Fairness in Algorithmic Decision

Thierry KIRAT, Directeur de recherche au CNRS –
Professeur attaché à l'Université PSL
Université Paris Dauphine-Paris Sciences et Lettres
Graduate Program Data Science – PSL Preparatory Week,
sept. 2021

Summary

PART 1. Insights into technique and political philosophy

PART 2. Discrimination

PART 3. Explicability of algorithmic decisions

PART I

Fairness: insights into technique and political philosophy

- Plurality of fairness metrics
- Sources of bias
- Incompatibities between fairness metrics
- Political philosophy – utility and fairness / normative egalitarian considerations

Plurality of fairness metrics

See :

Dooa Abu Elyoues, "Contextual Fairness, Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness" (September 1, 2019). Journal of Law, Technology and Policy, forthcoming

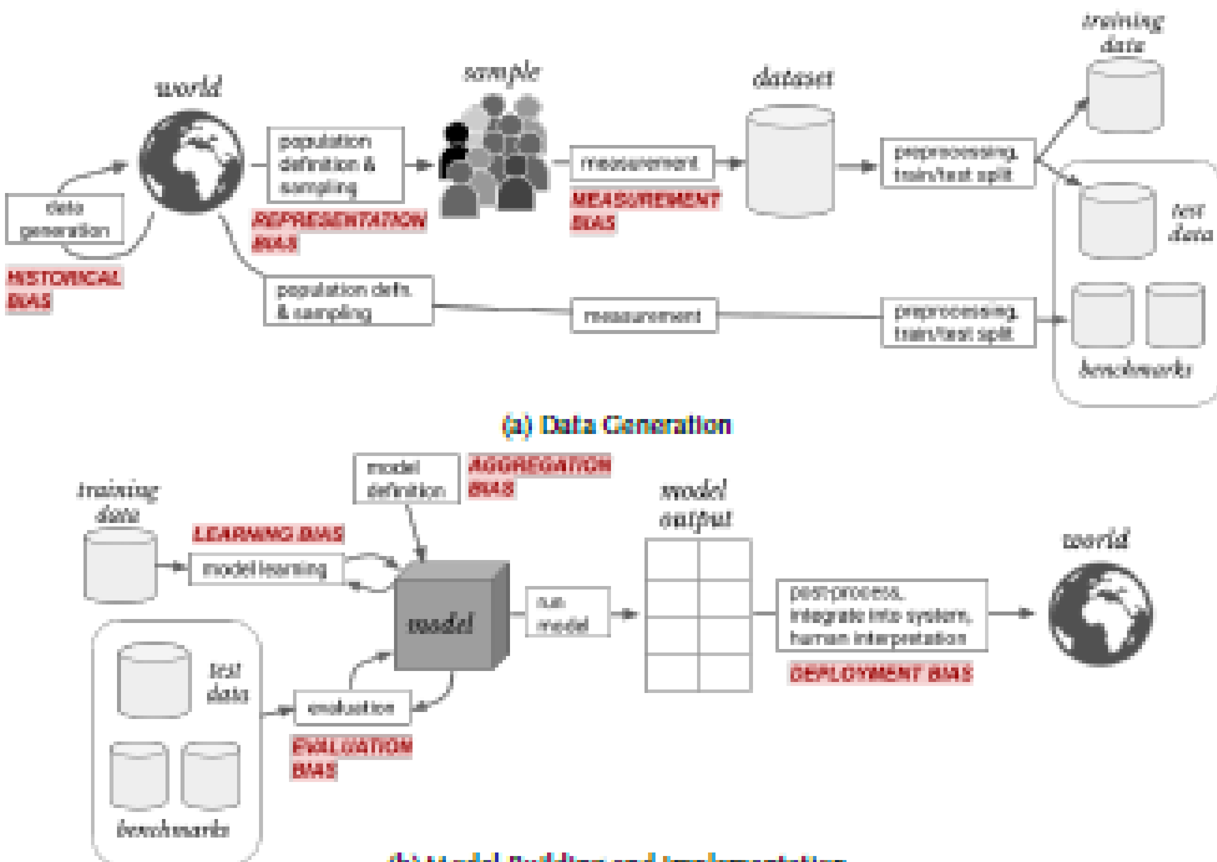
Arvind Narayanan, "Tutorial: 21 fairness definitions and their politics", March, 2018
https://www.youtube.com/watch?v=jIXIuYdnyyk

Table 1: Notions of fairness and summary of their corresponding legal mechanisms.

