# Algorithmic Fairness Through (Wasteful[1]) Counterfactual Analysis and Optimal Transport

**Ewen Gallic**

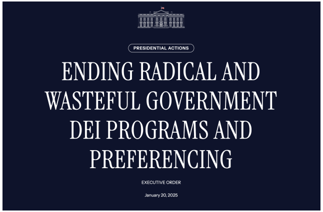
joint with Arthur Charpentier and Agathe Fernandes Machado

Séminaire interne CRM-CNRS, 18 Février 2025

[1]Not in my opinion.

# Motivations

*The Biden Administration forced* ***illegal and immoral discrimination programs****, going by the name “****diversity, equity, and inclusion”*** *(DEI), into virtually all aspects of the Federal Government, in areas ranging from airline safety to the military. [. . . ] The public release of these plans demonstrated* ***immense public waste and shameful discrimination****. That ends today. Americans deserve a government committed to serving every person with equal dignity and respect [. . . ]*





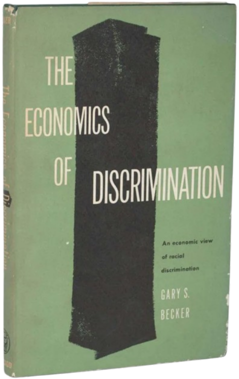
Executive orders 14151 (“Ending Radical and Wasteful Government DEI Programs and Preferencing” Jan 20, 2025) and 14173 (“Ending Illegal Discrimination And Restoring Merit-Based Opportunity” Jan 21, 2025)

# Broad Framework

- We want to make **predictions** on an outcome variable (e.g., claim frequency, loan default risk, recidivism).
- To do so, we use a **statistical model**, or a machine learning model fed with **historical data**.
- To comply with regulations, we want to obtain a model that **does not discriminate** with respect to a **sensitive attribute**.

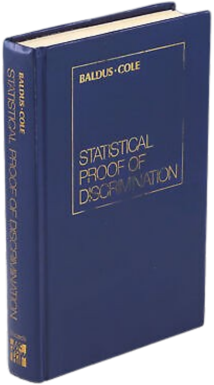

Digital illustration of fairness and machine learning generated using DALL-E 3. Retrieved from ChatGPT Interface.

# What is Discrimination? An Economic Perspective




- In economics, following Becker (1957), **discrimination**: situations in which individuals are treated differently based on attributes such as race, gender, etc., rather than their productivity or other relevant characteristics.
  - **Disparate treatment** (or taste-based discrimination): intentional discrimination, where individuals are treated differently explicitly because of a **protected characteristic**.
  - **Disparate impact**: policy, practice, or decision that appears neutral on the surface disproportionately affects members of a **protected group**, even without intentional discrimination.
- From a Law perspective: **direct** vs. **indirect** discrimination (Campbell and Smith, 2023)

# What is Discrimination? A Statistical Perspective

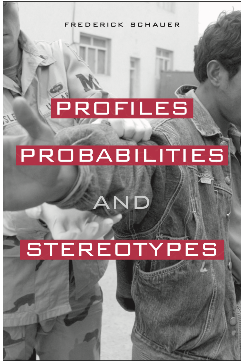



- **Statistical discrimination** (see, e.g., Baldus and Cole, 1980): individuals are treated differently based on group-level statistical averages, rather than their individual characteristics. They do not arise from prejudice or bias but from **decision-makers relying on imperfect information** and using group membership as a **proxy** for individual traits.
- Some forms of discrimination are considered unacceptable (Hellman, 2008).
- Fisher (1936): separating or classifying observations into distinct groups based on measured characteristics. In this context, discrimination is purely a statistical operation with no connotation of social bias or inequality.
- However, statistical discrimination may lead to:
  - Reinforcement of Biases (through lack of opportunities).
  - Legal and Ethical Concerns.

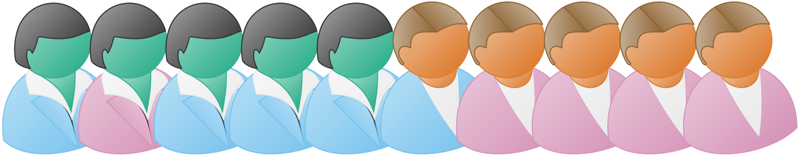
# A Focus on the Actuarial Context: Risk Discrimination

In this talk, we will focus on predictive models rooted in **actuarial science**.

