TELECOM SudParis

INSTITUT POLYTECHNIQUE DE PARIS

HOW TO DESIGN FAIR AND UNBIASED ALGORITHMS?

DATA AI 951 – AI ETHICS

SOPHIE CHABRIDON
PROFESSOR IN COMPUTER SCIENCE

Institut Mines-Télécom

Outline

Introduction

- Algorithmic decision-making (ADM) systems widely used in various domains and already substantially affect humans lives for credit lending, health, criminal justice and any kind of recommendations
- Intuitively algorithms are expected to be more objective and impartial than humans
- However many cases already demonstrate the presence of bias in ADM systems and lack of fairness
- Provide transparent and explainable algorithms to facilitate fairness audits. Consider self-explainable solutions
- Making algorithms fair and unbiased is a very active research domain

Outline

What is a bias? I

- Humans have bias which can manifest in deviating perception, thinking, judgment or remembering
- What about algorithms?
- **Technical bias**: systematic deviation from a true state. In statistics, an estimator is biased when there is an error that causes it to not converge to the true value that it is trying to estimate
  - Taken into account by algorithms designers as they directly reduce the performance of the algorithms
  - Become problematic when they disadvantage a specific group and lead to **discrimination**
- **Societal bias**:
  - An algorithm can be very accurate technically while being biased from a societal point of view
  - Reproduces, via the algorithm, biases already present in the society
  - Often leads to arbitrate to the disadvantage of already disadvantaged populations
- Bias, if not controlled for, may cause unfairness and **discrimination**

What is a bias? II

- Consider **algorithmic biases** including both technical biases and societal biases
- **Technical bias**
  - Well known to statisticians and can be measured
  - Reduce the performance of the algorithm, hindering the achievement of its objective
  - Mitigating technical biases has a cost and requires effort, but often also a clear benefit for developers
- **Societal bias**
  - Less well defined or formalized
  - Following societal biases may allow algorithms to perform better
  - When it comes to advertising or job postings, sticking to stereotypes can maximize the number of clicks on the ads
  - Societal biases can be deliberate to meet a business strategy

Categories of bias

- Synthetizing the analysis in [Tolan, 2018], [Bertail et al., 2019] and [Institut Montaigne, 2020], we identify three main categories of bias
- **Statistical**: bias in data, omitted variable bias, selection bias, endogeneity bias
- **Psychological**: emotional and cognitive bias
- **Economic** bias serving a business strategy

Statistical biases I

Bias in Data

- Most common source of technical bias, especially in terms of data quality or representation
- Biases pre-exist algorithms and mainly come from the data they use
- Quality of data and relevance for the pursued objective must be verified
- "Garbage in - Garbage out" principle
- With reinforcement learning, this issue is amplified by self-consuming data produced by the algorithm
- If labels on training images are wrong, the final result will be biased
- If training samples are not coming from real data, an algorithm may not be able to predict a real phenomenon

Statistical biases II

Omitted variable bias

- Not always possible to collect complete data
- Some data may replaced with proxies or approximated

- For example, soft skills such as leadership or emotional intelligence are difficult to measure and may be negatively correlated with academic performance. A selection algorithm considering only academic performance would surely fail to identify certain high-potential individuals

Statistical biases III

Selection bias

- Occurs when the learning sample is not representative of the population concerned
- Possible when proper randomization is not achieved

- Well-known example of commercial image analysis programs. Difficulty for classifying the gender of dark-skinned individuals, a shortcoming that is potentially due to the relative dearth of dark-skinned faces in popular facial analysis datasets [Buolamwini and Gebru, 2018]

Statistical biases IV

Endogeneity bias

- An endogenous variable is a variable whose value is determined by the model
- In econometrics, an endogenous random variable is correlated with the error term, while an exogenous variable is not
- Important reason for incorrect causal inferences
- Potential sources, of endogeneity: omission of variables, errors in variables or simultaneous causality
- In practice, it happens in situations where there is a need to anticipate the future and where historical data become useless to predict some outcome
- Example: Consider person with a bad reimboursement history in a bank, with an overdrawn account. If this person lives as a couple and starts a family, he or she will surely change lifestyle and starts saving money. A human is able to understand this change and anticipate a future situation, this is more difficult for an algorithm based solely on the past

