Data Science for Public Policy

AI and Fairness

Dr. Sergio Galletta

ETHZ Zurich

02/05/2024

AI can support more efficient decisions. But...

AI can support more efficient decisions. But...

What if the human generated data were biased?
All human decisions are biased to some degree.

AI can support more efficient decisions. But...

What if the human generated data were biased?
All human decisions are biased to some degree.

$\downarrow$

AI models are trained to replicate human decisions.

AI can support more efficient decisions. But...

What if the human generated data were biased?
All human decisions are biased to some degree.

$\downarrow$

AI models are trained to replicate human decisions.

$\downarrow$

AI models are biased?

Data can be biased

- Education:
  - Teachers (grading essays) might be biased against some groups of students
    $\rightarrow$ so will essay grading software based on teacher's grades.

Data can be biased

- Education:
  - Teachers (grading essays) might be biased against some groups of students
    $\rightarrow$ so will essay grading software based on teacher's grades.
- Medicine:
  - medical decisions (e.g. diagnosis/treatment) might be biased against some groups of patients
    $\rightarrow$ so will a medical robot trained to make those decisions.

Data can be biased

- Education:
  - Teachers (grading essays) might be biased against some groups of students
    $\rightarrow$ so will essay grading software based on teacher's grades.
- Medicine:
  - medical decisions (e.g. diagnosis/treatment) might be biased against some groups of patients
    $\rightarrow$ so will a medical robot trained to make those decisions.
- Criminal risk scoring:
  - Blacks and whites who are otherwise identical are treated the same;

Data can be biased

- Education:
  - Teachers (grading essays) might be biased against some groups of students
    $\rightarrow$ so will essay grading software based on teacher's grades.
- Medicine:
  - medical decisions (e.g. diagnosis/treatment) might be biased against some groups of patients
    $\rightarrow$ so will a medical robot trained to make those decisions.
- Criminal risk scoring:
  - Blacks and whites who are otherwise identical are treated the same;
  - But blacks tend to be rated as more risky:
    - longer criminal histories **produced by biased system** (Skeem and Lovenkamp 2016).
    - recidivism is measured as "is re-arrested"; blacks **more likely to be re-arrested due to policing bias**.

Bias in AI Systems

There are no perfect solutions to the problem of bias in AI systems.

Bias in AI Systems

There are no perfect solutions to the problem of bias in AI systems.

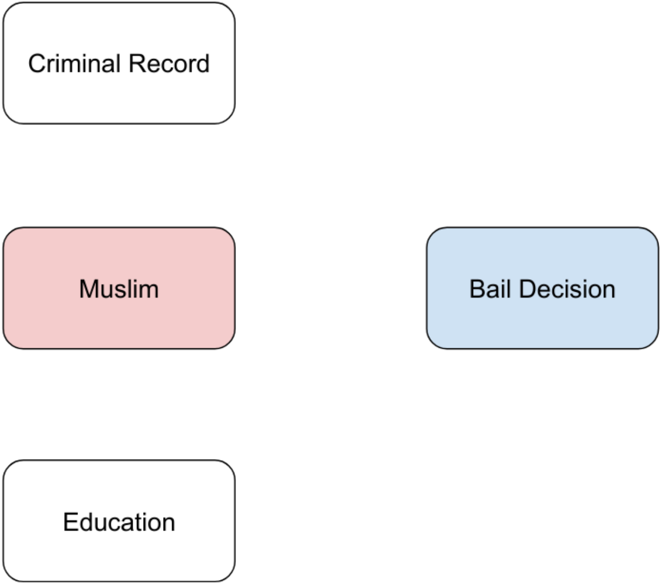
But the baseline is not a perfect AI model – the relevant comparison is a biased human decision.

- When human are biased, it is difficult to detect.
- A major benefit of using algorithms in decision-making $\rightarrow$ we can more easily detect when the system is biased toward some groups
- Further, AI systems can be used to detect bias among human decision-makers.