Notion	Sub-notion	Corresponding Legal Mechanism
Individual Fairness	The unaware approach	Equal opportunity as colorblindness
	Fairness through awareness	Equal opportunity based on similarities, and levels of scrutiny
Group fairness	Decoupling	Affirmative action (as separate but equal)
	Statistical or conditional parity	Affirmative action (preferably through critical diversity)
	Equal opportunity	Affirmative action (as equal opportunity)
	Equalized odds	Achieving equality by equalizing the false positive and false negative errors
	Calibration	Achieving equality by statistical significance
	Multicalibration	Achieving equality by statistical significance, and accounting for intersectionality
Causal Reasoning	Counterfactual fairness	Due process

Sources of Bias (1)

Suresh & Guttag, "A framework for understanding unattended consequences of machine learning" (2019)



(a) Data Generation

(b) Model Building and Implementation

Sources of Bias (2)

1. **historical bias** : occurs when the world as it is leads a model to produce outcomes that are not wanted
2. **Representation bias**: occurs when certain parts of the input space are underrepresented
3. **Measurement bias** : occurs when proxies are generated differently across groups, or the granularity(or quality of data) varies across groups…

Sources of Bias (3)

4. **Aggregation bias** : occurs when a one-size-fits-all model is used for groups with different conditional distributions $P(X \mid Y)$

5. **Evaluation bias** : occurs when the evaluation and/or benchmark data for an algorithm doesn't represent the target population

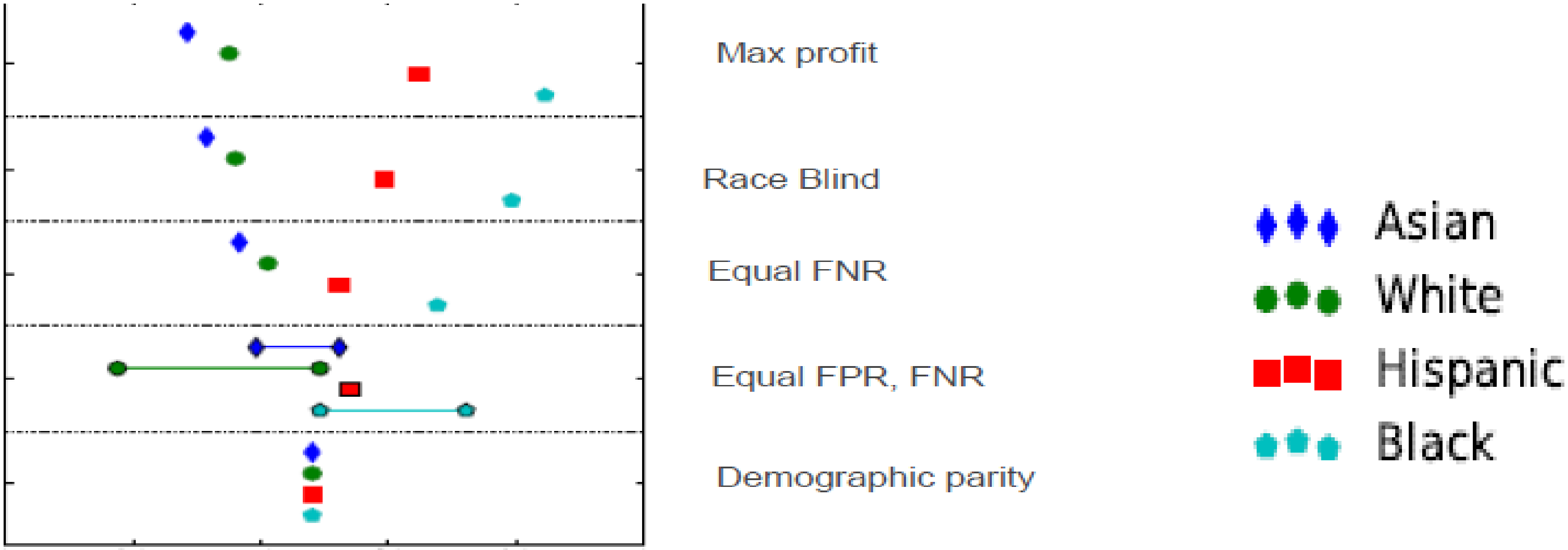
The Prediction Problem

		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$		

Fairness metrics: incompatibilities

See : Hardt, Price & Srebro, Equality of opportunity in machine learning, 2016 : access to bank credit by origin (FICO dataset, USA)

-> score scale : increasing risk of default



Success rate of loan applications for 4 racial groups

Fairness metrics: incompatibilities

Tutorial:

Martin Wattenberg, Fernanda Viégas, and Moritz Hardt, *Attacking discrimination with smarter machine learning* (companion to Hardt, Price & Srebro, “Equality of opportunity in machine learning”, 2016)

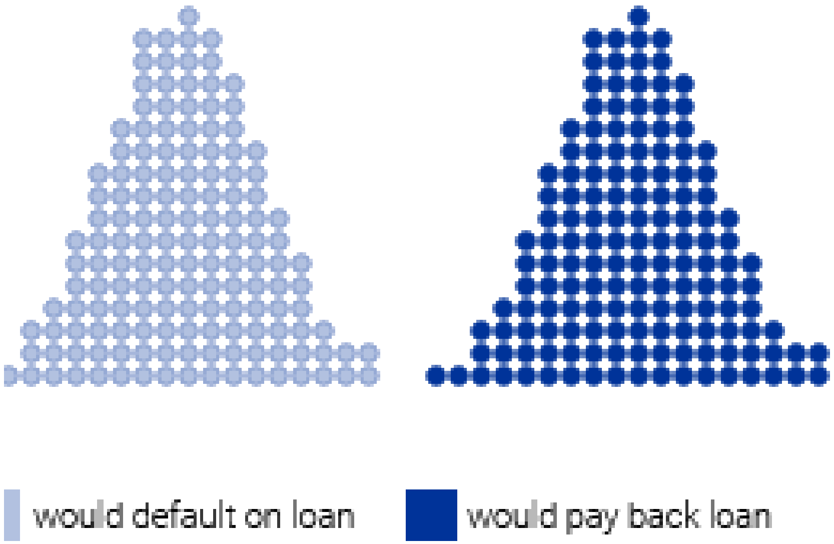
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