Psychological biases I

Emotional (or affective) biases

- Distorsion in the manner an information is treated, in contrast to a rational behavior
- In general, humans refuse to believe in unpleasant realities
- **Panurge or bandwagon bias**
  - An individual might blindly follow others regardless of the consequences
  - Case of a developer following a popular model without checking its relevance and accuracy
- **Confirmation bias**
  - Contribute to overconfidence in personal beliefs
  - Contrary evidence can even maintain or strengthen beliefs
  - Developers may favor their own vision of the world

Psychological biases II

Cognitive biases

- **Stereotypes**
  - Related to generalization
  - Treat persons according to the group they belong to, and the traits usually associated with that group, rather than on their individual characteristics
  - Very few stereotypes are openly acknowledged, however, implicit stereotypes are widespread
  - Only 10% of the population explicitly acknowledges being biased [Bolukbasi et al., 2016]
  - Famous American examples: Facebook's real estate ads, where Facebook was differentially showing ads for housing by gender and race, and Amazon's résumé rating, not gender-neutral
  - The Democrats proposed an Algorithm Accountability Act to House and Senate in 2019

Psychological biases III

Cognitive biases

- **Illusory correlation bias**
  - Perceiving a relationship between variables even when no such relationship exists
  - May be formed because rare or novel occurrences are more salient and therefore tend to capture one's attention

- Cognitive biases may lead developers to choose variables according to their own perception of a phenomenon, transposing their own bias to the algorithms

Economic biases

- An algorithm may contain a bias voluntarily or involuntarily for reasons of business strategy
- Example:

  An algorithm that simply optimizes the cost-effectiveness of a job posting, may display fewer advertisements to young women than to young men

  Advertising space for young women is more expensive than advertising space for young men. This happens because young women are a prized demographic and are more expensive to show ads to [Lambrecht and Tucker, 2019]

  Less costly for the algorithm to prefer men for these job ads

  The commercial strategy to recruit while minimizing recruitment costs could thus lead to discrimination against women

Outline

Some definitions for classification applications

A classifier $\hat{Y}$ is a mapping from the space of possible values for variable X to the space of values of the target variable Y

Common classification criteria:

Event	Condition	Resulting Probability $P\{event\|condition\}$
$\hat{Y} = 1$	$Y = 1$	True positive rate or **Recall**
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

Addition classification criteria, swapping event and condition:

Event	Condition	Resulting Probability $P\{event\|condition\}$
$Y = 1$	$\hat{Y} = 1$	Positive predicted value or **Precision**
$Y = 0$	$\hat{Y} = 0$	Negative predicted value

How to define fairness?

- Many definitions of fairness exist, to be used in different contexts and application domains
- Tutorial at ACM FAT* 2018 conference discusses 21 definitions of fairness [Narayanan, 2018]
- [Berk et al., 2018] considers that at least 6 kinds of fairness can be used in the domain of criminal justice

- Distinguish group and individual fairness
  - **Group fairness** has been extensively studied, related to statistical approaches
  - **Individual fairness** requires to consider similar individuals similarly
- We first investigate fairness criteria based on the concepts of **Independence**, **Separation** and **Sufficiency** as proposed in [Barocas et al., 2020]

Concepts of Independence, Separation and Sufficiency

[Barocas et al., 2020]

- Most proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the classifier or score R.
- Consider fairness criteria as categories of different conditional independence statements between these random variables

Independence	**Separation**	**Sufficiency**
$R \perp A$ or $A \perp R$	$R \perp A \mid Y$	$Y \perp A \mid R$

Independence

$$R \perp A \text{ or } A \perp R$$

- Generally speaking, independence holds when A and R have no mutual information
- Sensitive attribute A is statistically independent of the score R

