Foundations of machine learning

Fairness and machine learning

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2025

Outline

- Targeted treatment assignment and supervised learning.
- Fairness as predictive parity and taste-based discrimination.
- Limitations of this notion of fairness.
- Alternative notions of fairness / discrimination.
- Social welfare as a unifying framework for many theories of justice.
- The causal impact of algorithms on inequality / social welfare.
- Case study: Predictive incarceration.

Takeaways for this part of class

- Public debate and the computer science literature:
  Fairness of algorithms, understood as the absence of **discrimination**.
- We argue: Leading definitions of fairness have three limitations:
  1. They legitimize inequalities justified by "merit."
  2. They are narrowly bracketed; only consider differences of treatment within the algorithm.
  3. They only consider between-group differences.
- Two alternative perspectives:
  1. What is the causal impact of the introduction of an algorithm on **inequality**?
  2. Who has the **power** to pick the objective function of an algorithm?

Fairness in algorithmic decision making – Setup

- Binary treatment W, treatment return M (heterogeneous), treatment cost c. Decision maker's objective
$$\mu = E[W \cdot (M - c)].$$

- All expectations denote averages across individuals (not uncertainty).

- M is unobserved, but predictable based on features X. For $m(x) = E[M|X = x]$, the optimal policy is
$$w^*(x) = 1(m(X) > c).$$

Examples

- Bail setting for defendants based on predicted recidivism.
- Screening of job candidates based on predicted performance.
- Consumer credit based on predicted repayment.
- Screening of tenants for housing based on predicted payment risk.
- Admission to schools based on standardized tests.

Definitions of fairness

- Most definitions depend on **three ingredients**.
  1. Treatment W (job, credit, incarceration, school admission).
  2. A notion of merit M (marginal product, credit default, recidivism, test performance).
  3. Protected categories A (ethnicity, gender).
- I will focus initially on the following **definition of fairness**:

$$\pi = E[M|W=1, A=1] - E[M|W=1, A=0] = 0$$

"Average merit, among the treated, does not vary across the groups a."

This is called "predictive parity" in machine learning,
the "hit rate test" for "taste based discrimination" in economics.

- "Fairness in machine learning" literature: **Constrained optimization**.

$$w^*(\cdot) = \underset{w(\cdot)}{\text{argmax}}\, E[w(X) \cdot (m(X) - c)] \qquad \text{subject to} \qquad \pi = 0.$$

Fairness and $\mathscr{D}$'s objective

Observation

Suppose that W, M are binary ("classification"), and that

1. $m(X) = M$ (perfect predictability), and
2. $w^*(x) = 1(m(X) > c)$ (unconstrained maximization of $\mathscr{D}$'s objective μ).

Then $w^*(x)$ satisfies predictive parity, i.e., $\pi = 0$.

In words:

- If $\mathscr{D}$ is a firm that is maximizing profits and observes everything then their decisions are fair by assumption.
  - No matter how unequal the resulting outcomes within and across groups.
- Only deviations from profit-maximization are "unfair."

Three normative limitations of “fairness” as predictive parity

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

Three normative limitations of “fairness” as predictive parity

1. They legitimize and perpetuate **inequalities justified by “merit.”**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

Three normative limitations of "fairness" as predictive parity

1. They legitimize and perpetuate **inequalities justified by "merit."**
   Where does inequality in M come from?

2. They are **narrowly bracketed**.
   Inequality in W in the algorithm,
   instead of some outcomes Y in a wider population.

3. Fairness-based perspectives **focus on categories** (protected groups)
   and ignore within-group inequality.

Alternative measures of fairness (1)

Measures that share the same limitations:

- Equality of true positives:

$$E[W|M=1,A=1]-E[W|M=1,A=0].$$

- Equality of false positives:

$$E[W|M=0,A=1]-E[W|M=0,A=0].$$

- Balance for the negative class:

$$E[M|W=0,A=1]-E[M|W=0,A=0]$$

(Like predictive parity, but for $W=0$.)

Alternative measures of fairness (2)

Measures which share only some of these limitations:

- Disparate impact and demographic parity:

$$\frac{E[W|A=1]}{E[W|A=0]}, \qquad E[W|A=1]-E[W|A=0].$$

- Conditional statistical parity:

$$E[W|A=1,X'=x']-E[W|A=0,X'=x']$$

  for a subset of features X' considered "legitimate" sources of inequality. (Cf. Oaxaca-Blinder decompositions.)

- Individual fairness:

$$E[W|X=x_i]-E[W|X=x_j] \text{ for } d(i,j)\approx 0,$$

  for a measure of distance $d(i,j)$ between individuals.

