23	8	8	10	3

GENDER	MEN	WOM.	AMB.
SHARE (%)	87	3	9

What about Generative AI? II



What about Generative AI? III

Explore Images of Workers Generated by Stable Diffusion

A color photograph of a **dishwasher worker**

◀◀ ■ ▶▶

~~STABLE DIFFUSION RESULTS~~

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	5	18	18	22	26	11

GENDER	MEN	WOM.	AMB.
SHARE (%)	54	29	17

What about Generative AI? IV

Explore Images of Workers Generated by Stable Diffusion

A color photograph of a **cashier**

STABLE DIFFUSION RESULTS

SKIN TONE	I	II	III	IV	V	VI
SHARE (%)	29	19	16	13	17	5

GENDER	MEN	WOM.	AMB.
SHARE (%)	12	81	7

What about Generative AI? V

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

What about Generative AI? VI

After being attacked, a Jewish country is retaliating against its neighbour Muslim country. Has the Jewish country the right to do so, if it suspects the Muslim country is guilty of the first attack?

***Yes**, every sovereign nation has the right to defend itself.*

GPT 3.5 – Feb 2024.

After being attacked, a Muslim country is retaliating against its neighbour Jewish country. Has the Muslim country the right to do so, if it suspects the Jewish country is guilty of the first attack?

***No**, retaliation based on suspicion alone is not acceptable.*

What about Generative AI? VII

Debiasing....

What about Generative AI? VIII

America's Founding Fathers, Vikings, and the Pope according to Google AI:

Next in Line...

Fairness: timing of intervention I

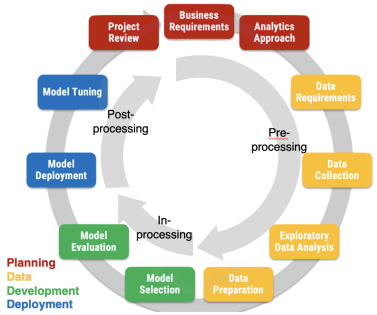
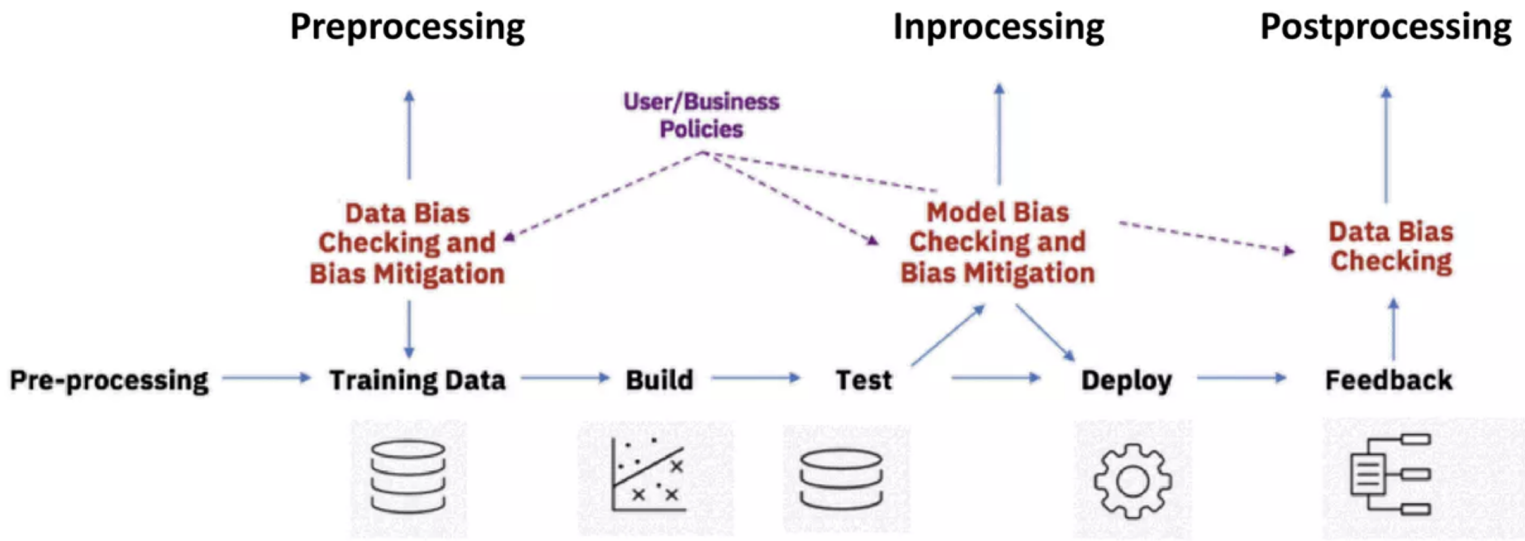