"***To be an actuary is to be a specialist in generalization***, *and actuaries engage in a form of decisionmaking that is sometimes called actuarial. Actuaries guide insurance companies in* ***making decisions about large categories*** *(teenage males living in northern New Jersey) that have the effect of attributing to the entire category certain characteristics (carelessness in driving) that are probabilistically indicated by membership in the category, but that still may not be possessed by a particular member of the category (this particular teenage male living in northern New Jersey).*" (Schauer 2006, p. 4)



# Assessing Risk for Managing Solvency



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10% | 20% | 10% | 10% | 10% | 10% | 20% | 20% | 20% | 20% |

- To cover future claims, insurance companies must set their premiums.
- The pricing exercise boils down to a fair allocation problem in Game Theory (Nash, 1950; Shapley, 1953; Harsanyi, 1959)
- A solution: **actuarially fair premiums** Arrow (1963):
  - individuals with similar risk levels pay similar amounts (**horizontal equity**),
  - those with higher risks pay correspondingly higher premiums (**vertical equity**).

# Actuarial Fairness: Is it "Fair?"

To be **actuarially fair**, the **premiums should be equal to the expected loss of the insured risks** (Arrow, 1963)

"*In the insurance industry, the concept of actuarial fairness serves to establish what could be adequate, **fair premiums**. Accordingly, premiums paid by policyholders should **match as closely as possible their risk exposure** (i.e. their expected losses). Such premiums are the product of the probabilities of losses and the expected losses.*" (Landes, 2014)

"*Since the insurer assumes the individual insured's risk of loss, **the premium should be fundamentally based upon the expected value of an insured's losses**.*" (Walters, 1981)

J Bus Ethics
DOI 10.1007/s10551-014-2120-0

## How Fair Is Actuarial Fairness?

**Xavier Landes**

Received: 15 August 2013 / Accepted: 20 February 2014


**Abstract** Insurance is pervasive in many social settings. As a cooperative device based on risk pooling, it serves to attenuate the adverse consequences of various risks (health, unemployment, natural catastrophes and so forth) by offering policyholders coverage against the losses implied by adverse events in exchange for the payment of premiums. In the insurance industry, the concept of actuarial fairness serves to establish what could be adequate, fair premiums. Accordingly, premiums paid by policyholders should match as closely as possible their risk exposure (i.e. their expected losses). Such premiums are the product of the probabilities of losses and the expected losses. This article presents a discussion of the fairness of actuarial fairness through three steps: (1) defining the concept based on its formulation within the insurance industry; (2) determining in which sense it may be about fairness; and (3) raising some objections to the actual fairness of actuarial fairness. The necessity of a normative evaluation of actuarial fairness is justified by the influence of the concept on the current reforms of public insurance systems and the fact that it highlights the question of the repartition of the gains and burdens of social cooperation.

**Keywords** Cooperation · Expected utility · Insurance · Fairness · Premiums · Responsibility

Insurance is an important mechanism of cooperation for modern industrialized societies. The principle is that individuals gather resources against risk. By doing so, they are said to 'pool' their risks. Therefore, insurance is usually characterized as a risk-pooling device. Fundamentally, insurance is a cooperative mechanism that transforms the significance and implications of random events by mutualizing risks and their adverse consequences.[1]

From an individual point of view, risks imply uncertainty. For instance, what is *your* chance (not the average probability of the population you belong to) of being run over by a car? Of being afflicted by cancer? Of becoming unemployed or outliving your personal savings? These questions cannot be answered at a strictly individual level. Individuals experience uncertainty regarding accidents of various sorts: disease, chronic pathology, economic downturn, death and so forth. From a collective point of view, though, uncertainty can be converted into probabilities. In the case of car accidents, for instance, instead of pure uncertainty, I may know that I have 1:100 odds of getting involved in an accident and, perhaps, 1:1,000 odds of dying. Risk pooling provides the opportunity to calculate the probabilities of a particular set of events. In return, the expected costs for such risks can be calculated and spread over the policyholders through premiums.

In that sense, insurance is about transforming uncertain adverse events with uncertain outcomes into statistical events with certain outcomes: the expected losses that the payment of the premiums reflects. Following Knightian

X. Landes (✉)
University of Copenhagen, Copenhagen, Denmark
e-mail: xavier.landes@gmail.com

[1] Any discussion of insurance should distinguish between insurance as the general cooperative arrangement among individuals based on risk pooling and specific arrangements where an insurer acts as an intermediary between policyholders. In other words, a distinction should be made between the concept and different conceptions of insurance. This article is mainly about the concept of insurance as a cooperative mechanism. Even if mentions are made of cases where an insurer acts as an intermediary, the locus of the article remains the *philosophy of insurance*, i.e. the normative principles that should prevail when individuals decide to pool their risks in the face of uncertainty. We are grateful to an anonymous referee who brought this point to our attention.

Published online: 07 March 2014
Springer

# Toy Example: Risk Estimation

Assume we want to predict **claim frequency** using a Poisson regression model, using three predictors.

Further assume that the number of claims $y$ has a Poisson distribution with a conditional mean that depends on some features $\boldsymbol{X}$ according to the following structural model:

$$E(y_i|\boldsymbol{X}_i) = \exp\left(\boldsymbol{X}_i\beta\right)$$

**The set of predictors $\boldsymbol{X}$** contains three features :

- A binary variable indicating whether the insured lives in an urban area.
- The insured's age.
- The insured's gender.