Practice problem

- Which of these measures of fairness do you find more or less appealing?
- Why? For which contexts or applications?

Fairness and discrimination

Inequality and social welfare

Case study

References

Social welfare

- The framework of fairness / bias / discrimination contrasts with perspectives focused on *consequences for social welfare*.
- Common presumption for most theories of justice:

  Normative statements about society
  are based on statements about individual welfare
- Formally:
  - Individuals $i = 1, \ldots, n$
  - Individual i's welfare Y_i
  - Social welfare as function of individuals' welfare

$$SWF = F(Y_1, \ldots, Y_n).$$

Practice problem

- **Who is to be included** among $i = 1, \ldots, n$?
  - All citizens? All residents? All humans on earth?
  - Future generations? Animals?
- **How to measure individual welfare** Y_i?
  - Opportunities or outcomes?
  - Utility? Resources? Capabilities?
- **How to aggregate** to SWF?
  How much do we care about
  - Trevon vs. Emily, Sophie vs. José?
  - Millionaires vs. homeless people?
  - Sick vs. healthy people?
  - Groups that were victims of historic injustice?

The impact on inequality or welfare as an alternative to fairness

- Outcomes are determined by the **potential outcome equation**

$$Y = W \cdot Y^1 + (1 - W) \cdot Y^0.$$

- The **realized outcome** distribution is given by

$$p_{Y,X}(y,x) = \left[p_{Y^0|X}(y,x) + w(x) \cdot \left(p_{Y^1|X}(y,x) - p_{Y^0|X}(y,x) \right) \right] \cdot p_X(x).$$

- What is the impact of $w(\cdot)$ on a **statistic** ν?

$$\nu = \nu(p_{Y,X}).$$

  Examples: Variance, quantiles, between group inequality.

- Cf. Distributional decompositions in labor economics!

When fairness and equality are in conflict

- Fairness is about **treating** people of the same "**merit**" independently of their **group** membership.
- Equality is about the (counterfactual / causal) **consequences** of an algorithm for the distribution of **welfare** of different **people**.

Examples when they are in conflict:

1. Increased surveillance / **better prediction** algorithms:
   Lead to treatments more aligned with "merit"
   Good for fairness, bad for equality.
2. Affirmative action / **compensatory interventions** for pre-existing inequalities:
   Bad for fairness, good for equality.

Influence function approximation of the statistic ν

$$\nu(p_{Y,X}) - \nu(p^*_{Y,X}) = E[IF(Y,X)] + o(\|p_{Y,X} - p^*_{Y,X}\|).$$

- $IF(Y,X)$ is the influence function of $\nu(p_{Y,X})$.

  Formally: The Riesz representer of the Fréchet derivative of ν.

- The expectation averages over the distribution $p_{Y,X}$.

The impact of marginal policy changes on profits, fairness, and inequality

Proposition

Consider a family of assignment policies $w(x) = w^*(x) + \varepsilon \cdot dw(x)$. Then

$$\partial_\varepsilon \mu = E[dw(X) \cdot l(X)], \qquad \partial_\varepsilon \pi = E\left[dw(X) \cdot p(X)\right], \qquad \partial_\varepsilon \nu = E[dw(X) \cdot n(X)],$$

The impact of marginal policy changes on profits, fairness, and inequality

Proposition

Consider a family of assignment policies $w(x) = w^*(x) + \varepsilon \cdot dw(x)$. Then

$$\partial_\varepsilon \mu = E[dw(X) \cdot l(X)], \qquad \partial_\varepsilon \pi = E\left[dw(X) \cdot p(X)\right], \qquad \partial_\varepsilon \nu = E[dw(X) \cdot n(X)],$$

where

$$\begin{aligned}
l(X) &= E[M|X = x] - c, \\
p(X) &= E\left[(M - E[M|W = 1, A = 1]) \cdot \frac{A}{E[WA]}\right. \\
&- \left.\qquad (M - E[M|W = 1, A = 0]) \cdot \frac{(1-A)}{E[W(1-A)]}\right|X = x\Bigg], \\
n(x) &= E\left[IF(Y^1, x) - IF(Y^0, x)|X = x\right].
\end{aligned}$$

Uses of the proposition

1. Elucidate the **tension** between objectives.
   - Profits vs. fairness vs. equality vs. welfare?
   - Suppose $\pi < 0$, $n(x) > 0$ is positive, while $p(x) < 0$.
     Then increasing $w(x)$ is good for welfare and bad for fairness.
   - $\Rightarrow$ Characterizes which parts of the feature space drive
     the tension between alternative objectives.

