Towards Ethical Intelligence: Navigating Fairness and Bias in AI

Roberta Calegari

Alma Mater Studiorum–Università di Bologna, Italy

BIAS 2023 — 3rd Workshop on Bias and Fairness in AI
Workshop at ECML PKDD 2023
Turin, Italy

22 September 2023

AEQUITAS
unbias AI

Funded by
the European Union

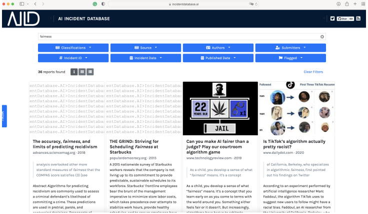
www.aequitas-project.eu
info@aequitas-project.eu

Next in Line. . .

1. Why fairness?
2. Goal of the talk
3. Fairness and bias in AI
4. Fairness Awareness
5. Enforcing Fairness
6. Our advancements in the field
7. Conclusions, Challenges and Opportunities

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/



36 reports found

The accuracy, *fairness*, and limits of predicting recidivism
advances.sciencemag.org · 2018

> analysis overlooked other more standard measures of *fairness* that the COMPAS score satisfies (3) [see

Abstract Algorithms for predicting recidivism are commonly used to assess a criminal defendant's likelihood of committing a crime. These predictions are used in pretrial, parole, and

THE GRIND: Striving for Scheduling *Fairness* at Starbucks
populardemocracy.org · 2015

A 2015 nationwide survey of Starbucks workers reveals that the company is not living up to its commitment to provide predictable, sustainable schedules to its workforce. Starbucks' frontline employees bear the brunt of the management imperative to minimize store labor costs, which takes precedence over attempts to stabilize work hours, provide healthy

Can you make AI fairer than a judge? Play our courtroom algorithm game
www.technologyreview.com · 2019

> As a child, you develop a sense of what "*fairness*" means. It's a concept

As a child, you develop a sense of what "fairness" means. It's a concept that you learn early on as you come to terms with the world around you. Something either feels fair or it doesn't. But increasingly,

Is TikTok's algorithm actually pretty racist?
www.dailydot.com · 2020

> of California, Berkeley, who specializes in algorithmic *fairness*, first pointed out his findings on Twitter

According to an experiment performed by artificial intelligence researcher Marc Faddoul, the algorithm TikTok uses to suggest new users to follow might have a racial bias. Faddoul, an AI researcher from

EG-TAI: TAI Requirements & AI Act

Main pillars

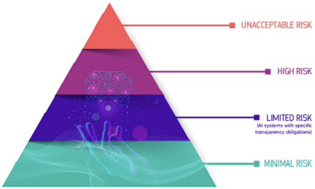
- lawfulness
- ethics
- robustness

$\underbrace{\text{ethics} \quad \text{robustness}}$

Seven specific requirements – dimensions to be audited – of an AI system:

1. human agency and oversight
2. technical robustness and safety
3. privacy and data governance
4. transparency (traceability, explainability)
5. diversity, non-discrimination and *fairness*
6. societal and environmental well-being
7. accountability

AI Act



Next in Line. . .

1. Why fairness?
2. Goal of the talk
3. Fairness and bias in AI
4. Fairness Awareness
5. Enforcing Fairness
6. Our advancements in the field
7. Conclusions, Challenges and Opportunities

Outline

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

Next in Line...

1 Why fairness?

2 Goal of the talk

3 Fairness and bias in AI

4 Fairness Awareness

5 Enforcing Fairness

6 Our advancements in the field

7 Conclusions, Challenges and Opportunities

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

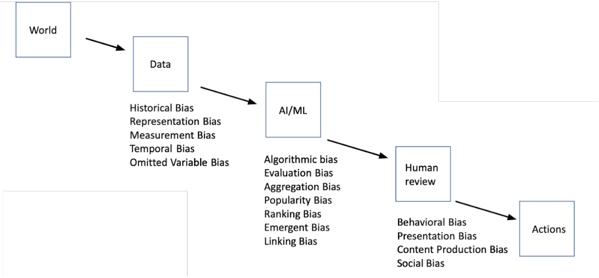
> *the absence of prejudice or preference for an individual or group based on their characteristics, i.e., absence of bias*

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



World

Data

Historical Bias
Representation Bias
Measurement Bias
Temporal Bias
Omitted Variable Bias

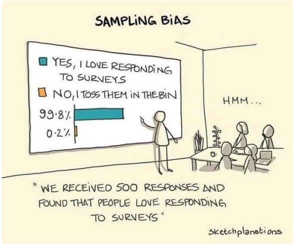
AI/ML

Algorithmic bias
Evaluation Bias
Aggregation Bias
Popularity Bias
Ranking Bias
Emergent Bias
Linking Bias

Human review

Behavioral Bias
Presentation Bias
Content Production Bias
Social Bias

Actions


SAMPLING BIAS
YES, I LOVE RESPONDING TO SURVEYS
NO, I TOSS THEM IN THE BIN
99·8%
0·2%
HMM...
"WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS"
sketchplanations

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. . .