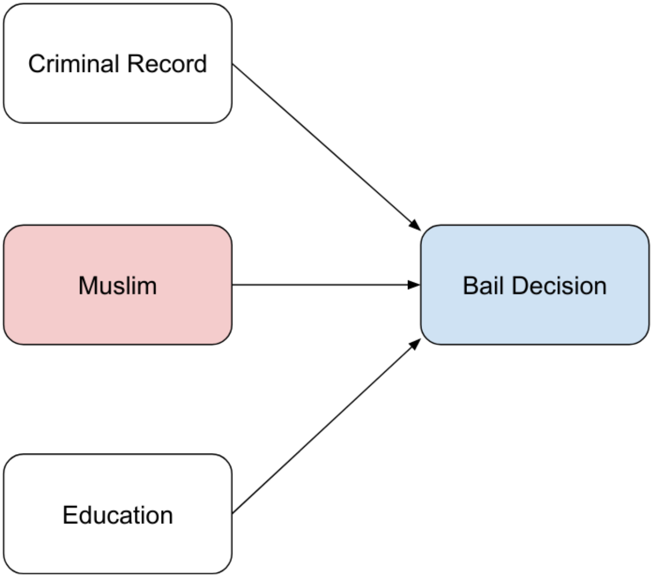
Making algorithms fair

- What type of information should be allowed to be included in the model?
- Can we solve fairness issue by simply refusing to allow models to access critical features (e.g., race or gender)?
- Fairness in the input vs. fairness in the output

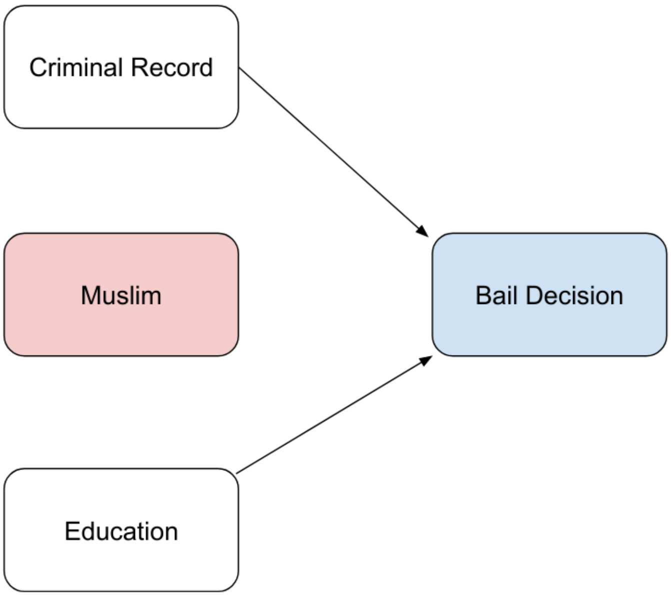
Making algorithms fair



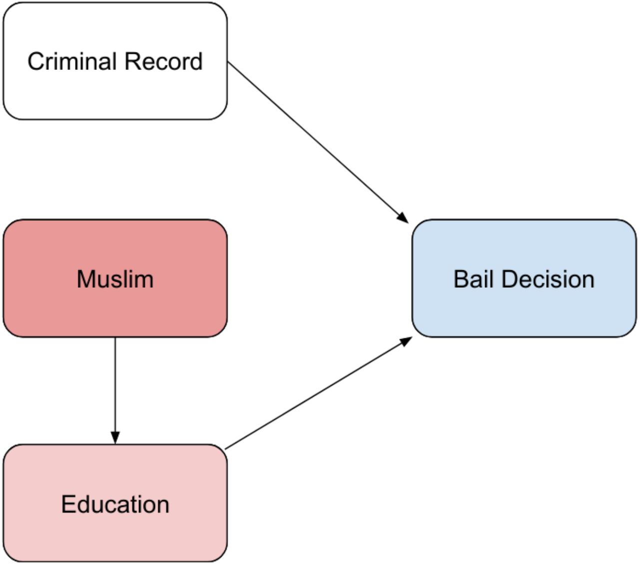
Standard Approach: Use All Data



“Fairness through Unawareness”



Problem: Indirect Discrimination



Making algorithms fair

- It has become virtually impossible to enforce fairness by trying to restrict the inputs given to the machine
- There are too many ways in which the model will find and use proxies for the forbidden information
- If someone knows what kind of car you drive, what kind of phone and computer you own and a few of your favorite websites, they might correctly predict gender, race, income...