- Example in a hiring process: the result of getting a given job is independent of the gender attribute

Separation

$$R \perp A \mid Y$$

- Sensitive attribute A may be correlated with the target Y
- R is separated from A by the target variable Y, i.e. R is conditionally independent of A given Y
- All groups should experience the same false negative rate and the same false positive rate
- Can be seen as equalizing odds and opportunities

Sufficiency

$$Y \perp A \mid R$$

- The score already subsumes sensitive attribute A for predicting the target
- Requires a parity of positive/negative predictive values across all groups
- Related to **calibration**
  - Sufficiency and calibration by group are equivalent notions
  - Methods for calibration can be applied in practice to achieve sufficiency

Group fairness

- Ensures that an algorithm does not arbitrarily disadvantage a certain group
- Consider **Independence** as the underlying concept
- Variants of **group fairness**: **demographic parity**, **statistical parity**, **disparate impact**, etc.

Group Fairness and other statistical fairness criteria

The 3 concepts of Independence, Separation and Sufficiency allow to categorize fairness criteria [Barocas et al., 2020]

Independence	**Separation**	**Sufficiency**
Statistical parity	Equal opportunity	Cleary model
Group fairness	Equalized odds	Cond. use accuracy
Demographic parity	Cond. procedure accuracy	Predictive parity
Conditional statistical parity	Avoiding disparate mistreatment	Calibration within groups
Darlington criterion (4)	Balance for negative class	Darlington criterion (1), (2)
	Balance for the positive class	
	Equalized correlations	
	Predictive equality	
	Darlington criterion (3)	

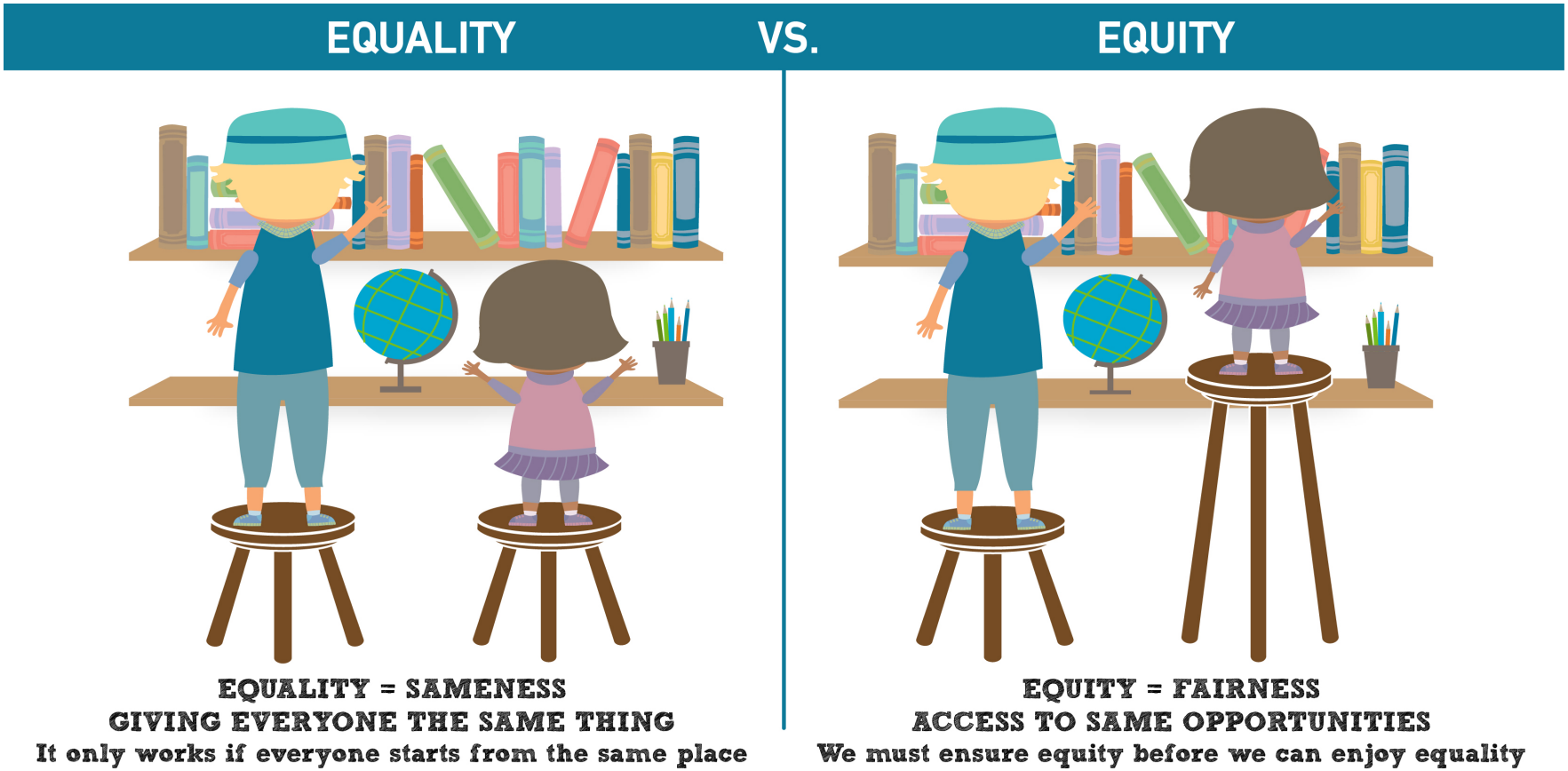
Fairness criteria - Impossibility result

