MAI4CAREU | Master programmes in Artificial Intelligence 4 Careers in Europe

European University Institute | DEPARTMENT OF LAW

www.eui.eu

Fairness in algorithmic decision making

Francesca Lagioia

Giovanni Sartor

European University Institute

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423

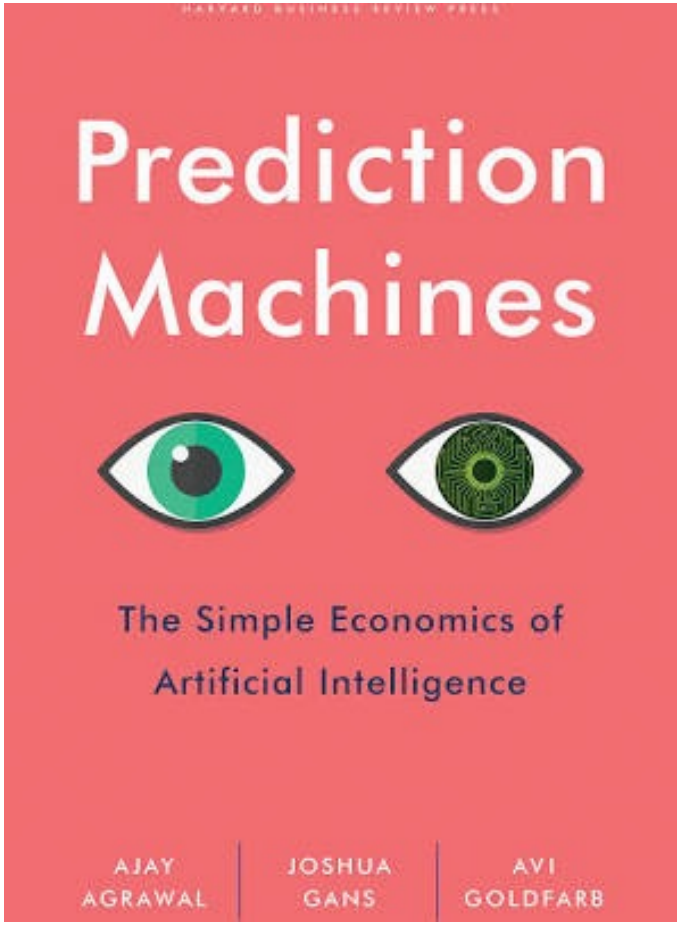
Outline

- AI in decision making concerning individuals
  - Possible causes of unfairness
- The principle of Fairness and its substantive dimension
- AI unfairness
  - The COMPAS predictive system and the Loomis case
  - A toy example and the criteria for assessing fairness

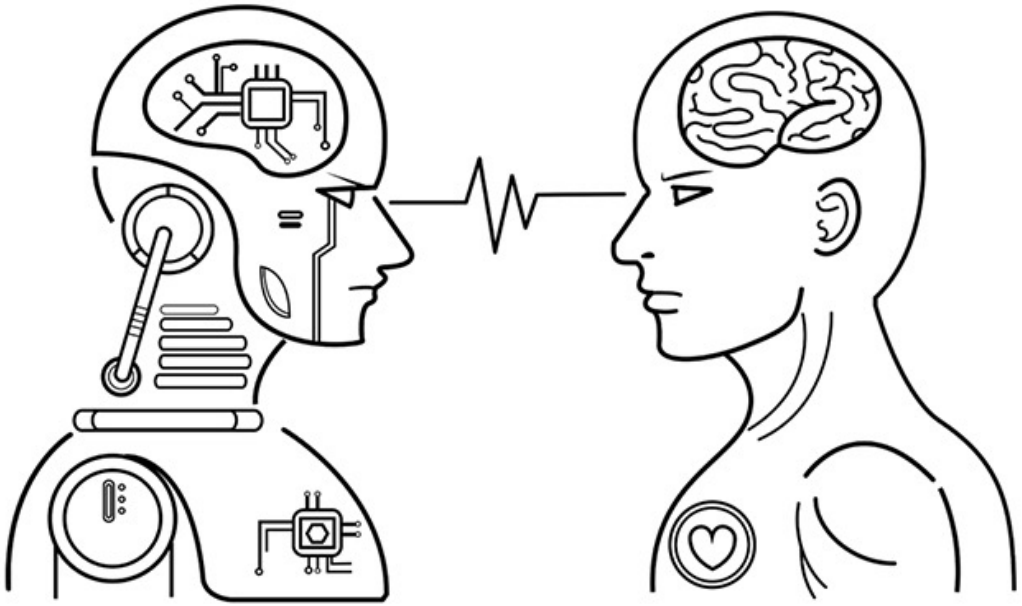
European University Institute
DEPARTMENT OF LAW

AI in decision making concerning individuals: fairness and discrimination

- The combination of AI and Big Data enables automated decision-making even in domains that require complex choices, based on multiple factors, and on non-predefined criteria.
- In recent years, a wide debate has taken place on prospects and risks of algorithmic assessments and decisions concerning individuals



Are AI systems better then humans in assessing us?