Making algorithms fair

- It has become virtually impossible to enforce fairness by trying to restrict the inputs given to the machine
- There are too many ways in which the model will find and use proxies for the forbidden information
- If someone knows what kind of car you drive, what kind of phone and computer you own and a few of your favorite websites, they might correctly predict gender, race, income...
- Simulations from Gillis and Spiess (2019) suggest that:
  - Excluding sensitive features from training doesn't necessarily reduce disparities due to correlations with other features.
  - Including forbidden characteristics may decrease disparity, especially when there's measurement bias in the data – e.g., credit scoring, gender feature and income.

Statistical parity

- One can focus on **fairness in the output**, i.e., look at fairness relative to the actual decisions or predictions a model makes

Statistical parity

- One can focus on **fairness in the output**, i.e., look at fairness relative to the actual decisions or predictions a model makes
- The simplest notion of fairness applied to predictions or decisions of a model is known as **statistical parity**

Statistical parity

- One can focus on **fairness in the output**, i.e., look at fairness relative to the actual decisions or predictions a model makes
- The simplest notion of fairness applied to predictions or decisions of a model is known as **statistical parity**
- First, one need to identify what group of individuals we want to protect based on an attribute e.g,. sex, race...

Statistical parity

- One can focus on **fairness in the output**, i.e., look at fairness relative to the actual decisions or predictions a model makes
- The simplest notion of fairness applied to predictions or decisions of a model is known as **statistical parity**
- First, one need to identify what group of individuals we want to protect based on an attribute e.g,. sex, race...
- Statistical parity requires that across the different groups the same fraction of individuals will face a certain action
    - E.g., we are concerned about discrimination against Black people in the granting of loans by a lender - statistical parity asks that the fraction of Black applicants that are granted loans be nearly the same as the fraction of Withe applicants that are granted loans.

Statistical parity

Statistical parity has limitations

Statistical parity

Statistical parity has limitations

- First, there statistical parity could be achieved with random selection (no need to know individual characteristics)
  - Yet, we can improve on that results by having an AI model that is constrained to take certain action with an equal rates across the different groups
  - Random selection could be good in an *exploration* phase

Statistical parity

Statistical parity has limitations

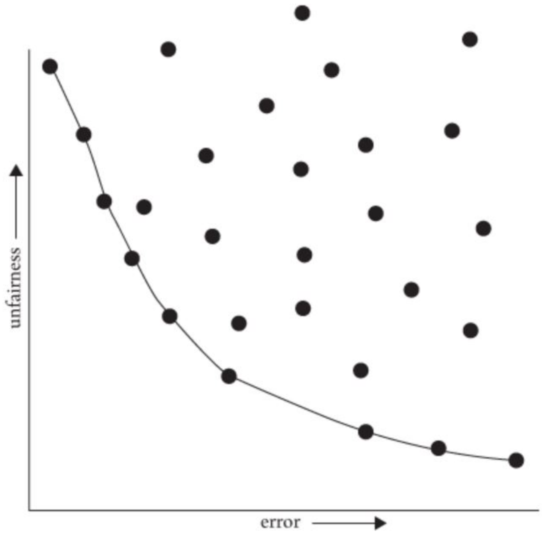
- First, there statistical parity could be achieved with random selection (no need to know individual characteristics)
  - Yet, we can improve on that results by having an AI model that is constrained to take certain action with an equal rates across the different groups
  - Random selection could be good in an *exploration* phase
- Second, statistical parity do not care about the outcome (e.g., the credit-worthiness of each applicant)
  - For example, to obey statistical parity the lender must either deny loans to some good applicants, or give loans to some bad applicants

Statistical parity

Statistical parity has limitations

- First, there statistical parity could be achieved with random selection (no need to know individual characteristics)
  - Yet, we can improve on that results by having an AI model that is constrained to take certain action with an equal rates across the different groups
  - Random selection could be good in an *exploration* phase
- Second, statistical parity do not care about the outcome (e.g., the credit-worthiness of each applicant)
  - For example, to obey statistical parity the lender must either deny loans to some good applicants, or give loans to some bad applicants
- Indeed, there is a **trade-off between fairness and accuracy**
  - How strong is the trade-off? It depends on the application

Pareto frontier



- The Pareto frontier makes our problem as quantitative as possible

"An Economic Approach to Regulating Algorithms" (Rambachan et al 2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

“An Economic Approach to Regulating Algorithms” (Rambachan et al 2020)

- Apply welfare economics to the design and regulation of algorithmic decision processes.
- Algorithmic decision-making has two components:
  - (1) training a prediction function, and (2) a decision rule based on the predictions.