- Important impossibility result for statistical definitions of fairness [Chouldechova, 2017, Kleinberg et al., 2017]
  - Consider a small number of protected groups (based on a protected attribute) and a binary classification task (TN: True Negatives, FP: False Positives, FN: False Negatives and TP: True positives)
  - Goal: parity of some statistical measure across all of these groups
  - Except in trivial settings, it is impossible to simultaneously equalize false positive rates, false negative rates, and positive predictive value for different groups
  - The impossibility remains for any three rates. [Chouldechova, 2017] suggests to use only two measures FP and FN rates
- Due to the trade-off between separation and sufficiency [Barocas et al., 2020]

Equality vs equity I

- Equality does not imply equity. Considering fairness simply as uniform distribution is not sufficient
- For equity, there is a need to consider fairness as justice [Schement, 2001]

Equality vs equity II



Equality vs equity III

- Equality
  - Equal distribution or sharing, without considering individual needs
  - Might create unfair opportunities for achieving success
- Equity
  - Need-based approach. Distribution to provide individuals with the resources they require for being successful
  - Related to impartiality and fairness
- Equity can be called positive discrimination, whereas equality might lead to negative discrimination
- Equity can help in minimizing gaps between people and groups
- Equity is subjective and cannot be measured with certainty, whereas equality is objective and can be measured

See [difference101, 2020]

Individual fairness

Importance of metrics

- In general, similar people should be treated similarly
- With a metric, it becomes possible to limit the amount by which the treatment of two different people can differ
- Need for different notions of similarity considering different objectives
- A fairness metric has to be related to the application domain