In many domains automated predictions and decisions are not only **cheaper**, but also **more precise and impartial** than human ones.

- AI can **avoid** typical **fallacies of human psychology** (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc,), and the widespread human **inability to process statistical data,** as well as **typical human prejudice** (concerning, e.g., ethnicity, gender, or social background).
- In many assessments and decisions —on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism— algorithmic systems have **often performed better**, according to usual standards, than human experts.

Or not?

Others have underscored the possibility that algorithmic decisions may be **mistaken or discriminatory.**

- Only in rare cases will algorithms engage in explicit unlawful discrimination, so-called **disparate treatment**, basing their outcomes on prohibited features (predictors) such as race, ethnicity or gender.
- More often a system's outcome will be discriminatory due to its **disparate impact**, i.e., since it disproportionately affects certain groups, without an acceptable rationale

Systems reproducing the strengths and weaknesses of humans in making judgments

Systems based on **supervised learning** may be trained on **past human judgements** and may therefore **reproduce** the strengths and weaknesses of the humans who made these judgements, including their **propensities to error and prejudice.**

- For example, a recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work. If past decisions were influenced by prejudice, the system will reproduce the same logic.

Prejudice in the training set

Prejudice baked into training sets may persist even if the inputs (the predictors) to automated systems do not include forbidden discriminatory features (e.g. ethnicity or gender.)

This may happen whenever **a correlation exists between discriminatory features and some predictors**

- Assume, for instance, that a prejudiced human resources manager did not hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods. A training set of decisions by that manager will teach the systems not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity. (Kleinberg et all (2019)).

Systems biased against groups

In other cases, a training set may be biased against a certain group, since the achievement of the outcome being predicted (e.g., job performance) is approximated through a **proxy** that has a disparate impact on that group.

- Assume, for instance, that the **future performance** of employees (the target of interest in job hiring) is only measured by the **number of hours worked in the office.** This outcome criterion will lead to past hiring of women —who usually work for fewer hours than men, having to cope with family burdens— being considered less successful than the hiring of men; based on this correlation (as measured on the basis of the biased proxy), the systems will predict a poorer performance of female applicants.

System's biases embedded in the predictors

In other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors.

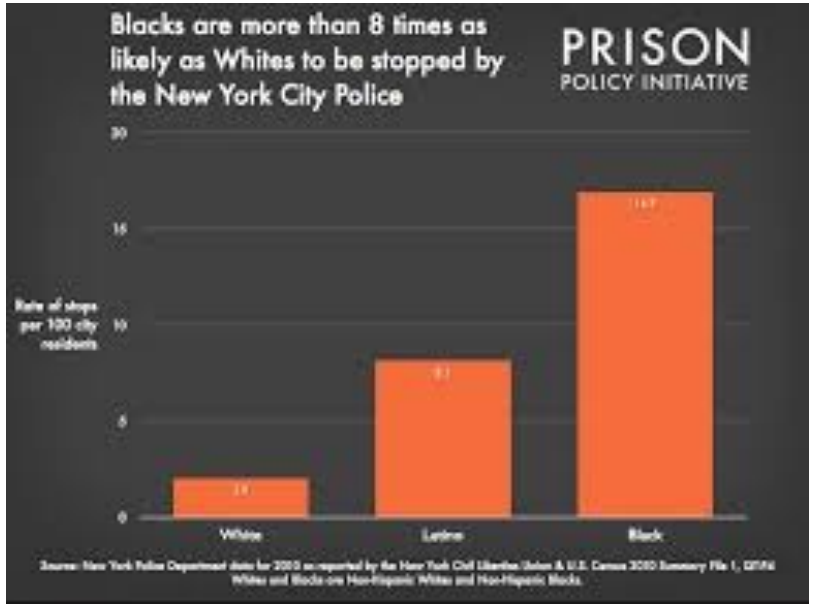
A system may perform unfairly, since it uses a favourable predictor (input feature) that only applies to members of a certain group (e.g., the fact of having attended a socially selective high-education institution).