Result 1 (social planner):

- The equity preferences of the social planner do not affect the training procedure for the prediction function.
- i.e., there should be no limit on the use of sensitive attributes.

“An Economic Approach to Regulating Algorithms” (Rambachan et al 2020)

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.

"An Economic Approach to Regulating Algorithms" (Rambachan et al 2020)

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.

"An Economic Approach to Regulating Algorithms" (Rambachan et al 2020)

Result 2 (private actors):

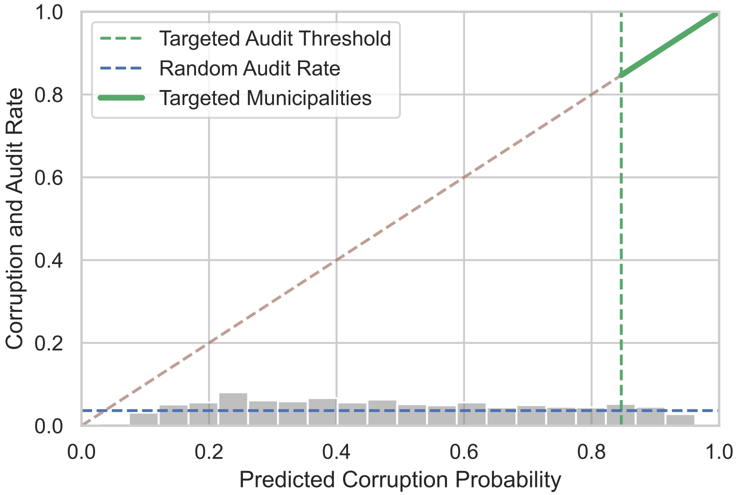
- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.

"An Economic Approach to Regulating Algorithms" (Rambachan et al 2020)

Result 2 (private actors):

- key factor is disclosure of decision process (data, ML training, and decision rule), which, unlike human decision-making, allows prejudicial treatment to be detected.
- without disclosure, algorithms will be just as biased as humans.
- with disclosure, discrimination decreases relative to humans, and government should impose no constraints on the use of sensitive attributes as predictors.
    - caveat: disclosure must include the data and ML training process, not just the decision rule.

Application: Using Machine Learning to Guide Audit Policy

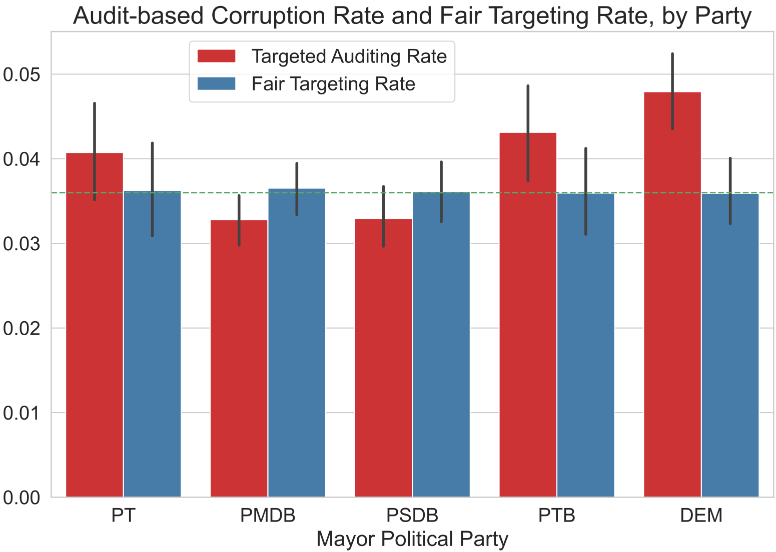


- Rather than sampling 200 municipalities uniformly from distribution, audit 200 with highest $\hat{y}_{it}$.

Application: Using Machine Learning to Guide Audit Policy

- Start with $\hat{y}_i$ for each municipality and the resulting corruption-risk ranking for all municipalities in a given year.
- Produce separate rankings by party.
- Within each party, audit the same share of municipalities.