1. Why fairness?
2. Goal of the talk
3. Fairness and bias in AI
4. **Fairness Awareness**
5. Enforcing Fairness
6. Our advancements in the field
7. Conclusions, Challenges and Opportunities

Fairness Awareness I

Two elements required

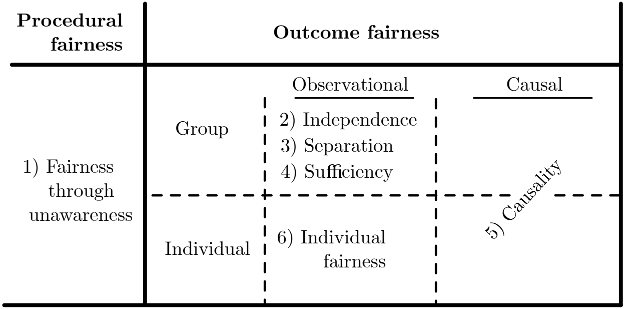
- Definition of *fairness notions* (context-dependent, social perspective)
- Quantitative mechanism to measure them

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

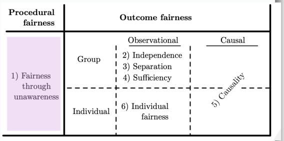


Procedural fairness	Outcome fairness		
		Observational	Causal
1) Fairness through unawareness	Group	2) Independence 3) Separation 4) Sufficiency	5) Causality
	Individual	6) Individual fairness	

Figure: Organising framework of algorithmic fairness metrics

Procedural Fairness

Procedural Fairness



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

- concept inherited from administrative law concerned with equality of treatment *within the process* that carries out a decision
- in the computational area
  - not including sensitive attributes in the AI algorithm → omission of sensitive attributes or *fairness through unawareness*

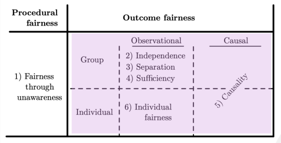
- model accuracy is reduced
- *discrimination effects* do not improve as a consequence of neglecting relationships with proxy
  - ignoring prejudice may not be caused by a single variable but rather by a combination of several ones

→ 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



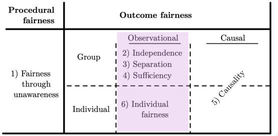
Procedural fairness	Outcome fairness		
		Observational	Causal
1) Fairness through unawareness	Group	2) Independence 3) Separation 4) Sufficiency	5) Causality
	Individual	6) Individual fairness	

- *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



Procedural fairness

Outcome fairness

Observational

Causal

Group

Individual

1) Fairness through unawareness

2) Independence
3) Separation
4) Sufficiency

5) Causality

6) Individual 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

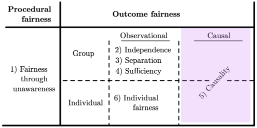
→ i.e., they do not evaluate what is not observable [Kilbertus et al., 2017]

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



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

Advantages & Limitations

- deeper understanding → more effective interventions
- more precise and effective fairness measures
- fairness by design: building fairness from the design phase

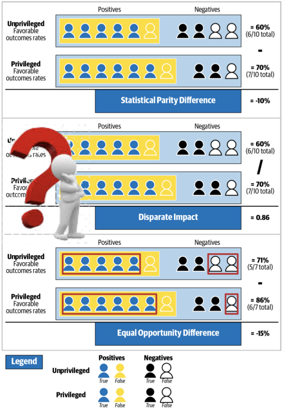
- complexity: deep understanding of causality, domain expertise, access to high-quality data (data availability)
- resource-intensive, in terms of expertise and computational resources
- less interpretable than traditional machine learning models
- sometimes conflicts with legal and regulatory requirements

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
- ...

Next in Line. . .

1. Why fairness?
2. Goal of the talk
3. Fairness and bias in AI
4. Fairness Awareness
5. Enforcing Fairness
6. Our advancements in the field
7. Conclusions, Challenges and Opportunities

Fairness: timing of intervention

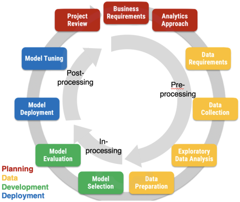


Figure: AI lifecycle & fairness intervention time

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

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).

Enforcing Fairness: Challenges

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

Next in Line. . .