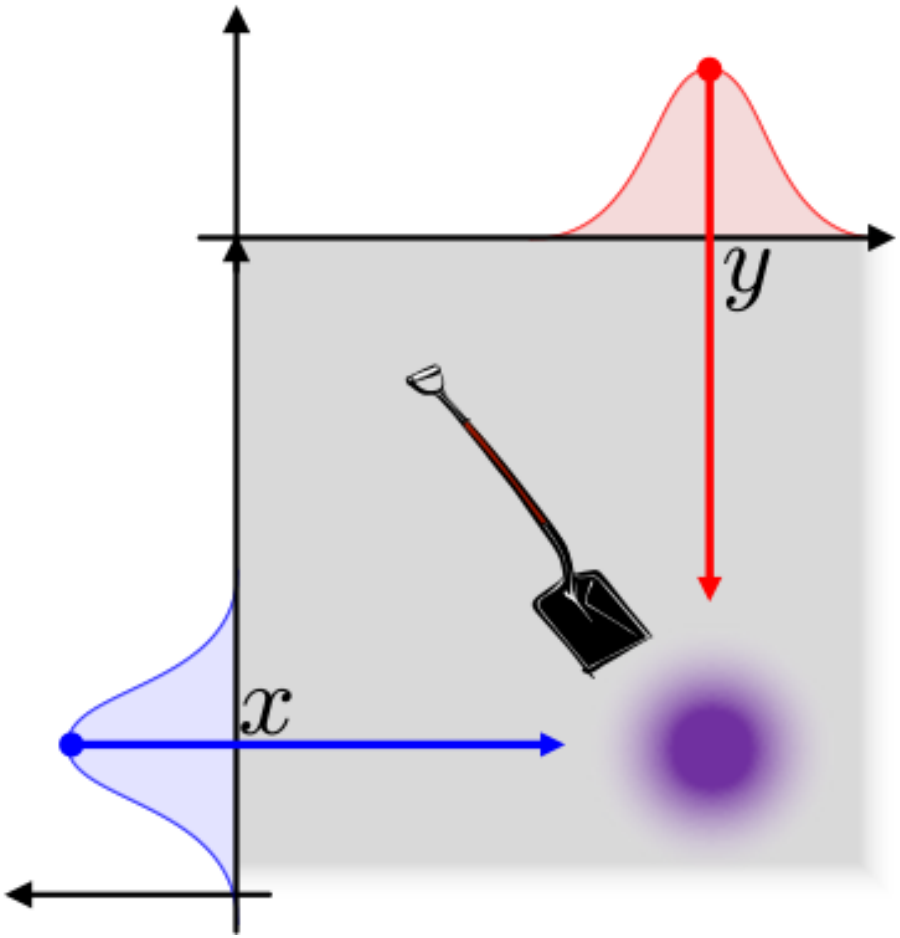
Fairness through awareness [Dwork et al., 2012]

1. Determine a task specific similarity metric
   - **Earthmover distance**: Determine the work necessary for a road mender to carry earth to fill holes in a road. After all earth heaps are moved to all holes, the work is established
2. Determine fairness by solving a linear optimization problem to minimize the loss of utility
3. Demonstrate that when the **Lipschitz** condition holds, i.e. the distance between the outcomes is not higher than the distance between the inputs

4. The paper also demonstrates that individual fairness implies group fairness as statistical parity under the condition that the Earthmover distance between the two considered groups is small

Fairness through awareness [Dwork et al., 2012]

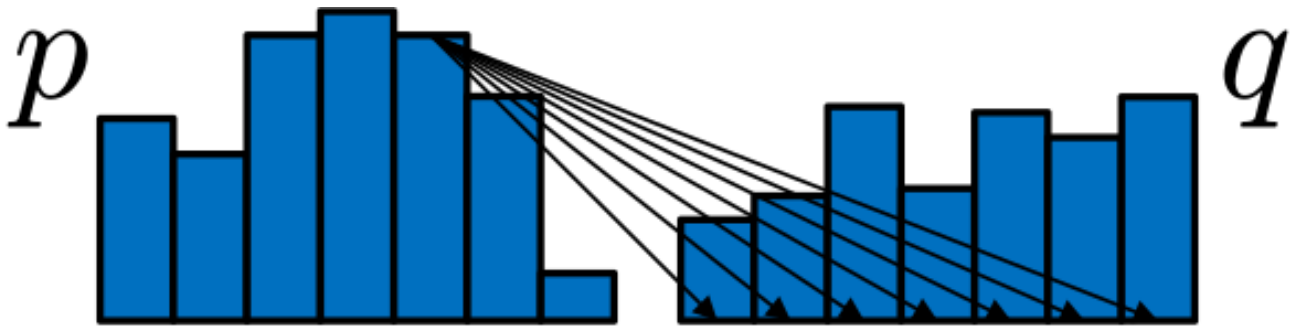
Earthmover distance



For distribution F, cost to move mass m from x to y:
$m \times f(x, y)$

Fairness through awareness [Dwork et al., 2012]