Audit Allocation with Fair Targeting



- Fair targeting equalizes the probability of targeted audits across parties.

Performance of Fair Targeting

	Random Audits	Targeted Audits		Fair Targeting	
	(1)	(2)		(3)	
Corruption Rate, if Audited	0.466	0.856	(0.016)	0.836	(0.017)
Audit Rate, if Corrupt	0.036	0.067	(0.001)	0.065	(0.001)
$\hookrightarrow$ Ratio over Random Audits		1.836	(0.035)	1.793	(0.037)

Other measures of fairness - Equality in metric values

- We could ask the rate of false negative to be roughly similar across groups (**equality of false negative**)
  - e.g., if we're predicting loan approval and considering fairness for gender, equality of false negatives would require that the rate at which the model **incorrectly denies** loans to creditworthy individuals is the same for men and women.
- We are evenly distributing the mistakes we make in the form of false rejections

Other measures of fairness - Equality in metric values

- We could ask the rate of false negative to be roughly similar across groups (**equality of false negative**)
  - e.g., if we're predicting loan approval and considering fairness for gender, equality of false negatives would require that the rate at which the model **incorrectly denies** loans to creditworthy individuals is the same for men and women.
- We are evenly distributing the mistakes we make in the form of false rejections
- Similarly, we could ask the rate of false positive to be roughly similar across groups (**equality of false positive**)
  - e.g., if we're predicting loan approval and considering fairness for gender, equality of false positives would require that the rate at which the model **incorrectly approves** loans to non-creditworthy individuals is the same for men and women.

Other measure of fairness - Equality in metric values

- Alternatively, one could try to equalize **positive predicted values**
- Individuals who qualify for a desirable outcome should have an equal chance of being correctly classified for this outcome
  - e.g., if we're predicting loan approval and considering fairness for gender, equalizing PPVs would require that the proportion of correct loan approvals is the same for men and women.

Trade-offs

From a computational/mathematical prospective we know:

- that increasing fairness, decreases model precision (e.g., firm profits or crime detection)
- fairness criteria cannot be achieved simultaneously (e.g., equality of false negative + equality of false positive + positive pred. value)

Trade-offs

From a computational/mathematical prospective we know:

- that increasing fairness, decreases model precision (e.g., firm profits or crime detection)
- fairness criteria cannot be achieved simultaneously (e.g., equality of false negative + equality of false positive + positive pred. value)

Algorithms cannot tell us which definition of fairness to use!

Trade-offs

From a computational/mathematical prospective we know:

- that increasing fairness, decreases model precision (e.g., firm profits or crime detection)
- fairness criteria cannot be achieved simultaneously (e.g., equality of false negative + equality of false positive + positive pred. value)

Algorithms cannot tell us which definition of fairness to use!

Policy makers play a crucial role in defining the rule of the game!

Principles and Objectives

Principles

- **Justice, equality, non-discrimination:** AI should respect rights, avoid bias, and treat all fairly.
- **Privacy, surveillance:** AI should respect privacy rights and not be used for unlawful surveillance.
- **Safety and reliability:** AI should be safe, reliable, and perform as intended.

Principles and Objectives

Principles

- **Justice, equality, non-discrimination:** AI should respect rights, avoid bias, and treat all fairly.
- **Privacy, surveillance:** AI should respect privacy rights and not be used for unlawful surveillance.
- **Safety and reliability:** AI should be safe, reliable, and perform as intended.

Objectives

- **Accuracy:** AI should produce accurate results.
- **Equity:** AI should promote fair distribution of benefits and harms.
- **Explainability:** AI decisions should be understandable and explainable.

Principles and Objectives

Principles

- **Justice, equality, non-discrimination:** AI should respect rights, avoid bias, and treat all fairly.
- **Privacy, surveillance:** AI should respect privacy rights and not be used for unlawful surveillance.
- **Safety and reliability:** AI should be safe, reliable, and perform as intended.

Objectives

- **Accuracy:** AI should produce accurate results.
- **Equity:** AI should promote fair distribution of benefits and harms.
- **Explainability:** AI decisions should be understandable and explainable.
- **Auditability, transparency:** AI should be inspectable and its processes open.
- **Responsibility, accountability:** Clear lines of responsibility and accountability for AI should exist.