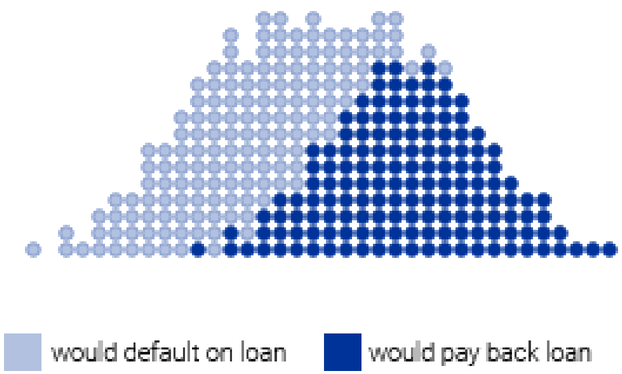
Attacking discrimination with smarter machine

Credit - Risk of default

Ideal : separating good and bad borrowers (left figure)

In practice the two groups can't be clearly separated (the FP problem….) (right figure)

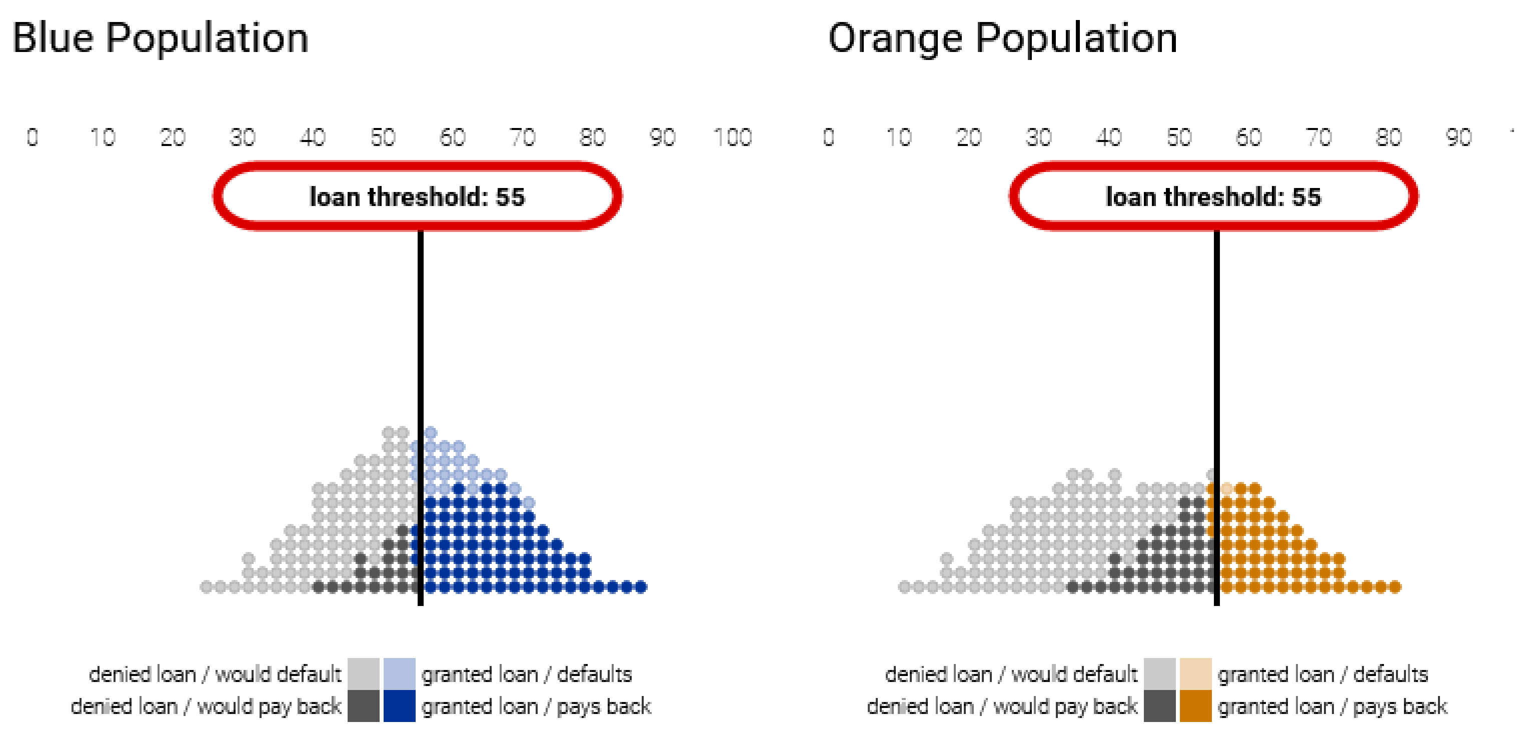




Attacking discrimination with smarter machine

Simulation # 1: <u>Group Unaware</u> - holds all groups to the same standard (same threshold on the risk-score scale)