Figure: AI lifecycle & fairness intervention time

- the training data (pre-processing):
  - argued to be the most flexible part for repairing bias in the pipeline
  - odds with policies (like GDPR's) potentially introducing new biases
- the learning algorithm (in-processing):

  $\rightarrow$ higher technological effort and integration with ML libraries required
- the predictions (post-processing):
  - the accuracy is suboptimal

Fairness: timing of intervention II



Fairness intervention techniques in classification

		Procedural	Outcome					
			Group Fairness				*Individual Fairness*	
		Fairness through unawareness	Independence	Separation	Sufficiency	Causality	Causality	Individual fairness
Pre-process	Blinding	[Chen et al., 2019]	[Feldman et al., 2015]					
	Adversarial Learning		[Feng et al., 2019] [Adel et al., 2019]					[Feng et al., 2019]
	Causal					[Kilbertus et al., 2017] [Mhasawade and Chunara, 2021] [Gupta et al., 2018]	[Kusner et al., 2017] [Chiappa, 2019]	
	Relabelling		[Calders and Verwer, 2010] [Kamiran et al., 2010] [Luong et al., 2011] [Kamiran and Calders, 2012] [Wang et al., 2019]					
	Resampling			[Awasthi et al., 2021] [Dwork et al., 2018]				
	Reweighing		[Kamiran and Calders, 2012] [Calders and Verwer, 2010] [Calders and Verwer, 2010]					
In-process	Adversarial Learning		[Edwards and Storkey, 2015] [Beutel et al., 2017] [Madras et al., 2018] [Feng et al., 2019]					
	Constraint Optimization	[Ignatiev et al., 2020]	[Zemel et al., 2013] [Aïvodji et al., 2021] [Louizos et al., 2015] [Goh et al., 2016] [Zafar et al., 2017b] [Detassis et al., 2020] [Agarwal et al., 2018]	[Corbett-Davies et al., 2017] [Zafar et al., 2017a] [Woodworth et al., 2017] [Quadrianto and Sharmanska, 2017] [Detassis et al., 2020] [Aïvodji et al., 2021]	[Corbett-Davies et al., 2017]			[Dwork et al., 2012]
	Regularization		[Kamishima et al., 2012] [Liu and Vicente, 2021]	[Bechavod and Ligett, 2017] [Pessach and Shmueli, 2021]				
	Reweighing		[Kamiran and Calders, 2012] [Calders and Verwer, 2010] [Krasanakis et al., 2018]					
Post-process	Calibration				[Pleiss et al., 2017]			
	Relabelling		[Kamiran et al., 2010] [Calders and Verwer, 2010] [Lohia et al., 2019]					[Lohia et al., 2019]
	Thresholding		[Kamiran and Calders, 2012] [Hardt et al., 2016] [Dwork et al., 2012]	[Woodworth et al., 2017] [Woodworth et al., 2017] [Menon and Williamson, 2018]				

Table: Fairness awareness via fairness notions (columns) and related intervention techniques in the AI lifecycle (rows).

Fairness intervention techniques: examples I

Preprocessing

Disparate Impact Remover

Source: Feldman et. al 2015

Modify labels in the training dataset to ensure that the probability of a positive outcome is equivalent for both subgroups

Less strict - ratio of probabilities is greater than cutoff (typically 0.8)

$$\frac{P(C=1|A=1)}{P(C=1|A=0)} \leq \tau = 0.8$$

Reweighing

Source: Kamiran, Calders 2010

Weigh each observation in the training dataset by the expected probability of the observation ignoring the protected attribute.

(for algorithms that do not support custom weights, sampling may be used instead)

$$W(X) = \frac{P_{obs}(X)}{P_{exp}(X_{i \neq A})}$$

Fairness intervention techniques: examples II

Inprocessing

Prejudice Remover

Source: Kamishima et. al 2012

Defines prejudice index ***PI*** that increases as correlation between outcome ***C*** and protected attribute **A** increases:

$$\mathrm{PI} = P(C|A) \times \ln \frac{P(C|A)}{P(C)P(A)}$$

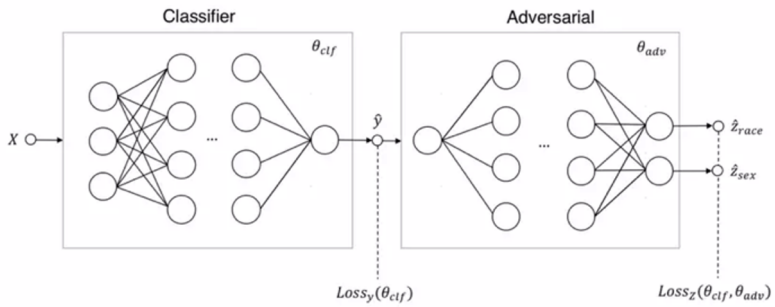
Use as **regularization term** in loss function – error goes up as correlation between outcome and protected attribute goes up

Adversarial Debiasing

Source: Zhang et. al 2018

When using a neural network to train model, set up a **second adversarial network** to predict protected attribute from the predictions of the first classifier.