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices
  - Expands the power of large corporations
  - Negative externalities, tendency to concentration

Governance Strategies

- Industry-driven approach;
  - Reduces regulatory red tape, could help innovation
  - No central authority to enforce best-practices
  - Expands the power of large corporations
  - Negative externalities, tendency to concentration
- Regulator-driven approach:
  - could reduce externalities and concentration
  - significant technical knowledge/skills needed to be effective
  - could limit innovation and expansion of digital economy
  - could collude with industry leaders
- Recent discussion on regulating AI pushed by LLMs deployment (not always about fairness)

Fairness auditing

- Many countries are making efforts to integrate fairness auditing into legislation
- Critical algorithms should be assessed on a yearly basis and the results reported to the government
- Companies would also be required to document how their algorithms are build, how the algorithm makes determinations and all of the determinations made

Fairness auditing

Auditing for

- hidden influences of sensitive variables on other variables
- distribution of model errors for the different classes of a sensitive variable (and repair with statistical distortion)
- false positive rate across different groups
- more general fairness statistics

NYC Law Automated Employment Decision Tools (AEDTs)

- From January (July) 1, 2023, New York City employers will be prohibited from using AI in employment decision-making processes without providing bias audit of the tool
- Automated tools employers include
  - resume scanners that prioritize applications based on certain content;
  - employee monitoring software that analyzes employee performance;
  - virtual assistants and video technologies that evaluate candidates' mannerisms and other characteristics.

NYC Law Automated Employment Decision Tools (AEDTs)

Employers must:

1. Subject AEDTs to a bias audit (by independent auditors) within one year of its use
2. Ensure that the results of such audits are publicly available
3. Provide particular notices to job candidates regarding the employer's use of these tools
4. employee monitoring software that analyzes employee performance;
5. Allow candidates or employees to potentially request alternative evaluation processes as an accommodation

Tools for Auditing

- Usually large consultancies or specialized players that offer fairness auditing in form of a report/tools
- Fairness audits are often times offered by the open source community
- There exist many tools and frameworks to independently audit machine learning models (e.g., Aequitas, IBM Fairness 360, Fairlearn)

Tools for Auditing

- Usually large consultancies or specialized players that offer fairness auditing in form of a report/tools
- Fairness audits are often times offered by the open source community
- There exist many tools and frameworks to independently audit machine learning models (e.g., Aequitas, IBM Fairness 360, Fairlearn)
- We just launched our company for AI certification - CertifAI

Tools for Auditing

- Usually large consultancies or specialized players that offer fairness auditing in form of a report/tools
- Fairness audits are often times offered by the open source community
- There exist many tools and frameworks to independently audit machine learning models (e.g., Aequitas, IBM Fairness 360, Fairlearn)
- We just launched our company for AI certification - CertifAI
- However, while there are many tools to test the model for fairness, it's much harder to find information on how to tackle the problem of an unfair model

How to ensure fairness?

- **Pre-processing methods**: These methods aim to remove bias from the data before it's used to train the algorithm. This could involve modifying the data to ensure that outcomes are independent of protected attributes, or generating synthetic data to balance the representation of different groups.

How to ensure fairness?

- **Pre-processing methods**: These methods aim to remove bias from the data before it's used to train the algorithm. This could involve modifying the data to ensure that outcomes are independent of protected attributes, or generating synthetic data to balance the representation of different groups.
- **In-processing methods**: These methods incorporate fairness constraints directly into the learning process. For example, a machine learning model could be trained to not only minimize prediction error, but also to minimize disparity in outcomes between different groups.

How to ensure fairness?

- **Pre-processing methods**: These methods aim to remove bias from the data before it's used to train the algorithm. This could involve modifying the data to ensure that outcomes are independent of protected attributes, or generating synthetic data to balance the representation of different groups.
- **In-processing methods**: These methods incorporate fairness constraints directly into the learning process. For example, a machine learning model could be trained to not only minimize prediction error, but also to minimize disparity in outcomes between different groups.
- **Post-processing methods**: These methods adjust the predictions of a trained model to improve fairness. For example, thresholds for decision-making could be adjusted for different groups to equalize false positive or false negative rates.

How to ensure fairness?