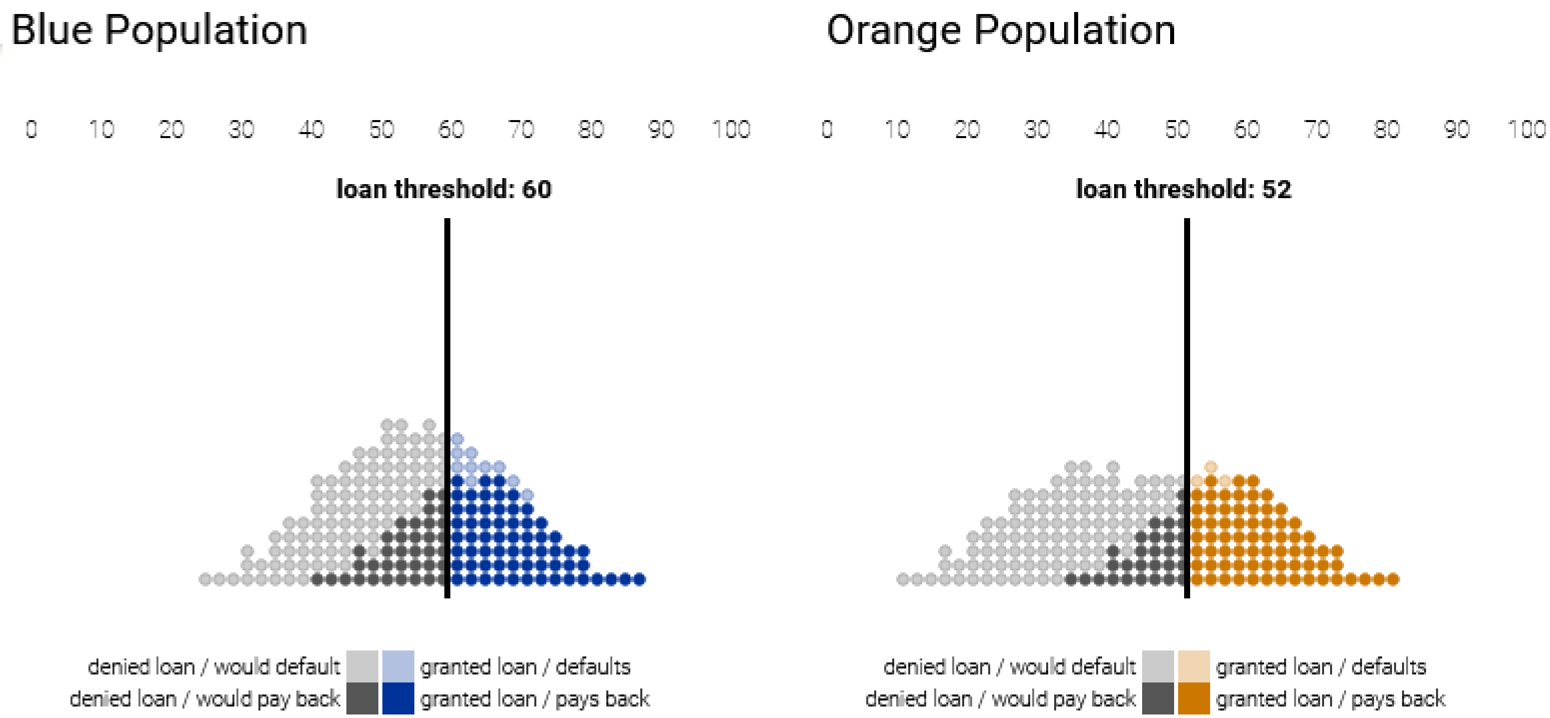
Both groups have the same threshold, but the orange group has been given fewer loans overall. Among people who would pay back a loan, the orange group is also at a disadvantage (FN).



Attacking discrimination with smarter machine

Simulation # 2: <u>Demographic parity -</u> If the goal is for the two groups to receive the same number of loans, then a natural criterion is demographic parity, where the bank uses loan thresholds that yield the same fraction of loans to each group. -> the "positive rate" is the same across both groups (37% of applicants obtain in loan in each group)