# Toy Example: Risk Estimation

The predicted value will thus be:

$$\begin{cases} \hat{y}(\text{man}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{\text{urban}} + \hat{\beta}_2 \text{age} + \hat{\beta}_3\right] \\ \hat{y}(\text{woman}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{\text{urban}} + \hat{\beta}_2 \text{age}\right] \end{cases}$$

Hence:

$$\hat{y}(\text{man}) = \exp\left[\hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{\text{urban}} + \hat{\beta}_2 \text{age} + \underbrace{\boxed{\hat{\beta}_3}\, \mathbf{1}_{\text{man}}}_{\times e^{\beta_3} \text{ ceteris paribus}}\right] = \hat{y}(\text{woman}) \cdot \boxed{\exp[\beta_3]}$$

If $\beta_3$ is small, $e^{\beta_3} \approx 1 + \beta_3$. Thus, if $\beta_3 = 0.2$, it corresponds to $+20\%$ for men.

# Toy Example: Risk Estimation

- In the toy example, the estimates indicate that men are at higher risks than women:
  - **Gender** is a statistical predictor.
- With such insight from the data, should the premium paid by men to an insurance company be higher than that paid by women?
- In other words, should the insurance company **discriminate** by gender in such a context?
  - risk-based discrimination
  - discrimination w.r.t. a **sensitive attribute**.

# Policymakers Point of View: Europe

- Europe: Court of Justice of the European Union – 2011

"*At the moment, a careful young male driver pays more for auto insurance* ***just because he is a man****. Under the ruling,* ***insurers can no longer use gender as the sole determining risk factor to justify differences in individuals' premiums****. But the premiums paid by careful drivers – male and female – will continue to decrease based on their individual driving behaviour. The ruling does not affect the use of other legitimate risk-rating factors (such as, for example, age or health status) and prices will continue to reflect risk.*" (Commission, 2011 through Frezal and Barry, 2019)

# Policymakers Point of View: Québec

- Québec: Charte des droits et libertés de la personne (C-12, Article 20.1)

  "*Dans un contrat d'assurance ou de rente, un régime d'avantages sociaux, de retraite, de rentes ou d'assurance ou un régime universel de rentes ou d'assurance, une distinction, exclusion ou préférence fondée sur l'âge, le sexe ou l'état civil est* ***réputée non discriminatoire*** *lorsque son utilisation est légitime et que le* ***motif qui la fonde constitue un facteur de détermination de risque, basé sur des données actuarielles****.*"

# Fair Discrimination in Insurance: an Oxymoron

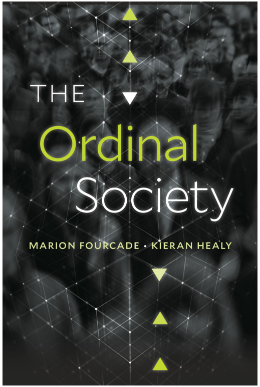
"*what is unique about insurance is that even statistical discrimination (the act by which an insurer uses a characteristic of an insured or potential insured as a statistic for the risk it poses to an insurer), which by definition is absent any malicious intentions, poses significant moral and legal challenges. Why? Because* ***on the one hand, policy makers*** *would like insurers to* ***treat their insureds equally, without discriminating*** *based on* ***race, gender, age, or other characteristics****,* ***even if it makes statistical sense*** *to discriminate.* [...] ***On the other hand****, at the core of* ***insurance business lies discrimination between risky and non-risky insureds****.* ***But riskiness often statistically correlates with the same characteristics policy makers would like to prohibit insurers from taking into account****.*" (Avraham, 2017)

# Individual Characteristics

- In our example, **gender** may be a statistical predictor, but from the European legislation perspective using it leads to a **direct discrimination**.
- Here, **gender** is not a **causal predictor**. It does not reflect **individual behavior**.
- In the era of **big data** and **artificial intelligence**, a naive solution consists in hiding the sensitive attribute, and use a **machine learning model** trained on additional (hopefully behavioral) data:
  - explicability issues
  - proxy discrimination issues (Pedreshi et al., 2008; Dwork et al., 2012).

# Individual Characteristics

"***shifting from socialized to individualized risk also transforms the very purpose of insurance**. [...] the most significant **sources of risk**—and thus the proper allocation of responsibility—**may lie outside the individual** in the natural or social environment. The fact that these structural forces cannot easily be measured does not mean that they can be conveniently ignored. Doing so not only **excludes people unfairly** but **also threatens** the way that insurance systems can act as a prosaic but intensely practical manifestation of **solidarity**.*" (Fourcade and Healy, 2024)



# Individualization, Actuarial Fairness and Accuracy

- We can also question the "accuracy" of individual predictions.
- Recall that following Arrow (1963):

  "**actuarially fair premiums**" = "**expected losses**"

- But, with different models and different portfolio, we can have different premiums.
  - There is no law of one price in insurance.