Uses of the proposition

1. Elucidate the **tension** between objectives.
   - Profits vs. fairness vs. equality vs. welfare?
   - Suppose $\pi < 0$, $n(x) > 0$ is positive, while $p(x) < 0$.
     Then increasing $w(x)$ is good for welfare and bad for fairness.
   - $\Rightarrow$ Characterizes which parts of the feature space drive
     the tension between alternative objectives.
2. Solve for **optimal assignment** subject to constraints.
   - E.g. maximize μ subject to $\pi = 0$.
   - Then $w(x) = 1(l(x) > \lambda p(x))$.

Uses of the proposition 1, continued

3. **Power and inverse welfare weights**
   - For a given $w(\cdot)$, what objective is implicitly maximized?
   - What are the weights for different individuals that rationalize $w(\cdot)$?

Uses of the proposition 1, continued

3. **Power and inverse welfare weights**
   - For a given $w(\cdot)$, what objective is implicitly maximized?
   - What are the weights for different individuals that rationalize $w(\cdot)$?
4. **Algorithmic auditing**.
   - Similar to distributional decompositions in labor economics.
   - Cf. Fortin and Lemieux (1997); Firpo et al. (2009).

Power

- Both fairness and equality are about differences between people who are **being treated**.
- Elephant in the room:
  - Who is on the **other side** of the algorithm?
  - Who gets to be the decision maker $\mathscr{D}$ – who gets to pick the objective function μ?
- Political economy perspective:
  - **Ownership of the means of prediction**.
  - Data and algorithms.

Case study

- Compas risk score data for recidivism.
- From Pro-Publica's reporting on algorithmic discrimination in sentencing.

Mapping our setup to these data:

- A: race (Black or White),
- W: risk score exceeding 4,
- M: recidivism within two years,
- Y: jail time,
- X: race, sex, age, juvenile counts of misdemeanors, fellonies, and other infractions, general prior counts, as well as charge degree.

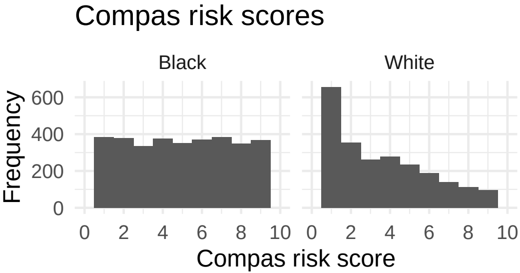
Counterfactual scenarios

Compare three scenarios:

1. "Affirmative action:" Adjust risk scores ± 1, depending on race.
2. Status quo.
3. Perfect predictability: Scores equal 10 or 1, depending on recidivism in 2 years.

For each: Impute counterfactual

- W: Counterfactual score bigger than 4.
- Y: Based on a causal-forest estimate of the impact on Y of risk scores, conditional on the covariates in X.
- This relies on the assumption of conditional exogeneity of risk-scores given X. Not credible, but useful for illustration.


Compas risk scores
Black
White
Frequency
600
400
200
0
0 2 4 6 8 10
Compas risk score

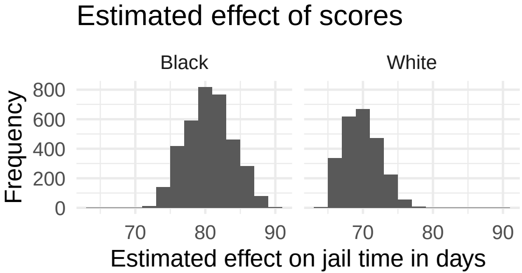

Estimated effect of scores
Black
White
Frequency
800
600
400
200
0
70 80 90
Estimated effect on jail time in days

Table: Counterfactual scenarios, by group

	Black			White		
Scenario	(Score>4)	Recid|(Score>4)	Jail time	(Score>4)	Recid|(Score>4)	Jail time
Aff. Action	0.49	0.67	49.12	0.47	0.55	36.90
Status quo	0.59	0.64	52.97	0.35	0.60	29.47
Perfect predict.	0.52	1.00	65.86	0.40	1.00	42.85

Table: Counterfactual scenarios, outcomes for all

Scenario	Score>4	Jail time	IQR jail time	SD log jail time
Aff. Action	0.48	44.23	23.8	1.81
Status quo	0.49	43.56	25.0	1.89
Perfect predict.	0.48	56.65	59.9	2.10

References

- *Pessach, D. and Shmueli, E. (2020). Algorithmic fairness.* arXiv preprint arXiv:2001.09784
- *Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision making.* ACM Conference on Fairness, Accountability, and Transparency.
- *Kasy, M. (2016). Empirical research on economic inequality.* `http://inequalityresearch.net/`
- *Roemer, J. E. (1998).* Theories of distributive justice. *Harvard University Press, Cambridge.*