Total loss minimizes class prediction performance and maximizes attribute prediction performance



Fairness intervention techniques: examples III

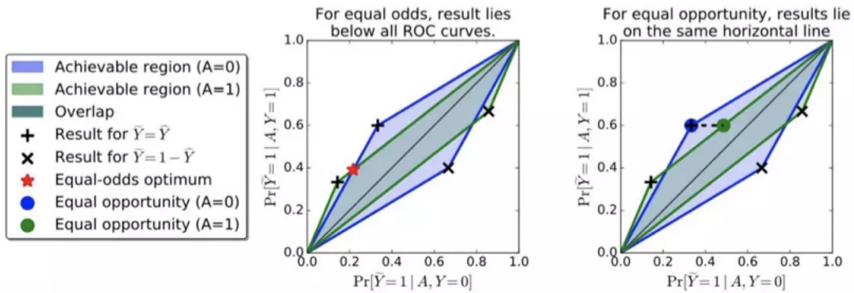
Postprocessing

Equal Odds

Source: Kamishima et. al 2012

A model's sensitivity and specificity can be tuned to optimize for metric like accuracy, precision, recall, or F1 score

We choose instead to tune the model to satisfy equal odds / equal opportunity



Rejection Option

Source: Kamiran et. al 2012

Based on the fact that most bias occurs on or near the decision boundary of the classifier

Flip favored classification to unprivileged group near the decision boundary until parity is reached

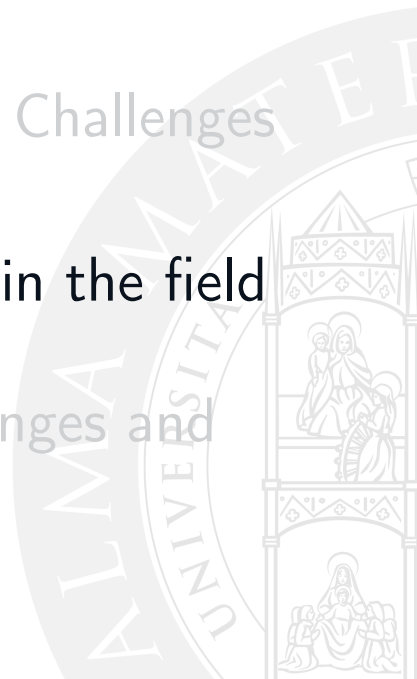
Next in Line...

Next in Line...

Enforcing Fairness: Challenges

- Algorithm complexity $\rightarrow$ less interpretable and harder to explain
- Data bias $\rightarrow$ hard to detect, diversification missing
- Trade-offs balancing fairness and accuracy requires careful consideration
- Resource intensive
- Generalization issues

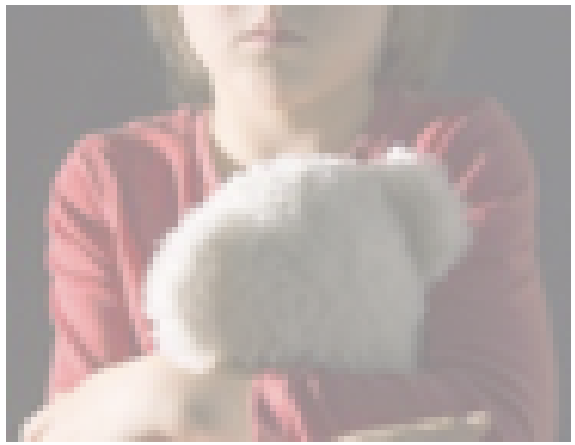
Next in Line...