Earthmover distance



$$
\begin{aligned}
min_T \ & \Sigma_{ij} \, T_{ij} \, f(x_i, y_j) \\
s.t. \ & \Sigma_j \, T_{ij} = p_i \\
& \Sigma_i \, T_{ij} = q_j \\
& T \geq 0
\end{aligned}
$$

Outline

By not using protected attributes?

About direct and indirect discrimination

- Is it possible to avoid discrimination by restraining the algorithm from taking into account protected attributes?
- Fairness as blindness means systems are designed to be blind to protected attributes
- With Big Data [Barocas and Selbst, 2016], protected attributes are almost always correlated with other attributes or with the predicted outcome itself
- [Lipton et al., 2018] shows it does not prevent disparate impact if other features are predictive or partly predictive of group membership
- Some research direction suggests to use the protected group feature during training but not for prediction

By not using protected attributes?

How to deal with protected attributes in an accountable way?

- Regulation requires that ADM systems are not discriminatory, and to minimize the use of private data
- Is it possible to guarantee both fairness and privacy?
- [Kilbertus et al., 2018] proposes to encrypt sensitive attributes and to use secure multi-party computation
  - Encrypted data exchanged between individual users and service providers
  - Allows also fairness certification while keeping sensitive attributes encrypted to both the regulator and the service provider
- [Žliobaitė and Custers, 2016] demonstrated the necessity of using sensitive attributes to guarantee fairness and some contradiction between fairness and privacy
  - In order to guarantee non-discrimination, for instance, with respect to race, the sensitive racial information needs to be used in the model building process

Approaches for bias mitigation

[Ntoutsi et al., 2020] identifies three families of bias mitigation methods

- **Preprocessing methods** focusing on the data
- **Inprocessing methods** focusing on the ML algorithm
- **Post-processing methods** on the ML model

Pre-processing methods

- Modify the original data distribution
  - By altering class labels of instances close to the decision boundary
  - Probabilistic modification while controlling the distortion and preserving utility
- Assign different weights to instances based on their group membership
- Balance the protected and unprotected groups in the training set using heuristics

Pre-processing methods

Focus on diversity

- [Drosou et al., 2017] considers different ways that **diversity** may help to remove data and selection biases
- With sample bias, training data is not representative of the overall population
- Diversity relates to the quality of a collection of items, or of a composite item
- Ensures that different kinds of objects are present in the output of an algorithmic process
- Important for both ethical reasons (mitigate risks of exclusion) and utilitarian reasons (to enable more powerful, accurate and engaging data analysis)
- Various measures based on distance, coverage, novelty

In-processing methods

- For classification:
  - Incorporate the model's discrimination behavior in the objective function through regularization or constraints
  - Suppose latent target labels and iteratively train towards those classes by altering weights of the instances
  - Optimizes for balanced error instead of overall error to account for class imbalance

- For unsupervised learning:
  - Fair-PCA (principal component analysis) approach by forcing equal reconstruction errors for both protected and unprotected groups
  - Fair clustering with approximately equal representation for each protected group in every cluster

Post-processing methods

Modify the classification model once it has been learned from data

- White-box approaches: alter the model's internals
  - Correct the confidence of classification rules
  - Modify probabilities in Bayes models
  - Change the class label at leaves of decision trees
- Recently, in-processing methods are preferred to white-box post-processing approaches

- Black-box approaches: alter the model's predictions
  - Enforce proportionality of decisions among protected versus unprotected groups
  - By differentiating the decision boundary itself over groups
  - By wrapping a fair classifier on top of a black-box base classifier