  "*The Law states that identical goods must have identical prices. [...] Economic theory teaches us to expect the Law to hold exactly in competitive markets with no transactions costs and no barriers to trade.*" (Lamont and Thaler, 2003)

# Individualization, Actuarial Fairness and Accuracy

- Premiums are based on an **estimation of the expected loss** that maximizes **accuracy**:

$$\underbrace{\bar{y} = \underset{\gamma\in\mathbb{R}}{\text{arg min}}\left\{\sum_{i=1}^{n}(y_i-\gamma)^2\right\}}_{\text{average loss}} \text{ or } \underbrace{\mathbb{E}[Y] = \underset{\gamma\in\mathbb{R}}{\text{arg min}}\left\{\sum_{y}(y-\gamma)^2\,\mathbb{P}[Y=y]\right\}}_{\text{empirical losses}}$$

average loss / empirical losses

least squares

i.e., we want to minimize the error between observed loses $y$ and predictions $\hat{y}$.

- If the prediction is a binary outcome $y \in \{0,1\}$ (e.g., accident, default), it is hard to assess if $\hat{y} = 8.2740164\%$ is accurate or not.

# Individualization, Actuarial Fairness and Accuracy

Does accuracy for a single individual make any sense?

"*When we speak of the 'probability of death', the exact meaning of this expression can be defined in the following way only.* ***We must not think of an individual, but of a certain class as a whole****, e.g., 'all insured men forty-one years old living in a given country and not engaged in certain dangerous occupations'. A probability of death is attached to the class of men or to another class that can be defined in a similar way. We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail.* ***The phrase 'probability of death', when it refers to a single person, has no meaning at all for us.***" (von Mises, 1957) (p. 11)

# Individualization, Actuarial Fairness and Accuracy

Is the predicted value well estimated? "*among patients with an* ***estimated risk of 20%****, we expect 20 in 100 to have or to develop the event*" (Van Calster et al., 2019)

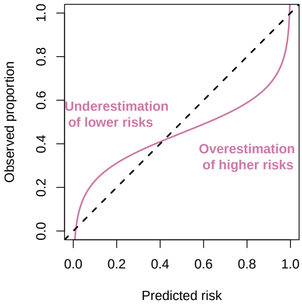
- If 40 out of 100 in this group are found to have the disease, the risk is **underestimated**.
- If 10 out of 100 in this group are found to have the disease, the risk is **overestimated**.

The prediction $\hat{m}(\boldsymbol{X})$ of $Y$ is a **well-calibrated** prediction if:

20 out of 100 (proportion $y = 1$)

$$\mathbb{E}[Y \mid \hat{Y} = \hat{y}] = \hat{y}, \quad \forall \hat{y}$$

estimated risk $\hat{y} = 20\%$



# Individualization, Actuarial Fairness and Accuracy

A model will be:

- Globally **well balanced** if:

$$\mathbb{E}[\underbrace{\hat{Y}}_{\text{premium collected}}] = \mathbb{E}[\underbrace{Y}_{\text{losses paid}}]$$

- Locally well balanced, or **well-calibrated** if:

$$\mathbb{E}[\hat{Y} \mid \hat{Y} = \hat{y}] = \mathbb{E}[Y \mid \hat{Y} = \hat{y}] = \hat{y}, \quad , \forall \hat{y}$$

For more details on calibration see Fernandes Machado et al. (2024a,b)

# Road Map

# Quantifying Unfairness

# What is Algorithmic Fairness?

- Let $m : \mathcal{X} \to \mathcal{Y}$ be a predictive model that predicts an outcome $Y$ (e.g., claims) w.r.t. a **sensitive attribute** $S \in \mathcal{S}$ (e.g., gender, race) using features $\boldsymbol{X}$.
- Regulations may prohibit **discrimination** on the sensitive attribute, requiring $m$ to be fair w.r.t. to $S$.
- **Approaches** to **evaluate** and, if necessary, **mitigate** the unfairness of model predictions $\hat{Y} = m(\boldsymbol{X})$ for $S$:
  - **Group fairness**: compare $\hat{Y}$ between groups defined by $S$, e.g., salary for males vs. salary for females (Barocas et al., 2023; Hardt et al., 2016).
  - **Individual fairness**: focus on a specific individual in the disadvantaged group (Dwork et al., 2012).
  - **Counterfactual fairness**: causality-based fairness (Plečko and Meinshausen, 2020; Plečko et al., 2024)

# Mitigation

Some techniques can be used to prevent models from perpetuating biases with respect to the sensitive attribute. These techniques can be applied at several stages (Hajian and Domingo-Ferrer, 2013)

1. **Preprocessing**: transform source data to remove biases before model training.
2. **In-processing**: modify algorithms to embed fairness constraints during training.
3. **Postprocessing**: alter models after training to correct unfair outcomes.