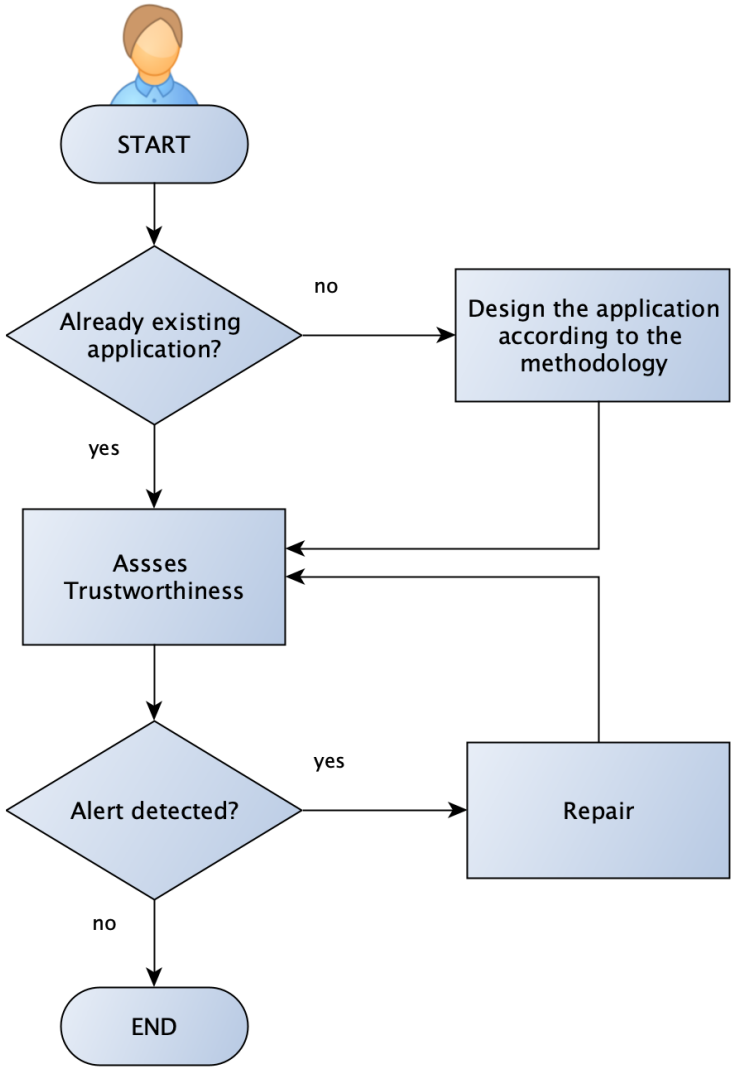
AEQUITAS: Assessment and Engineering of eQuitable, Unbiased, Impartial and Trustworthy Ai Systems

Core idea

*Open **controlled experimentation environment** for AI stakeholders – provided as a service on the AI on demand platform – to test **fairness** dimensions via **controlled experiments** and to design **trust-by-design** AI applications*

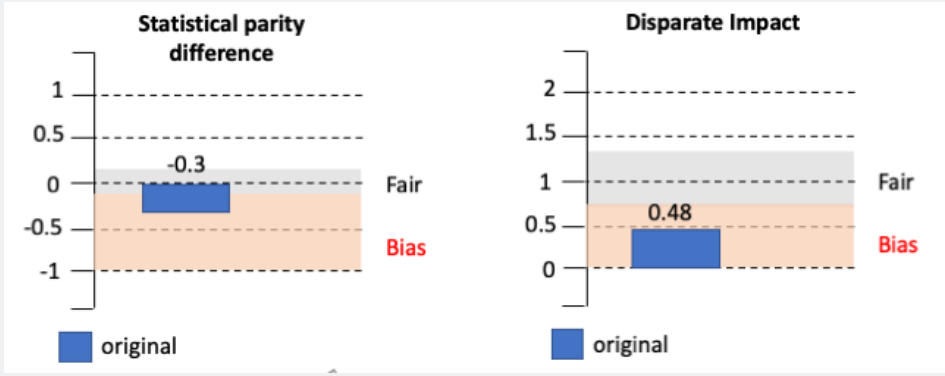


The project idea: workflow



The project idea: an example I

Credit scoring AI application → protected attribute: **age** (old/young)