Unfairness may also result from taking biased human judgements as predictors (e.g., recommendation letters).

Data set that does NOT reflect the statistical composition of the population

Finally, unfairness may derive from **a data set that does reflect the statistical composition of the population.**

- Assume for instance that in applications for bail or parole, previous criminal record plays a role, and that members of a certain groups are subject to stricter controls, so that their criminal activity is more often detected and acted upon. This would entail that members of that group will generally receive a less favourable assessment than members of other groups having behaved in the same ways.



- Members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set,
- This will reduce the accuracy of predictions for that group (e.g., consider the case of a firm that has appointed few women in the past and which uses its records of past hiring as its training set).

Challenging the unfairness of automated decision-making

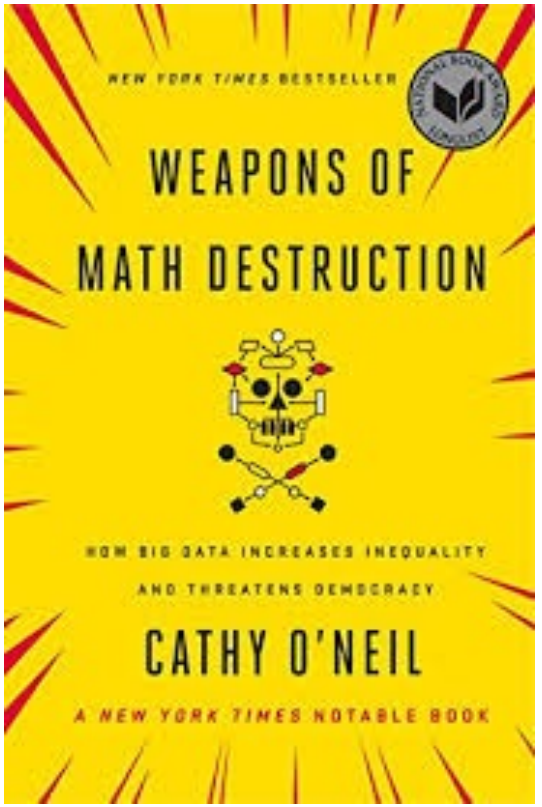
It has been observed that it is difficult to challenge the unfairness of automated decision-making.

Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operation, giving rise to additional **costs and uncertainties**.

In fact, predictions of machine-learning systems are based on **statistical correlations**, **against which it may be difficult to argue** on the basis of individual circumstances, even when exceptions would be justified.

Weapons of math distruction

"An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone's life upside down. And yet when the person fights back, "suggestive" countervailing evidence simply won't cut it. The case must be ironclad. The human victims of WMDs, we'll see time and again, are held to a far higher standard of evidence than the algorithms themselves". (O'Neil (2016))



Or not?

[W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central **trade-offs among competing values.** Algorithms are not only a threat to be regulated; with the right safeguards in place, they have **the potential to be a positive force for equity**

(Kleinberg, Ludwig, Mullainathan, e Sunstein (2018, 113)).

Challenging the unfairness of automated decision-making

These criticisms have been countered by observing that **algorithmic systems,** even when based on machine learning, are **more controllable** than human decision-makers, their **faults can be identified** with precision, and **they can be improved** and **engineered to prevent unfair outcomes**.

Should we exclude the use of automated decision-making?

It seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making.

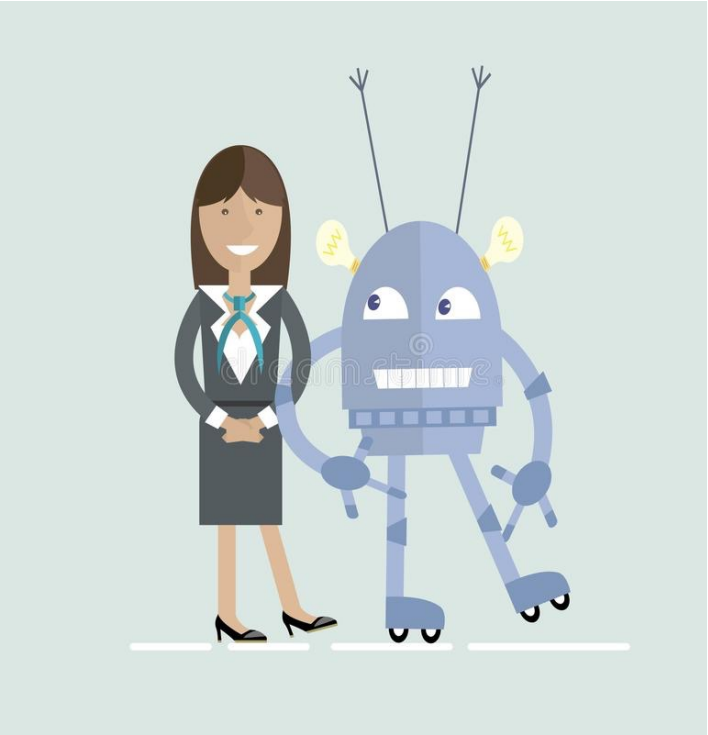
The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker.