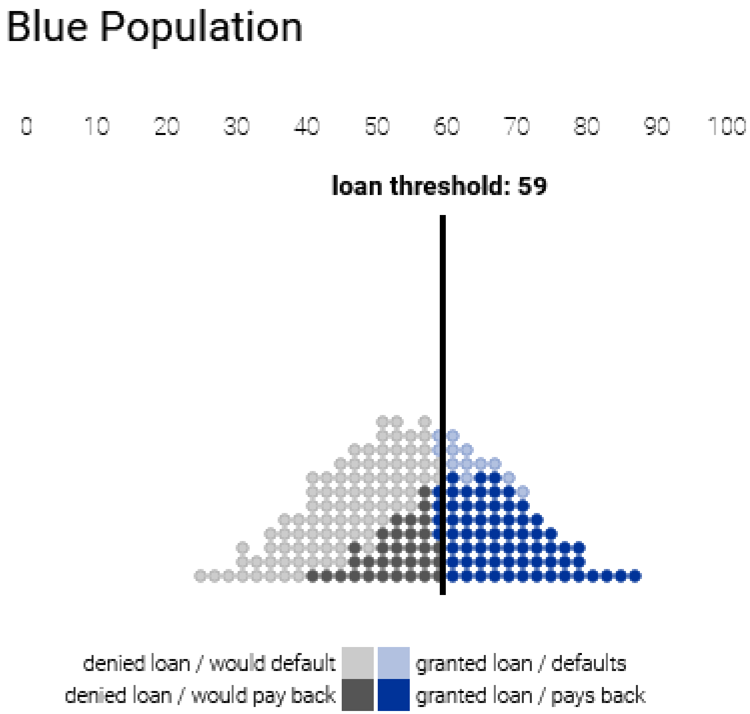
The number of loans given to each group is the same, but among people who would pay back a loan, the blue group is at a disadvantage.

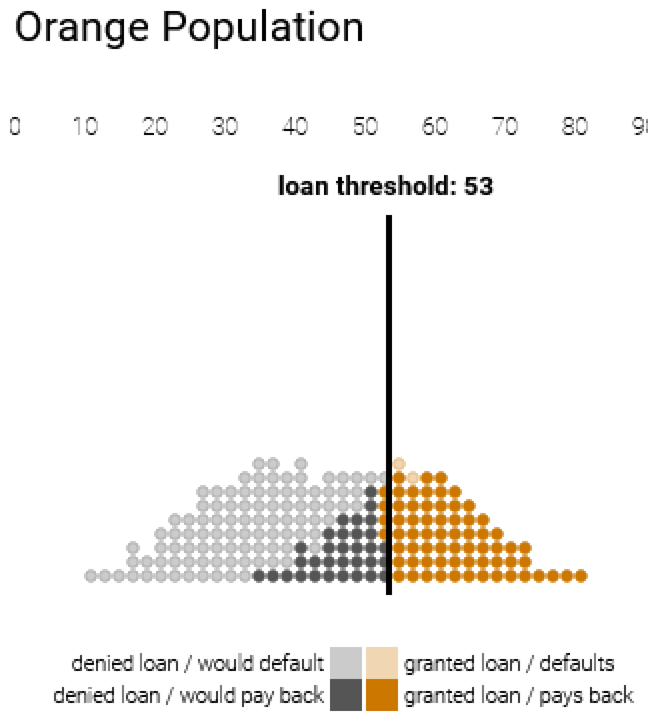


Attacking discrimination with smarter machine

Simulation #3 : <u>Equal opportunity</u> : The constraint is that of the people who can pay back a loan, the same fraction in each group should actually be granted a loan -> the true positive rate is identical between groups

Among people who would pay back a loan, blue and orange groups do equally well.





Political philosophy

See Reuben Binns, "Fairness in Machine Learning: Lessons from Political Philosophy", Proceedings of Machine Learning Research 81:1–11, 2018 Conference on Fairness, Accountability, and Transparency

" 'fairness' as used in the fair machine learning community is best understood as a placeholder term for a variety of normative egalitarian considerations"

Issue : examine how egalitarian norms might provide an account of why and when algorithmic systems can be considered unfair

Political philosophy : utilitarism

- **To satisfy one fairness criteria one must sacrifice some utility**. Ex : minimize the false positive rate of criminal reoffenders (high risk + no reoffending)=> risk of reducing public security level (release of truly high risk inmates) (Narayanan) ; Conversely if utility criteria prevails (e.g. public security) false positive rate are to be kept at a high level
- **Society values differently false positive and false negative** (Abu Elyounes 2019)
- Corbett-Davis, Pierson, Feller Avi, Sharad, « Algorithmic decision making and the cost of fairness », 2017 : utilitarist-inspired analysis of COMPAS -> "**there is tension between reducing racial disparities and improving public safety**". Incompatibility between maximisation of public security and equal treatment of individuals whatever their race. Algorithmic fairness is a problem of constrained optimisation (in reference to diverse fairness metrics : statistical parity, predictive equality, conditional statistical parity). The optimal algorithm that results require applying multiple, race-specific thresholds to individuals' risk scores.

- **Cost-benefit** approach : does the marginal social benefit of additional fairness (e.g. less group discrimination) outweigh the marginal cost ? (see Corbett-Davis & al. 2017).

Political philosophy : egalitarianism (1)

Variants of egalitarianism

Welfarism (Cohen 1989) :

a) *Hedonic welfare* : “welfare as enjoyment, or, more broadly, as a desirable or agreeable state of consciousness”. Limit : metrics of welfare
b) *Welfare as preference satisfaction (or fulfillment)* : “preferences order states of the world, and where a person's preference is satisfied if a state of the world that he prefers obtains, whether or not he knows that it does”. Limit : heterogeneity of preferences and resource needs (if Peter prefers champaign and Allan prefers beer, Peter needs more resources to fulfill his preference than Peter)
c) *Equality of opportunity for welfare* (Richard Arneson).