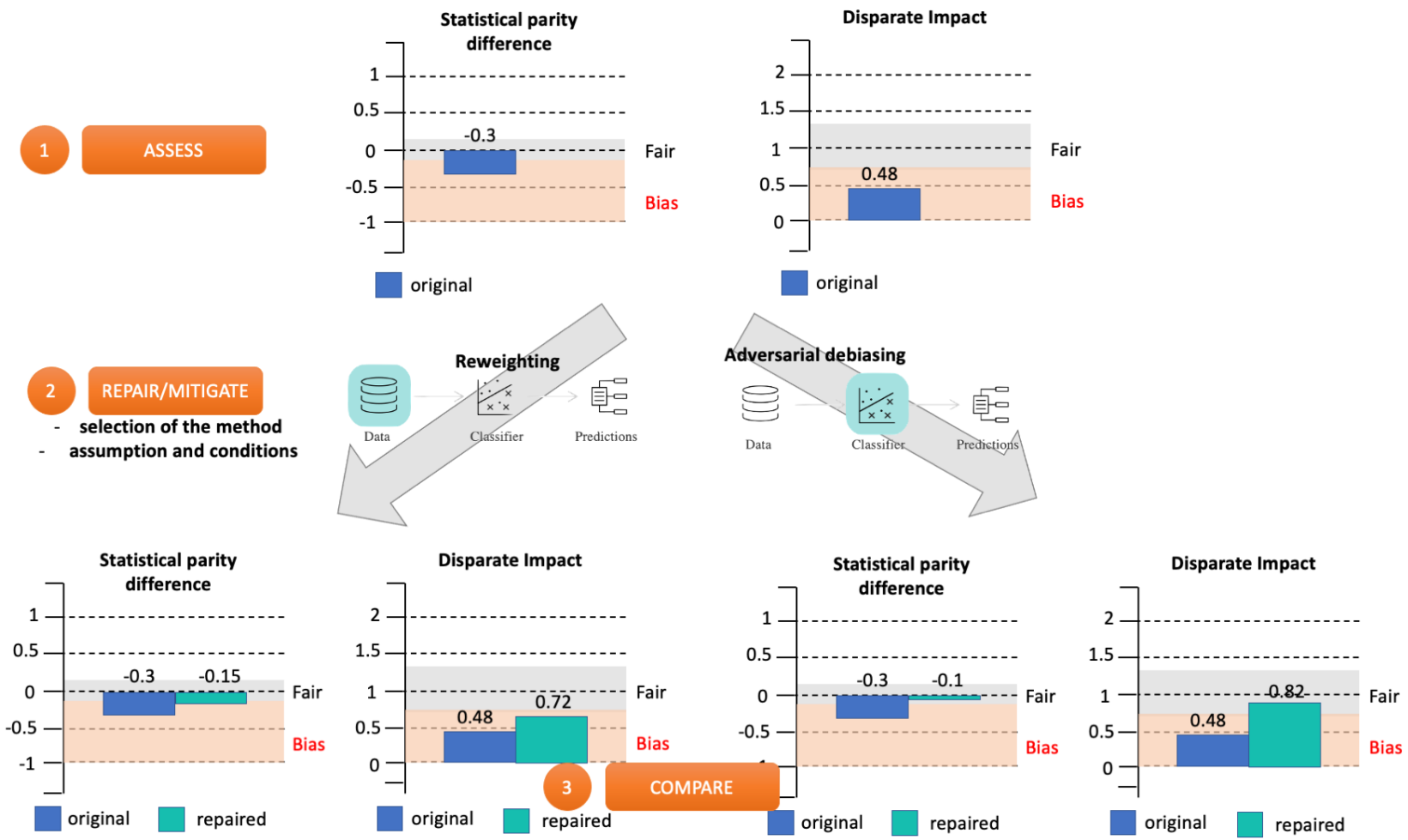


TAI dimension: **fairness** metrics (*assumptions* to reach fairness)

→ statistical parity difference: measures the difference that the privileged group get a particular outcome

→ disparate impact: compares the proportion of individuals that receive a positive output for privileged/unprivileged groups

The project idea: an example II



GEOFFair: GEOmetric Framework for Fairness I

GEOmetric Framework for Fairness

- represents distributions, ML models, fairness constraints, and hypothesis spaces as vectors and sets
- enables visualization, allowing us to gain insights into the data or the model operation
- enables studying fairness properties in ML

GEOFFair: main definitions

Definition (Ground Vector $y^+ \in \mathcal{Y}^n$)

- data that can be observed and used as ground truth
- paired with the input vector x

Definition (Gold Vector $y^* \in \mathcal{Y}^n$)

- "unbiased" data
- $y^+ = b(y^*)$, where $b : \mathcal{Y}^n \to \mathcal{Y}^n$ is called the biased mapping

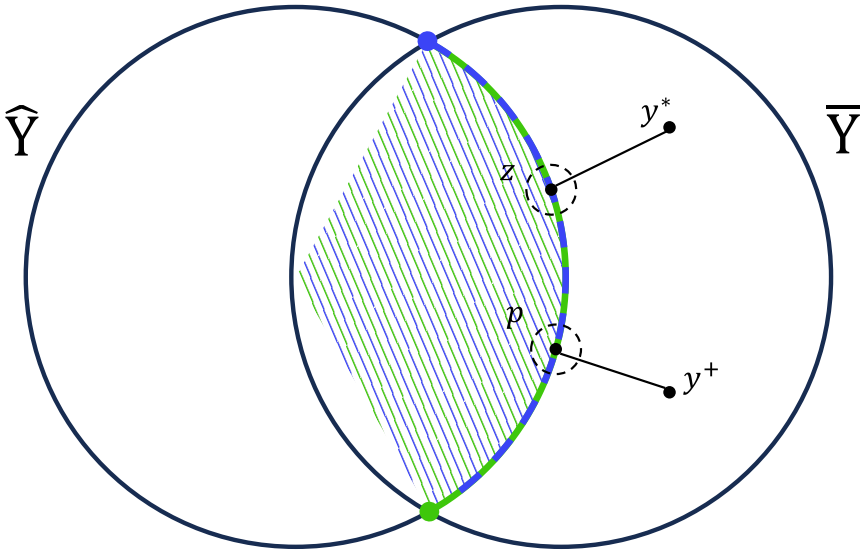
Definition (Hypothesis Space, $\hat{\mathbb{Y}}$)