Humans + Algorithms?

In many cases, the best solution consists in **integrating human and automated judgements**, by enabling the affected individuals to request a **human review** of an automated decision as well as by favouring **transparency** and developing methods and technologies that enable human experts to analyse and review automated decision-making.

In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, e.g., medical diagnosis, financial investment, granting of loans, etc.

The future challenge will consist in finding the best combination between human and AI, taking into account the capacities and the limitations of both.



Substantive Fairness and AI

- Equal and just distribution of **benefits** and **costs**
- Individuals and groups free from unfair **bias**, **discrimination** and **stigmatisation**
- AI decision making: informational fairness + content fairness of inferences/decision (avoid prejudice, discrimination, etc.)
  - **appropriate mathematical or statistical procedures** for profiling,
  - **technical and organisational measures** to ensure correctness of personal data
  - **secure personal data** (potential risks, discriminatory effects, etc.)

The COMPAS system: AI and unfairness

- An actuarial risk assessment instrument to determine:
  - Risk of recidivism and appropriate correctional treatment
- Based on statistical algorithms
- Offenders are classified in three categories: high, medium, low risk
  - Multiple-choice test (137 questions)
  - Static risk variables (e.g., prior criminal history, education, etc.)
  - Dynamic risk variables (e.g., drug abuse, employment status)

The Loomis case

- In 2013 E. Loomis was charged with driving a stolen vehicle and fleeing from police
- The Distrectual Court ordered a presentencing investigation that included the COMPAS risk assessment
- Loomis was classified at high risk for recidivism and sentenced to 6 years imprisonment
- The decision was appealed by Loomis for violation of due process rights (e.g., basic rights of defence):

- ❑ COMPAS functioning is unknown
- ❑ Its validity can not be verified
- ❑ It discriminates on gender and race
- ❑ Statistical-based predictions violate the right to individualized decision.

European University Institute
DEPARTMENT OF LAW

The Loomis case

In 2016 the Supreme Court of Wisconsin rejected all defendant's arguments.

According to the Supreme Court:

- Statistical algorithms **does not violate the right to individualized decisions**
- They should be used to "**enhance a judge's evaluation** of other evidence in the formulation of an individualized sentencing
- **Prohibition to base decisions solely on risk scores + obligation to motivate** as **safeguards** of the defendant' rights.
- Considering **gender is necessary to achieve statistical accuracy**.
- Judges should be informed on the debate concerning COMPAS race discrimination

The challenges

In 2016 ProPublica published a study (Larson et al. 2016):

Sample: 11,757 defendants assessed by COMPAS (2013-2014)

Objective: evaluate COMPAS accuracy and fairness

Methodology: Comparison between predicted recidivism rates and the rate that actually occurred over 2-year period.

The challenges

ProPublica **Results:**

- Moderate-Low Predictive accuracy (61.2%)
- Black defendants were predicted at a higher risk than they actually were. Probability of high-risk misclassification (45% blacks vs. 23% whites)
- White defendants were often predicted to be less risky than they were. Probability of low-risk misclassification (48% whites vs. 28% blacks).