Resources-based (Dworkin) : a society is just it holds individuals responsible for their decisions and actions, but not for circumstances beyond their control, such as race, sex, skin-color, intelligence, and social position. Unequal distribution of resources is considered fair only when it results exclusively from the decisions and intentional actions of those concerned

Primary social goods (Rawls) : those that the citizens need as free people and as members of the society : civil rights, political rights, liberties, income and wealth, the social bases of self-respect, etc.

Capabilities (Sen) : Capabilities are the doings and beings that people can achieve if they so choose, such as being well-nourished, getting married, being educated, and travelling; functionings are capabilities that have been realized.

Political philosophy : egalitarianism (2)

Implications for AI

(1) « egalitarian norms might provide an account of why and when algorithmic systems can be considered unfair » (Binns, 2018, p. 6)

(2) diversity of egalitarian norms implied in algorithmic decisions

- loan decision, insurance : impact the distribution of ressources (***distributive harm***)
- exclusion from a social network : impact the capabilities or welfare (***representative harm***)

(3) Welfarism : preferences fulfillment => some individuals may prefer a racially-segregated society (requires a moral judgment about which preferences are to taken into account or excluded)

- Rawls : Maximim principle + veil of ignorance : the criteria of social justice requires a social contract which have to be set-up by individuals ignoring their future position (veil of ignorance). A just society benefits the least advantaged (maximin principle).
- Sen : a just society benefits the poorer (strengthening the poorer' capabilities).

PART II

Discrimination

- Forms
- Disparate impact : legal dimension, crossroad between ML and law
- Controversy about COMPAS’ system of prediction of recidivism

Forms

- Direct (intentional) discrimination
- Indirect (unintentional) discrimination : Disparate impact
- Individual versus group discrimination

=> This lecture concentrates on group discrimination/disparate impact

Disparate Impact

$$\mathrm{DI} = \frac{P(Y=1) \mid (S=0)}{P(Y=1) \mid (S=1)}$$

Legal concept :

Civil Rights Act 1964

Title VII : *prohibits employment discrimination based on race, color, religion, sex and national origin*

Title VI : *No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance*.

Age Discrimination in Employment Act, 1967, Fair Housing Act, 1967

Equal Employment Opportunity Commission (EEOC) :

80% Rule (Uniform Guidelines on Employee Selection Procedures 1978) : ratio of selection rates across groups.

Ratio < 0.8 : presumption of discrimination

Case-Law (federal courts) : from a expansion of DI doctrine in the benefit of plaintiffs in the 70's to a more restrictive interpretation (in the benefit of employers) since the 90's

Disparate Impact

Case-law (federal courts, USA). Some major rulings

***Griggs v. Duke Power (Supreme Court,* 1971)** : the Supreme Court made a significant advance in securing civil rights for African Americans. The company in question conducted intelligence tests and required employees to have completed college in order to be promoted to higher paying positions.

***Wards Cove Packing v. Atonis* (*Supreme Court,* 1989)** the Court placed a very important restriction on disparate impact actions by establishing the evidentiary rule that the plaintiff must establish (a) what precisely defined practice or rule caused the indirectly discriminatory impact, and (b) that the employer refused to implement practices or rules that would have satisfied the plaintiff's grievances. In addition, the accused company may argue that the rule or practice that caused the disproportionate impact was justified by the necessity of business.

***Ricci v, DeStefano (Supreme Court,* 2009):** the Mayor of New Haven, Connecticut, cancelled a competition for the promotion of the city's firefighters because the success rate of white firefighters was twice that of African Americans. The court ruled in favor of the successful firefighters; it faulted the Mayor for canceling the competition without showing that its continuation could expose him to disparate impact liability.

Legal concepts & ML concepts (Xiang & Inioluwa Deborah Raji, arXiv: 1912.00761v1 [cs.CY] 25 Nov 2019