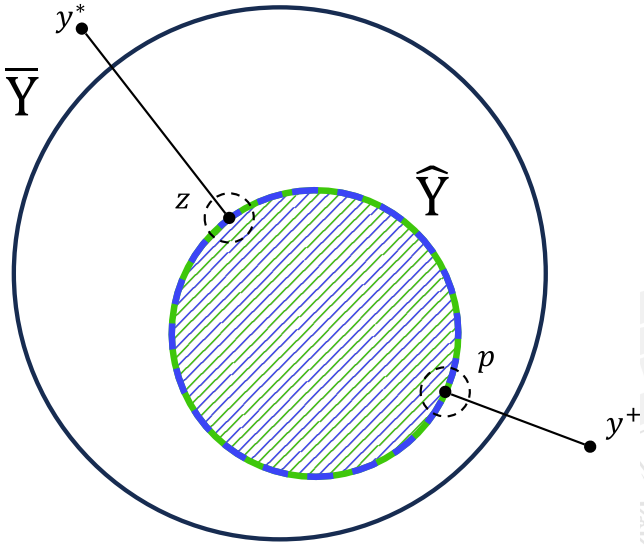
- set of possible outputs for the chosen class of ML models, i.e.
- $\hat{\mathbb{Y}} = \{y \in \mathcal{Y}^n \mid \exists f \in \mathcal{F} : f(x) = y\}$

Definition (Fair Space, $\overline{\mathbb{Y}} \subseteq \mathcal{Y}^n$)

- set containing all the output vectors aligned with the fairness requirements

GEOFFair: examples





FAiRDAS: Fairness-Aware Ranking as Dynamic Abstract System I

Fairness-Aware Ranking as Dynamic Abstract System

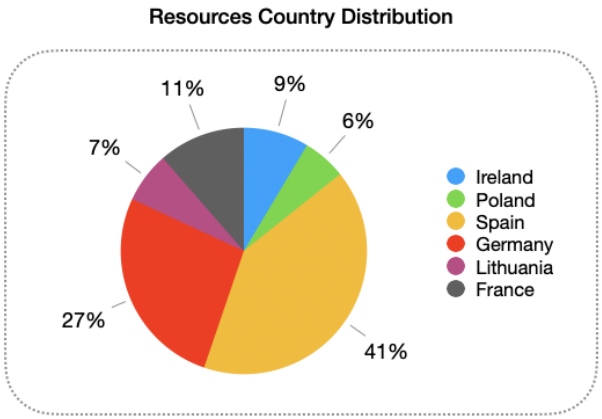
Long-term fairness as an abstract dynamical system

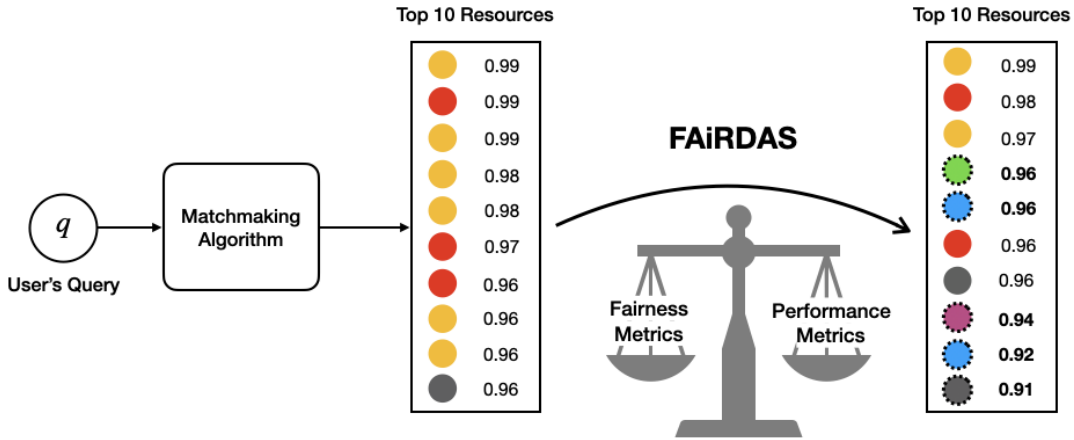
- define metrics of interests (fairness metrics, performance metrics, etc.)
- define the threshold for each metric
- Evolve the system in such a way the metrics remain below the thresholds