# How Can Fairness be Quantified?

We would like to **quantify unfairness** of a **supervised model** $\hat{m}(\cdot)$ trained on a set $\{(y_i, \boldsymbol{x}_i, s_i)\}_{i=1}^n$, where $y$ is the value to predict (i.e., the outcome), $\boldsymbol{x}$ is a set of (unprotected) predictors, $s$ is a **protected attribute**, and $i \in \{1, \ldots, n\}$ denotes an individual.

The outcome may be:

- **Binary** (classification task):
  - $\hat{y}_i = \mathbf{1}(\hat{m}(\boldsymbol{x}_i, s_i) > \text{threshold}) \in \{0, 1\}$
- **Continuous** (regression task):
  - $\hat{y}_i = \hat{m}(\boldsymbol{x}_i, s_i) \in [0, 1]$: a score
  - $\hat{y}_i = \hat{m}(\boldsymbol{x}_i, s_i) \in \mathbb{R}$: a premium

# Group Fairness Metrics in a Nutshell

**Demographic Parity** $\rightarrow \mathbb{E}[\underbrace{\hat{Y}}_{\text{score } \hat{y}} \mid \overbrace{S = A}^{\text{sensitive}}] \stackrel{?}{=} \mathbb{E}[\underbrace{\hat{Y}}_{\text{score } \hat{y}} \mid \overbrace{S = B}^{\text{sensitive}}]$

**Equalized Odds** $\rightarrow \mathbb{E}[\hat{Y} \mid \overbrace{Y = y}^{\text{outcome } y}, S = A] \stackrel{?}{=} \mathbb{E}[\hat{Y} \mid \overbrace{Y = y}^{\text{outcome } y}, S = B], \forall y$

**Calibration** $\rightarrow \mathbb{E}[Y \mid \hat{Y} = u, S = A] \stackrel{?}{=} \mathbb{E}[Y \mid \hat{Y} = u, S = B], \forall u$

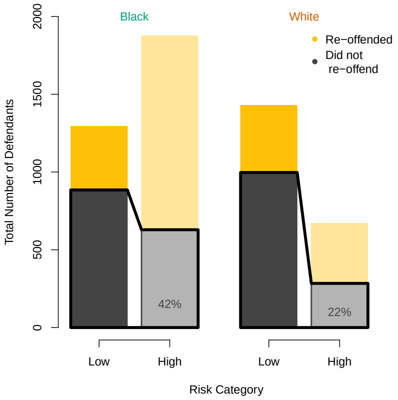
# Illustration With the COMPAS Dataset

- The algorithm "**C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions" attributes a score to each convicted individual in some states in the U.S.A, to estimate the likelihood of them committing a crime again if they are released from prison.
- This scoring classifier uses more than 100 predictors.
- **Race** is not one of them. However, when looking at the predicted values of the model, Angwin et al. (2016) claimed it was biased against Black people.
- The dataset they used is now available in an `R` packages: `fairness`.

# Equality of False Positive Rates?

Larson et al. (2016) looked at the **Equalized Odds**:

- For **Black people**, among those who did **not re-offend** ($y$), 42% were **wrongly classified** ($\hat{y} \neq y$).
- For **White people**, among those who did **not re-offend**, 22% were **wrongly classified**.
- Since 42% $\gg$ 22%: **unfair**.

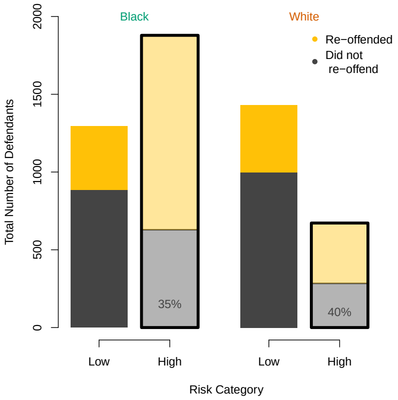



$$\mathbb{P}[\hat{Y} = \textbf{High} \mid Y = \textbf{no}, S = \textbf{Black}] = 42\% \overset{?}{=} \mathbb{P}[\hat{Y} = \textbf{High} \mid Y = \textbf{no}, S = \textbf{White}] = 22\%$$

# Another Metric, Another Result...

Dieterich et al. (2016): **predictive parity** (recidivism rate at each risk level)

- For **Black people**, among those who were **classified as high risk** ($\hat{y}$), 35% did **not re-offend** ($y$).
- For **White people**, among those who were **classified as high risk**, 40% did **not re-offend**.
- Since 35% $\approx$ 40%: **fair**.