	Anti-discrimination Law (American law)	Machine Learning
Procedural fairness	to arrive at just outcomes through iterative processes and the close examination of the set of governance structures in place to guide individual human decision-making Focus on processes & the system surrounding the algorithm and its use	refer to identifying the input features that lead to a particular model outcome, as a proxy for the "process" through which the model makes its prediction Focus on outcomes & specifics of the algorithm itself
Discrimination	Federal laws provide anti-discrimination protections in housing, employment, and other domains. The federal acts primarily define discrimination though motive, evidenced intent of exclusion, and causality, rather than simply outcomes .	Often presented as an unjust correlation between protected class variables and some metric of interest, such as outcomes, false positive rates, or a similarity metric
Protected Class/Sensitive Attribute	less commonly measured attributes can also be considered, such as sexual orientation, pregnancy, and disability status Aware of the implementation of law & possibility of "reversal" of the benefits of anti-discrimination law (Ricci v. DeStefano, 2009)	Protected attributes are presented as recorded or visible traits that should not factor into a decision, such as age, race, or gender. Unaware of the implementation of law
Anti-classification and anti-subordination	Anticlassification (or antidifferentiation principle) : holds that the government may not classify people either overtly or surreptitiously on the basis of a forbidden category such as their race Antisubordination (or equal citizenship, anti-caste) theorists contend that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups (Baldin & Siegel, 2003)	Anticlassification: classifications based on protected class attributes are impermissible. ML fairness community is actually quite familiar with this concept ("fairness through unawareness") Numerous works explicitly presenting anti-classification as a potential fairness objective Antisubordination is rarely called out as a motivation in ML fairness literature
Affirmative action	Landmark affirmative action cases have concluded that schools seeking to increase racial diversity cannot use racial quotas or point systems. Schools have dealt with this conundrum through greater opacity, seeking to be race conscious without making explicit how race factors into admissions decisions	The ML fairness community has articulated a goal of 'fair affirmative action,' which guarantees statistical parity (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible" and understands affirmative action to be cases in which we explicitly take demographics into account
Disparate treatment and disparate impact	Disparate treatment : the key legal question is whether the alleged discriminator's actions were motivated by discriminatory intent Disparate impact: disproportionate outcomes between sub-groups is illegal if intentional . In case of intentionality : liability incurred Key issue : intentionality	Disparate treatment is often explained as making use of the protected attribute in the decision-making process -> avoiding the use of protected class variables in debiasing techniques Disparate impact is understood as when outcomes differ across subgroups (even unintentionally) -> group fairness formulations Algorithm cannot possess intent by itself

Comparison between USA and European Union (statute law and case law)

	United States	E. U.
Main focus	Racial inequalities Workers' hiring and promotion	Salarial equality between men and women
Part-time work	Not taken into account	Taken into account
Burden of proof (from the plaintiff viewpoint)	Restrictive and limiting	Not very demanding
Justification of rules and practices with disparate impact by employers	Business necessity benefit the employers	Business necessity : balanced approach in the EUCJ case-law

Predictive criminal justice (USA): COMPAS

Correctional Offender Management Profiling for Alternative Sanctions, developed by the Northpointe Compagny (now Equivant)

Three-levels scores :

1. Pretrial Release Risk scale : Risk of not appearing in court and/or committing crimes between indictment and criminal sanction

2. ***General Recidivism Risk Scale* – GRRS:** Risk of re-offending after release. The scale takes into account the individual's criminal history and accomplices, drug use, juvenile delinquency…

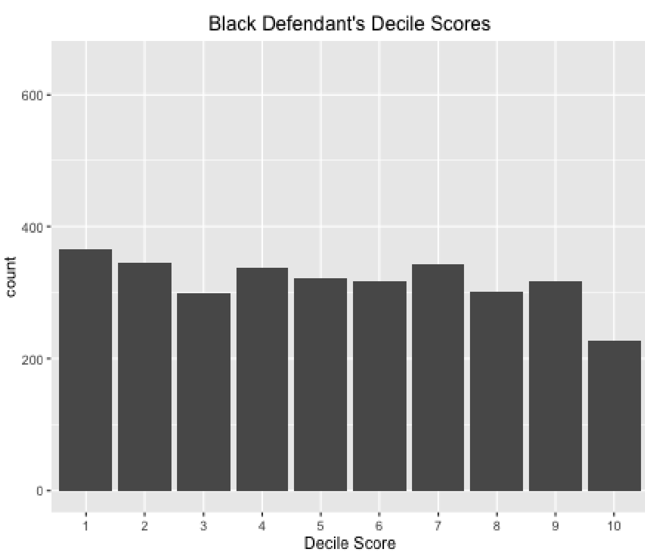
3. ***Violent Recidivism Risk Scale* – VRRS:** Risk of violent recidivism after release. Takes into account: the individual's history of violence, frequency of lawlessness, school problems, age of first arrest...