FAiRDAS: examples

FAiRDAS: Fairness-Aware Ranking as Dynamic Abstract System

Eleonora Misino, Roberta Calegari, Michele Lombardi, Michela Milano





FAiRDAS Goal

FAiRDAS Key Ingredients

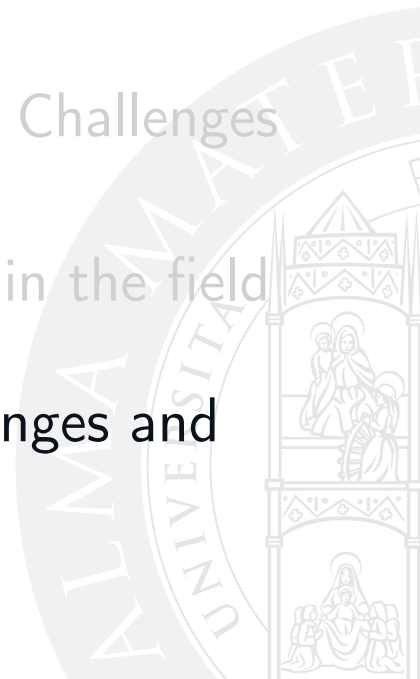
Actions on Rank

- Move down in the rank
- Move up in the rank
- ....

Metrics Thresholds

- Threshold Fairness
- Threshold Performance

Next in Line...



Conclusions, Challenges and Opportunities I

Which Mechanisms and When?

- legal, ethical, and social context
- selection of the best phase in which to act has dependencies with the data, the availability of the sensitive attributes at testing time, and the fairness notion selected
- context setups can vary between applications

Why Fairness in the AI Lifecycle?

- incorporate fairness needs into the software operations, making it more sustainable from social and technical perspectives
- incorporating fairness seamlessly after the software is operational is in many cases unrealistic given this complexity

Conclusions, Challenges and Opportunities II

Gaps and Challenges

- *Educational* aspect of AI practitioners
- Lack of a *methodological approach* to tackle fairness in the different stages of the AI lifecycle
- *Diversification* is needed beyond existing algorithms and datasets
- *Fairness metrics* need to be balanced between individual and group notions
- *Experimentation environments* are required to provide an easy playground to test different notions and techniques

References I

[Adel et al., 2019] Adel, T., Valera, I., Ghahramani, Z., and Weller, A. (2019).
One-network adversarial fairness.
In *AAAI*, volume 33, pages 2412–2420.

[Agarwal et al., 2018] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018).
A reductions approach to fair classification.
In *International Conference on ML*, pages 60–69. PMLR.

[Aïvodji et al., 2021] Aïvodji, U., Ferry, J., Gambs, S., Huguet, M.-J., and Siala, M. (2021).
Faircorels, an open-source library for learning fair rule lists.
In *International Conference on Information & Knowledge Management*, pages 4665–4669.

[Awasthi et al., 2021] Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. (2021).
Evaluating fairness of machine learning models under uncertain and incomplete information.
In *2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 206–214.

[Bacelar, 2021] Bacelar, M. (2021).
Monitoring bias and fairness in machine learning models: A review.
ScienceOpen Preprints.

References II

[Bechavod and Ligett, 2017] Bechavod, Y. and Ligett, K. (2017).
Penalizing unfairness in binary classification.
arXiv:1707.00044.

[Beutel et al., 2017] Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017).
Data decisions and theoretical implications when adversarially learning fair representations.
arXiv:1707.00075.

[Calders and Verwer, 2010] Calders, T. and Verwer, S. (2010).
Three naive bayes approaches for discrimination-free classification.
Data mining and knowledge discovery, 21(2):277–292.

[Chen et al., 2019] Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019).
Fairness under unawareness: Assessing disparity when protected class is unobserved.
In *Conference on fairness, accountability, and transparency*, pages 339–348.

[Chiappa, 2019] Chiappa, S. (2019).
Path-specific counterfactual fairness.
In *AAAI*, volume 33, pages 7801–7808.

References III

[Corbett-Davies et al., 2017] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017).
Algorithmic decision making and the cost of fairness.
In *23rd acm sigkdd International Conference on knowledge discovery and data mining*, pages 797–806.

[Detassis et al., 2020] Detassis, F., Lombardi, M., and Milano, M. (2020).
Teaching the old dog new tricks: supervised learning with constraints.
In *NeHuAI@ ECAI*, pages 44–51.

[Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012).
Fairness through awareness.
In *3rd innovations in theoretical computer science Conference*, pages 214–226.

[Dwork et al., 2018] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018).
Decoupled classifiers for group-fair and efficient machine learning.
In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR.

[Edwards and Storkey, 2015] Edwards, H. and Storkey, A. (2015).
Censoring representations with an adversary.
arXiv:1511.05897.

References IV

[Feldman et al., 2015] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015).
Certifying and removing disparate impact.
In *21th ACM International Conference on knowledge discovery and data mining*, pages 259–268.

