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
- Incompatibilities between fairness metrics
- Political philosophy – utility and fairness / normative egalitarian considerations
Plurality of fairness metrics

See:
**Arvind Narayanan**, “Tutorial: 21 fairness definitions and their politics”, March, 2018
https://www.youtube.com/watch?v=jIXluYdnyyk

<table>
<thead>
<tr>
<th>Notion</th>
<th>Sub-notion</th>
<th>Corresponding Legal Mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual Fairness</td>
<td>The unaware approach</td>
<td>Equal opportunity as colorblindness</td>
</tr>
<tr>
<td></td>
<td>Fairness through awareness</td>
<td>Equal opportunity based on similarities, and levels of scrutiny</td>
</tr>
<tr>
<td>Group fairness</td>
<td>Decoupling</td>
<td>Affirmative action (as separate but equal)</td>
</tr>
<tr>
<td></td>
<td>Statistical or conditional parity</td>
<td>Affirmative action (preferably through critical diversity)</td>
</tr>
<tr>
<td></td>
<td>Equal opportunity</td>
<td>Affirmative action (as equal opportunity)</td>
</tr>
<tr>
<td></td>
<td>Equalized odds</td>
<td>Achieving equality by equalizing the false positive and false negative errors</td>
</tr>
<tr>
<td></td>
<td>Calibration</td>
<td>Achieving equality by statistical significance</td>
</tr>
<tr>
<td></td>
<td>Multicalibration</td>
<td>Achieving equality by statistical significance, and accounting for intersectionality</td>
</tr>
<tr>
<td>Causal Reasoning</td>
<td>Counterfactual fairness</td>
<td>Due process</td>
</tr>
</tbody>
</table>
Sources of Bias (1)
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

<table>
<thead>
<tr>
<th>Predicted condition</th>
<th>True condition</th>
<th>Predicted condition positive</th>
<th>Condition positive</th>
<th>Condition negative</th>
<th>True positive</th>
<th>False positive, Type I error</th>
<th>Positive predictive value (PPV), Precision</th>
<th>False discovery rate (FDR)</th>
<th>False omission rate (FOR)</th>
<th>Negative predictive value (NPV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total population</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prevalence = ( \frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}} )</td>
<td></td>
<td></td>
<td>False discovery rate (FDR) = ( \frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}} )</td>
<td>False omission rate (FOR) = ( \frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}} )</td>
<td>Negative predictive value (NPV) = ( \frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}} )</td>
</tr>
<tr>
<td>Predicted condition positive</td>
<td>True positive</td>
<td>False positive, Type I error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted condition negative</td>
<td>False negative, Type II error</td>
<td>True negative</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Key Metrics**

- **True positive rate (TPR), Recall, Sensitivity, probability of detection, Power**
  \( \text{Power} = \frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}} \)

- **False positive rate (FPR), Fall-out, probability of false alarm**
  \( \text{FPR} = \frac{\Sigma \text{False positive}}{\Sigma \text{Condition positive}} \)

- **Positive likelihood ratio (LR+)**
  \( \text{LR}^+ = \frac{\text{TPR}}{\text{FPR}} \)

- **Negative likelihood ratio (LR−)**
  \( \text{LR}^- = \frac{\text{FNR}}{\text{TNR}} \)

- **Diagnostic odds ratio (DOR)**
  \( \text{DOR} = \frac{\text{LR}^+}{\text{LR}^-} \)

- **F₁ score**
  \( \text{F₁} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \)
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
Fairness metrics: incompatibilities

Tutorial:


**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: Group Unaware - 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: Demographic parity - 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: Equal opportunity: 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


“‘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: utilitarianism

- **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)


- **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 if 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 resources (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: Maximin 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

\[ DI = \frac{P(Y=1) \mid (S=0)}{P(Y=1) \mid (S=1)} \]

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)**:

- 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.
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Procedural fairness</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Discrimination</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Protected Class/Sensitive Attribute</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Affirmative action</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Disparate treatment and disparate impact</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
## Comparison between USA and European Union (statute law and case law)

<table>
<thead>
<tr>
<th></th>
<th>United States</th>
<th>E. U.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main focus</strong></td>
<td>Racial inequalities</td>
<td>Salarial equality between men and women</td>
</tr>
<tr>
<td></td>
<td>Workers’ hiring and promotion</td>
<td></td>
</tr>
<tr>
<td><strong>Part-time work</strong></td>
<td>Not taken into account</td>
<td>Taken into account</td>
</tr>
<tr>
<td><strong>Burden of proof (from the plaintiff viewpoint)</strong></td>
<td>Restrictive and limiting</td>
<td>Not very demanding</td>
</tr>
<tr>
<td><strong>Justification of rules and practices with disparate impact by employers</strong></td>
<td>Business necessity benefit the employers</td>
<td>Business necessity: balanced approach in the EUCJ case-law</td>
</tr>
</tbody>
</table>
Predictive criminal justice (USA): COMPAS

Correctional Offender Management Profiling for Alternative Sanctions, developed by the Northpointe Company (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 – VRSS**: 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**

<table>
<thead>
<tr>
<th></th>
<th>WHITE</th>
<th>AFRICAN AMERICAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Overall, Northpointe’s assessment tool correctly predicts risk from 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually to 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 Howard County, Md.]

---

**Bar Charts: Black and White Defendant’s Decile Scores**

The bar charts display the distribution of decile scores for black and white defendants. The x-axis represents the decile score, ranging from 1 to 10, while the y-axis represents the count of defendants in each decile.
Chouldechova: COMPAS scores are calibrated


See also: Julia Dressel & Hany Farid, “The accuracy, fairness, and limits of predicting recidivism”, Science Advances, 17 January 2018
COMPAS : not discriminatory?

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

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

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

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”

Thanks for your attention