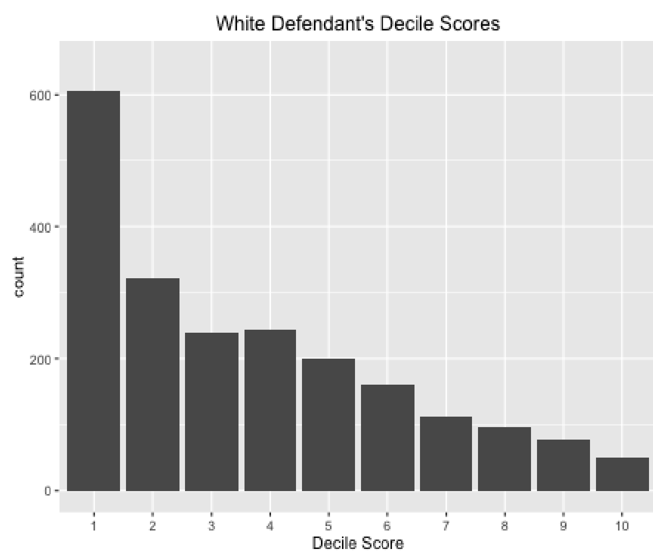
COMPAS : ProPublica critics

Prediction Fails Differently for Black Defendants

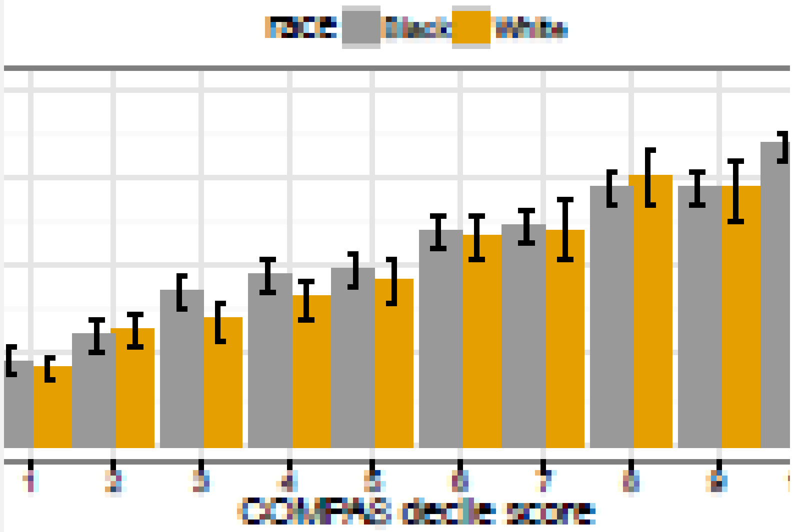
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)





Chouldechova : COMPAS scores are calibrated



Alexandra Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”, ArXiv: 1610.07524v1, 24 oct 2016

See also : Julia Dressel & Hany Farid, “ The accuracy, fairness, and limits of predicting recidivism”, Science Advances, 17 january 2018

COMPAS : not discriminatory ?

Overall accurary equality : The overall accuracy of the COMPAS label is the same, regardless of race

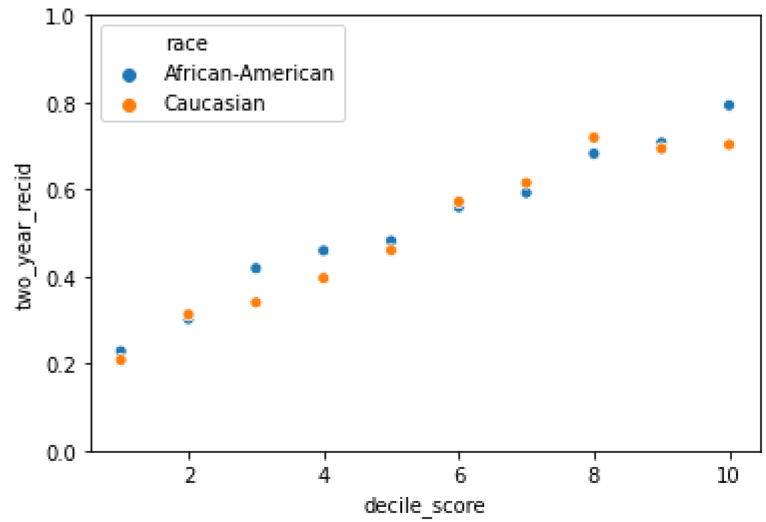
```
race
African-American    0.638258
Caucasian           0.669927
dtype: float64
```

Predictive Positive Value : The likelihood of recidivism among defendants labeled as medium or high risk is similar, regardless of race

```
race
African-American    0.629715
Caucasian           0.591335
Name: two_year_recid, dtype: float64
```

Calibration : For any given COMPAS score, the risk of recidivism is similar, regardless of race

ra
Af
Ca
Na



Farhan Rahman, COMPAS Case Study: Fairness of a Machine Learning Model,, Sep 7, 2020·
https://towardsdatascience.com/compas-case-study-fairness-of-a-machine-learning-model-f0f804108751

PART III

Explicability of algorithmic decisions

Explaining explicability… (1)

1. **What does « explicability mean »**?

- Global vs. Local explicability
- Explicability ex ante vs. ex post
- Technical vs. Decision process

2. **Explicability of what? Dataset, algorithm, model, outcome (decision, prediction)**

3. **Explicability for who? Expert, regulator, individual**

Explainability as a legal obligation?

Is it effective or practicable ?

EU law : GDPR, recital 71 : In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

French law :

- Loi n° 2018-493 : obligation to communicate the rules defining the processing + the main characteristics of its implementation (except if these rules are subject to secrets protected by law)
- Code des relations du public avec l'administration (CRPA, art. L. 311-3-1 : « the rules defining the processing and the main characteristics of its implementation shall be communicated by the Administration to the person concerned on request .
- CRPA, art. R. 311-1-2 : specifies the information to be provided in intelligible form.

Constraints : commercial secret ; black box

A complex algorithm with very good predictive capabilities is not necessarily explainable

- tension between accuracy (high reliability of predictions) and explicability
- Counterfactual explanation?

Explicability of algorithmic decisions

Counterfactual explanation

« *You have been refused credit by the bank. Your annual income is 30,000 euros. If your income had been 40,000 euros, you would have been granted credit* ».

"In the existing literature, "explanation" typically refers to an attempt to convey the internal state or logic of an algorithm that leads to a decision. In contrast, counterfactuals describe a dependency on the external facts that led to that decision"

See Sandra Wachter, Brent Mittelstadt & Chris Russell, COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR, Harvard Journal of Law & Technology 2018

Thanks for your attention