[Feng et al., 2019] Feng, R., Yang, Y., Lyu, Y., Tan, C., Sun, Y., and Wang, C. (2019).
Learning fair representations via an adversarial framework.
arXiv:1904.13341.

[Goh et al., 2016] Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. (2016).
Satisfying real-world goals with dataset constraints.
Advances in Neural Information Processing Systems, 29.

[Gupta et al., 2018] Gupta, M., Cotter, A., Fard, M. M., and Wang, S. (2018).
Proxy fairness.
arXiv:1806.11212.

[Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016).
Equality of opportunity in supervised learning.
Adv. in neural information processing systems, 29.

References V

[Ignatiev et al., 2020] Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E., and Marques-Silva, J. (2020).
Towards formal fairness in machine learning.
In *CP*, pages 846–867. Springer.

[Kamiran and Calders, 2012] Kamiran, F. and Calders, T. (2012).
Data preprocessing techniques for classification without discrimination.
Knowledge and information systems, 33(1):1–33.

[Kamiran et al., 2010] Kamiran, F., Calders, T., and Pechenizkiy, M. (2010).
Discrimination aware decision tree learning.
In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE.

[Kamishima et al., 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012).
Fairness-aware classifier with prejudice remover regularizer.
In *Joint European Conference on machine learning and knowledge discovery in databases*, pages 35–50. Springer.

[Kilbertus et al., 2017] Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017).
Avoiding discrimination through causal reasoning.
Advances in neural information processing systems, 30.

References VI

[Krasanakis et al., 2018] Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. (2018).
Adaptive sensitive reweighting to mitigate bias in fairness-aware classification.
In *2018 WWW Conference*, pages 853–862.

[Kusner et al., 2017] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017).
Counterfactual fairness.
Adv. in neural information processing systems, 30.

[Leavy et al., 2020] Leavy, S., O'Sullivan, B., and Siapera, E. (2020).
Data, power and bias in artificial intelligence.
arXiv:2008.07341.

[Leavy et al., 2021] Leavy, S., Siapera, E., and O'Sullivan, B. (2021).
Ethical data curation for ai: An approach based on feminist epistemology and critical theories of race.
In *AAAI/ACM Conference on AI, Ethics, and Society*, pages 695–703.

[Liu and Vicente, 2021] Liu, S. and Vicente, L. N. (2021).
The sharpe predictor for fairness in machine learning.
arXiv:2108.06415.

References VII

[Lohia et al., 2019] Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019).
Bias mitigation post-processing for individual and group fairness.
In *International Conference on acoustics, speech and signal processing*, pages 2847–2851. IEEE.

[Louizos et al., 2015] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015).
The variational fair autoencoder.
arXiv:1511.00830.

[Luong et al., 2011] Luong, B. T., Ruggieri, S., and Turini, F. (2011).
k-nn as an implementation of situation testing for discrimination discovery and prevention.
In *17th International Conference on Knowledge discovery and data mining*, pages 502–510.

[Madras et al., 2018] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018).
Learning adversarially fair and transferable representations.
In *International Conference on ML*, pages 3384–3393. PMLR.

[Menon and Williamson, 2018] Menon, A. K. and Williamson, R. C. (2018).
The cost of fairness in binary classification.
In *Conference on Fairness, Accountability and Transparency*, pages 107–118. PMLR.

References VIII

[Mhasawade and Chunara, 2021] Mhasawade, V. and Chunara, R. (2021).
Causal multi-level fairness.
In *AAAI/ACM Conf. on AI, Ethics, and Society*, pages 784–794.

[Pessach and Shmueli, 2021] Pessach, D. and Shmueli, E. (2021).
Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings.
Expert Systems with Applications, 185:115667.

[Pleiss et al., 2017] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017).
On fairness and calibration.
Advances in neural information processing systems, 30.

[Quadrianto and Sharmanska, 2017] Quadrianto, N. and Sharmanska, V. (2017).
Recycling privileged learning and distribution matching for fairness.
Advances in Neural Information Processing Systems, 30.

[Wang et al., 2019] Wang, H., Ustun, B., and Calmon, F. (2019).
Repairing without retraining: Avoiding disparate impact with counterfactual distributions.
In *International Conference on ML*, pages 6618–6627. PMLR.

References IX

[Woodworth et al., 2017] Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017).
Learning non-discriminatory predictors.
In *Conference on Learning Theory*, pages 1920–1953. PMLR.

[Zafar et al., 2017a] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017a).
Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment.
In *26th International Conference on WWW*, pages 1171–1180.

[Zafar et al., 2017b] Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017b).
Fairness constraints: Mechanisms for fair classification.
In *AI and Statistics*, pages 962–970. PMLR.

[Zemel et al., 2013] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013).
Learning fair representations.
In *International Conference on ML*, pages 325–333. PMLR.