$$\mathbb{P}[Y = \text{no} \mid \hat{Y} = \text{High}, S = \text{Black}] = 35\% \overset{?}{=} \mathbb{P}[Y = \text{no} \mid \hat{Y} = \text{High}, S = \text{White}] = 40\%$$

# Adjusting the Probability Threshold

We focus on **binary decisions** ($\hat{y} \in \{0, 1\}$).

$$\textbf{Demographic Parity} \rightarrow \mathbb{P}[\underbrace{\hat{Y} = 1}_{\text{binary decision } \hat{y}} \mid S = A] \overset{?}{=} \mathbb{P}[\underbrace{\hat{Y} = 1}_{\text{binary decision } \hat{y}} \mid S = B]$$

These decisions are usually based on **scores**, using a **threshold** $\tau$:

$$\textbf{Demographic Parity} \rightarrow \mathbb{P}[\underbrace{\hat{m}(\boldsymbol{X}, S)}_{\text{score } \hat{m}} > \tau \mid S = A] \overset{?}{=} \mathbb{P}[\underbrace{\hat{m}(\boldsymbol{X}, S)}_{\text{score } \hat{m}} > \tau \mid S = B]$$

Demographic Parity can be achieved **by setting different threshold in the groups**:

$$\textbf{Demographic Parity} \rightarrow \mathbb{P}[\hat{m}(\boldsymbol{X}, S) > \tau_A \mid S = A] = \mathbb{P}[\hat{m}(\boldsymbol{X}, S) > \tau_B \mid S = B]$$

It is then usually impossible to achieve **equalized odds** with this strategy.

# Counterfactual Fairness

1. **Affirmative actions**: "*The contractor will not discriminate against any employee or applicant for employment because of race, creed, color, or national origin. The contractor will* ***take affirmative action*** *to ensure that applicants are employed, and that employees are treated during employment, without regard to their race, creed, color, or national origin*" (John F. Kennedy, EO #10925, March 6, 1961)
   "*In order to get beyond racism, we must first take account of race. There is no other way. And* ***in order to treat some persons equally, we must treat them differently****.*" (Justice Harry Blackmun, Regents of Univ. of Cal. v. Bakke, 438 U.S. 265, 407, via Scalia (1979))
2. **Blindness**: "*The way to stop discrimination on the basis of race is to* ***stop discriminating on the basis of race****.*" (Chief Justice John G. Roberts, Jr, Parents Involved in Community Schools v. Seattle School District No. 1, via Turner (2015))

# Counterfactual Fairness

1. ***Ceteris paribus***: "*We capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly*" (Dwork et al., 2012).

   Similarity fairness is achieved if for all $i \neq j$ such that $\boldsymbol{x}_i = \boldsymbol{x}_j$ and $s_i \neq s_j$, then:

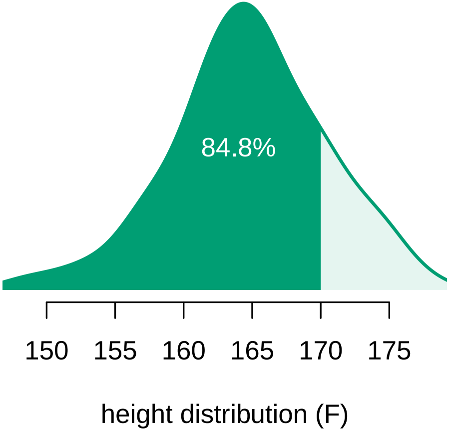
   $$m(\boldsymbol{x}_i, s_i = \mathbf{A}) = m(\boldsymbol{x}_j, s_j = \mathbf{B})$$

2. ***Mutatis mutandis***: build on the idea of counterfactuals: "*What would this woman earnings would have been had she been a man?*" (Kusner et al., 2017; Kilbertus et al., 2017a; De Lara et al., 2021; Charpentier et al., 2023; Fernandes Machado et al., 2024c)

# Building Counterfactuals

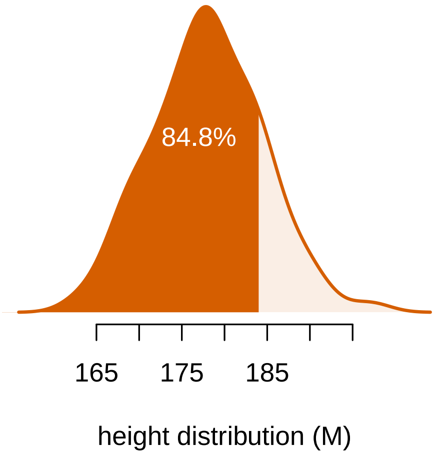
Consider the height of **females** and **males**.

- What is the counterfactual of a **female** with height 170cm (=5' 7") had she been a **male**?
- Within the distribution of **females**, this corresponds to a quantile level $\alpha = 84.8\%$.
  - $F_{\text{female}}(170) = 84.8\%$.