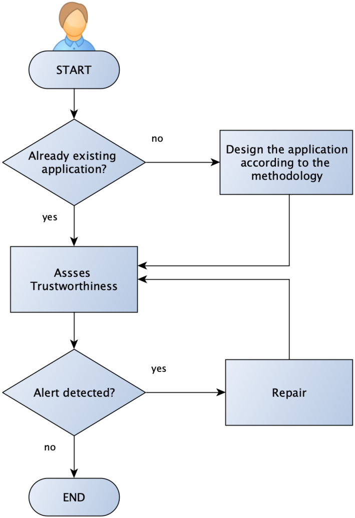
1. Why fairness?
2. Goal of the talk
3. Fairness and bias in AI
4. Fairness Awareness
5. Enforcing Fairness
6. Our advancements in the field
7. Conclusions, Challenges and Opportunities

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

Core idea

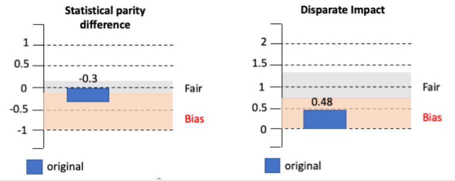
Open ***controlled experimentation environment*** *for AI stakeholders – provided as* **a service** *on the AI on demand platform – to test* ***fairness*** *dimensions via* ***controlled experiments*** *and to design* ***trust-by-design*** *AI applications*

The project idea: workflow


START
Already existing application?
no
Design the application according to the methodology
yes
Assses Trustworthiness
Alert detected?
yes
Repair
no
END

The project idea: an example I

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

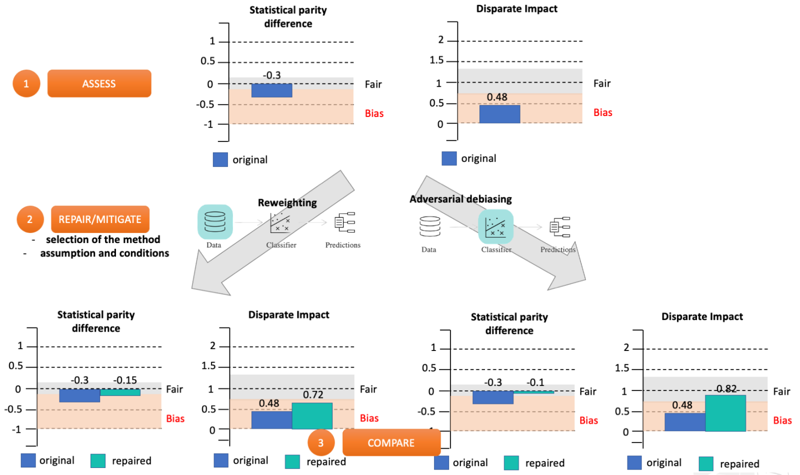


TAI dimension: **fairness** metrics (*assumptions* to reach fairness)

→ 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



1 ASSESS

2 REPAIR/MITIGATE

- selection of the method
- assumption and conditions

3 COMPARE

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

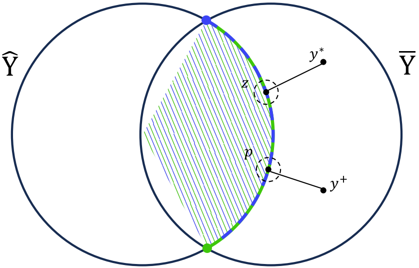
Definition (Hypothesis Space, $\hat{\mathbb{Y}}$)

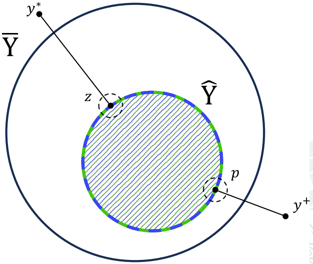
- set of possible outputs for the chosen class of ML models, i.e.
- $\hat{\mathbb{Y}} = \{y \in \mathcal{Y}^n \mid \exists f \in \mathcal{F} : f(x) = y\}$

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

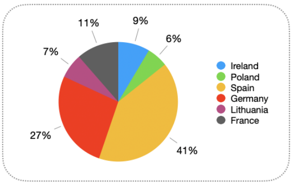
- define metrics of interests (fairness metrics, performance metrics, etc.)
- define the threshold for each metric
- Evolve the system in such a way the metrics remain below the thresholds

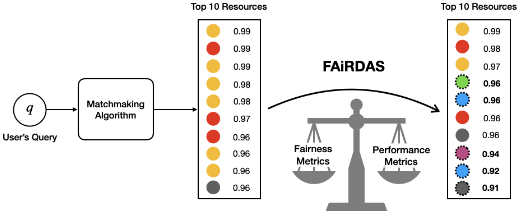
FAiRDAS: examples

FAiRDAS: Fairness-Aware Ranking as Dynamic Abstract System

Eleonora Misino, Roberta Calegari, Michele Lombardi, Michela Milano

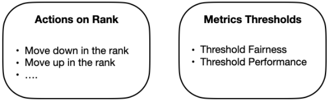
Resources Country Distribution





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. . .

1 Why fairness?

2 Goal of the talk

3 Fairness and bias in AI

4 Fairness Awareness

5 Enforcing Fairness

6 Our advancements in the field

7 Conclusions, Challenges and Opportunities

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.