European University Institute
DEPARTMENT OF LAW

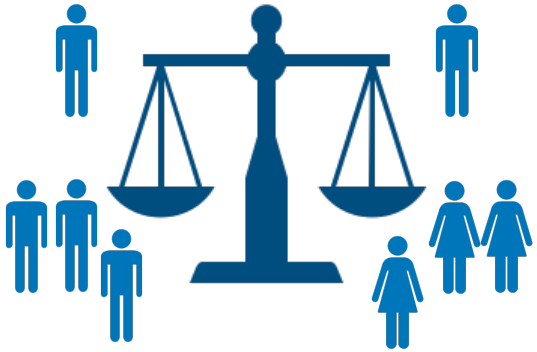
The rebuttals

According to Northpoint (Dieterich et al 2016) ProPublica made several statistical and technical errors

- The accuracy of COMPAS predictions > accuracy of human judgments
- General Recidivism Risk Scale is equally accurate for blacks and whites
- COMPAS is compliant with the principle of fairness
- It does not implement racial discrimination

The debate: Is COMPAS fair?

- Is it accurate?
- Is it fair to individuals?
- Is it fair to groups?



European University Institute
DEPARTMENT OF LAW

The case of SAPMOC

- 2000 defendants
  - 1000 blues
  - 1000 greens
- A single predictor:
  - If previous offences then probably recidivate
- Assumption 1
  - previous offenders: 75% recidivate
  - fist-time offenders: 25% recidivate
- Assumption 2
  - Blue: 75% previous offenders
  - Green 25% previous offenders

European University Institute
DEPARTMENT OF LAW

SAPMOC Assumptions

Real Outcomes			
	Recidivism	No Recidivism	Total
Previous Offence	750	250	1000
No Previous Offence	250	750	1000

SAPMOC Predictions			
	Recidivism	No Recidivism	Total
Previous Offence	1000	0	1000
No Previous Offence	0	1000	1000

Base Rate	Positives	Negatives
	(TP+FN)/(TP+FN+FP+TN)	(TN+FP)/(TP+FN+FP+TN)
Blue	62.5%	37.5%
Green	37.5%	62.5%

	Positives	True Positives	False Positives	Negatives	True Negatives	False Negatives
	(TP+FP)	(TP)	(FP)	(TN+FN)	(TN)	(FN)
Blu	750	562.5	187.5	250	187.5	62.5
Green	250	187.5	62.5	750	562.5	187.5

SAPMOC Accuracy

Accuracy	
(TP+TN)/(TP+FP+TN+FN)	
Blue	75,0%
Green	75,0%

SAPMOC FAIRNESS

- Statistical Parity
- Equality of Opportunity
- Calibration
- Conditional Use Error
- Treatment Equality

Statistical parity

- Each group should have an equal proportion of positives and negatives predictions

Statistical Parity	Positives	Negatives
	(TP+FP)/(TP+FP+TN+FN)	(TN+FN)/(TP+FP+TN+FN)
Blu	75,00%	25,00%
Green	25,00%	75,00%

Equality of opportunity

- The members of each group, which share the same features, should be treated equally in equal proportion.

Equality of opportunity	Positives	Negatives
	TP/(TP+FN)	TN/(TN+FP)
Blu	90,0%	50,0%
Green	50,0%	90,0%

Calibration

- The proportion of correct predictions should be equal within each group and with regard to each class.

Calibration	Positives	Negatives
	TP/(TP+FP)	TN/(TN+FN)
Blu	75,0%	75,0%
Green	75,0%	75,0%

Conditional use error

- The proportion between FP (FN) and the total amount of positive (negatives) predictions should be equal for the 2 groups.

False rate	Positives	Negatives
	FP/(TP+FP)	FN/(TN+FN)
Blu	25,0%	25,0%
Green	25,0%	25,0%

Treatment equality

- The ratio between errors in positive and negative predictions should be equal in all groups. .

Treatment Equality	Positives	Negatives
	FP/FN	FN/FP
Blu	300,0%	33,3%
Green	33,3%	300,0%

What about SAPMOC/COMPAS?

- Equal accuracy within groups
- Different base rate explains the violation of statistical parity, treatment equality, and equality of opportunities
- Violation of fairness criteria does not necessarily lead to unfairness
- Shall we impose statistical parity? (Lower accuracy + higher false rate + discrimination against individuals)
- Individuals fairness vs group fairness

Consideration on the Fairness in automated decision making

- Unpacking the decision
  - Unfairness in prediction (prohibited features, biased data set, biased proxy, etc.)
  - Unfairness in classification (threshold – affirmative actions)
  - Unfairness in decision (right/values optimization)
- Predictive systems as instruments to understand the reality

European University Institute
DEPARTMENT OF LAW

Looking to the future

- AI is too often perceived as a source of threats and Law is too often seen as difficult and sometimes even inaccessible for citizens
- The combination of AI and Law could be the key to protect citizens and make the Law accessible to the wider public

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423