# Building Counterfactuals

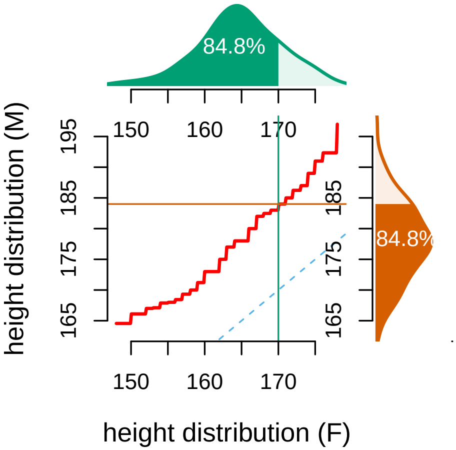
- The corresponding quantile in the height distribution of **males** is:
  - $F^{-1}_{\text{male}}(84.8\%) = 184\text{cm}\ (\approx 6')$.



# Building Counterfactuals

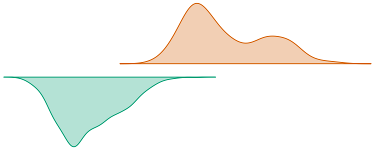
Counterfactual of a 170cm (=5' 7") **female** had she been a **male**?

$$T^\star(170) = \left( F_{\text{male}}^{-1} \circ F_{\text{female}} \right)(170)$$
$$= 184 \text{ cm } (\approx 6').$$



# Optimal Transport and Monge Mapping

- **Optimal Transport**: how to find the best way to transport mass from **one distribution** to **another** while minimizing a given cost.
- It involves constructing a **joint distribution** (coupling) between two marginal probability measures (Villani, 2003, 2009).
- Consider a measure $\mu_0$ (resp. $\mu_1$) on a metric space $\mathcal{X}_0$ (resp. $\mathcal{X}_1$). The goal is to move every elementary mass from $\mu_0$ to $\mu_1$ in the most "efficient way."



From Monge (1781): Mémoire sur la théorie des **déblais** et des **remblais**.

# Optimal Transport and Monge Mapping

### Definition

Let $\mathcal{X}_0$ and $\mathcal{X}_1$ be two metric spaces. Suppose a map $T : \mathcal{X}_0 \to \mathcal{X}_1$.
The push-forward of $\mu_0$ by $T$ is the measure $\mu_1 = T_{\#}\mu_0$ on $\mathcal{X}_1$ s.t.
$\forall B \subset \mathcal{X}_1, \quad T_{\#}\mu_0(B) = \mu_0(T^{-1}(B))$.

### Proposition

For all measurable and bounded $\varphi : \mathcal{X}_1 \to \mathbb{R}$,

$$\int_{\mathcal{X}_1} \varphi(x_1)\, dT_{\#}\mu_0(x_1) = \int_{\mathcal{X}_0} \varphi(T(x_0))\, d\mu_0(x_0) \quad .$$

# Optimal Transport and Monge Mapping

**Proposition**

If $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}^d$ and $\mu_0$ is atomless, then there exists $T$ such that $\mu_1 = T_{\#}\mu_0$.

**Definition: Monge problem, Monge (1781)**

If we further assume $\mu_0$ and $\mu_1$ are absolutely continuous w.r.t. Lebesgue measure, then we can find an "optimal" mapping, satisfying

$$\inf_{T_{\#}\mu_0 = \mu_1} \int_{\mathcal{X}_0} c(x_0, T(x0))\, d\mu_0(x_0),$$

for a general cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}^+$.

The optimal mapping is denoted $T^*$.

# Optimal Transport plans

In general settings, however, such a deterministic mapping $T^*$ between probability distributions may not exist.

**Kantorovich relaxation, Kantorovich (1942)**

The Kantorovich relaxation of Monge mapping is defined as

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathcal{X}_0 \times \mathcal{X}_1} c(\boldsymbol{x}_0, \boldsymbol{x}_1) \pi(\mathrm{d}\boldsymbol{x}_0, \mathrm{d}\boldsymbol{x}_1),$$

for a general cost function $c : \mathcal{X}_0 \times \mathcal{X}_1 \to \mathbb{R}^+$ and $\Pi(\mu_0, \mu_1)$ the set of all couplings of $\mu_0$ and $\mu_1$.

This problem always admits solutions and focuses on couplings rather than deterministic mappings.

# Univariate Optimal Transport Map

Suppose here that $\mathcal{X}_0 = \mathcal{X}_1$ is a compact subset of $\mathbb{R}$.