Outline

Research projects

- UnBias: Emancipating Users Against Algorithmic Biases for a Trusted Digital Economy
- ReEntrust: Rebuilding and Enhancing Trust in algorithms
- AI Now Institute, NYU, New York, USA
- No Bias project: Artificial Intelligence without Bias, Horizon 2020 European project

Initiatives for standards and Education

- Preparation of IEEE Standard P7003 for Algorithm Bias Considerations

- MOOC "Objectif IA" (in French) by Open Classroom [OpenClassroom and Institut Montaigne, 2020]

Outline

Conclusion

- Tests and thorough evaluations are required to determine whether algorithms are biased, but especially whether they are more or less biased than humans they replace or assist
- Automated Decisions systems present new risks but also new opportunities to reduce discrimination, as compared to human decisions
- ADM require to articulate decision-making objectives and to make explicit the tradeoffs between desiderata
- Researchers need to clearly understand the application domain to effectively formulate and test hypotheses about sources and mechanisms of unfairness

References I

Barocas, S., Hardt, M., and Narayanan, A. (2020).
Fairness and Machine Learning – Limitations and Opportunities.
Book in progress.

Barocas, S. and Selbst, A. D. (2016).
Big Data's Disparate Impact.
California Law Review, 104(3):671–732.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018).
Fairness in Criminal Justice Risk Assessments: The State of the Art.
Sociological Methods & Research.

Bertail, P., Bounie, D., Clémençon, S., Waelbroeck, P., and Fondation Abeona (2019).
Algorithmes : biais, discrimination et équité.
https://hal.telecom-paris.fr/hal-02077745.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016).
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.
In *Conf. on Neural Information Processing Systems, Barcelona, Spain*, pages 4349–4357.

Buolamwini, J. and Gebru, T. (2018).
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.
In *Conference on Fairness, Accountability and Transparency FAT, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Chouldechova, A. (2017).
Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.
Big Data, 5(2):153–163.

References II

difference101 (2020).
Equality vs Equity.

Drosou, M., Jagadish, H. V., Pitoura, E., and Stoyanovich, J. (2017).
Diversity in Big Data: A Review.
Big Data, 5(2):73–84.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012).
Fairness through awareness.
In *ACM Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA*, pages 214–226.

Institut Montaigne (2020).
Algorithms: Please mind the bias!

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., and Weller, A. (2018).
Blind Justice: Fairness with Encrypted Sensitive Attributes.
In *Int. Conference on Machine Learning (ICML), Stockholm, Sweden*, volume 80, pages 2635–2644. PMLR.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017).
Inherent Trade-Offs in the Fair Determination of Risk Scores.
In *8th Innovations in Theoretical Computer Science ConferenceITCS, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Lambrecht, A. and Tucker, C. (2019).
Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads.
Management Science, 65(7):2966–2981.

References III

Lipton, Z. C., Wang, Y., and Smola, A. J. (2018).

Detecting and Correcting for Label Shift with Black Box Predictors.

In *Int. Conference on Machine Learning (ICML), Stockholm, Sweden*, volume 80, pages 3128–3136. PMLR.

Narayanan, A. (2018).

Tutorial: 21 fairness definition and their politics.

ACM FAT* (Fairness, Accountability and Transparency) Conference, youtube.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernández, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., and Staab, S. (2020).

Bias in data-driven artificial intelligence systems - an introductory survey.

Wiley Interdiscip. Review on Data Mining and Knowledge Discovery, 10(3).

OpenClassroom and Institut Montaigne (2020).

Objectif IA : initiez-vous à l'intelligence artificielle url.

Schement, J. R. (2001).

Imagining fairness: Equality and equity of access in search of democracy.

Libraries and Democracy: The Cornerstones of Liberty, American Library Association.

Tolan, S. (2018).

Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges.

Joint Research Centre Conference and Workshop Reports, Publications Office of the European Union, Digital Economy Working Paper, JRC113750.

References IV

Žliobaitė, I. and Custers, B. (2016).

Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law*, 24(2):183–201.