As shown in Santambrogio (2015), the optimal Monge map $T^\star$ for some strictly convex cost $c$ such that $T^\star_\# \mu_0 = \mu_1$ is:

$$T^\star = \underbrace{F_1^{-1}}_{\text{generalized inverse (quantile function)}} \circ \overbrace{F_0}^{\text{cdf associated with } \mu_0},$$

# Counterfactual Fairness in Brief: Links with Causal Inference

| | Sex | Name | Treatment | Weight (Outcome) | | | | Height | ... |
|---|---|---|---|---|---|---|---|---|---|
| | | | $t_i$ | $y_i$ | $y^\star_{i,T\leftarrow A}$ | $y^\star_{i,T\leftarrow B}$ | TE | $x_i$ | ... |
| 1 | H | Alan | **A** | 75 | **75** | **64** | **11** | 172 | ... |
| 2 | F | Britney | **B** | 52 | **67** | **52** | **15** | 161 | ... |
| 3 | F | Aya | **B** | 57 | **71** | **57** | **14** | 163 | ... |
| 4 | H | Amir | **A** | 78 | **78** | **61** | **17** | 183 | ... |

Difference in the **potential outcomes** (or treatment effect):

$$\mathbf{TE} = y^\star_{i,T\leftarrow B} - y^\star_{i,T\leftarrow A}$$

If $s_i = \mathbf{A}$:

- the **observed value** is $y^\star_{i,T\leftarrow A}$
- the **counterfactual** is $y^\star_{i,T\leftarrow B}$

For More details on causal inference, see, e.g., Imbens and Rubin (2015); Pearl and Mackenzie (2018); Cunningham (2021); Chernozhukov et al. (2024)

# Counterfactual Fairness in the *ceteris paribus* case

Counterfactual fairness is achieved, on average, if:

$$\text{ATE} = \mathbb{E}\left[ Y_{S \leftarrow A} - Y_{S \leftarrow B} \right] = 0$$

A decision satisfies counterfactual fairness if "*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*." (Kusner et al., 2017)

Counterfactual fairness for an individual with characteristics $\boldsymbol{x}$ is achieved if:

$$\text{CATE}(\boldsymbol{x}) = \mathbb{E}\left[ Y_{S \leftarrow A} - Y_{S \leftarrow B} \mid \boldsymbol{X} = \boldsymbol{x} \right] = 0$$

Approach based on **causal graphs** (Plečko and Meinshausen, 2020; Plečko et al., 2024)

# Counterfactual Fairness in the *mutatis mutandis* case

- The **protected attribute** may affect another variable in a manner accepted as non-discriminatory (resolving variable, Kilbertus et al., 2017b).
- The ***mutatis mutandis*** version of the CATE writes:

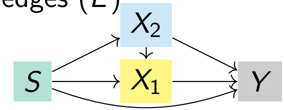
$$\mathbb{E}\left[\boxed{Y_{S\leftarrow A}} \mid \boldsymbol{X}=\boldsymbol{x}\right] - \mathbb{E}\left[\boxed{Y_{S\leftarrow B}} \mid \underbrace{\boxed{\boldsymbol{X}=\boldsymbol{x}^{\star}_{S\leftarrow B}}}_{\text{transported characteristics}}\right]$$

- In this version, $\boldsymbol{X} \mid \textbf{A}$ is transported to $\boldsymbol{X} \mid \textbf{B}$ (see Plečko and Meinshausen, 2020; Plečko et al., 2024; De Lara et al., 2021; Charpentier et al., 2023), according to an assumed **causal structure**.
- In Fernandes Machado et al. (2024c), we propose to unify the **causal graph** & **optimal transport** approaches, using a **sequential transport** approach.

# Counterfactuals with Sequential Transport

# Graphical Models and Causal Networks

- A **Directed Acyclic Graph** (DAG) $\mathcal{G} = (V, E)$ models relationships between variables as nodes ($V$) and edges ($E$)




- Such a causal graph imposes some ordering on variables, referred to as "**topological sorting**" Ahuja et al. (1993). Here,

$$S \to X_2 \to X_1 \to Y \quad .$$

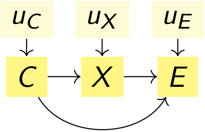
- The joint distribution of $X = (X_1, \ldots, X_d)$ satisfies the **Markov property**:

$$\mathbb{P}[x_1, \cdots, x_d] = \prod_{j=1}^{d} \mathbb{P}[x_j | \text{parents}(x_j)],$$

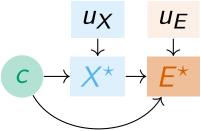
where parents$(x_i)$ are the immediate causes of $x_i$.

# Counterfactual for Non Linear Models

- From Pearl (2000), let $C, X, E$ be absolutely continuous, and consider $i$ where $E_i = h_i(\text{parents}(E_i), U_i)$ with $\text{parents}(E_i) = \boldsymbol{x}$ fixed.
- Define $h_{i|\boldsymbol{x}}(u) = h_i(\boldsymbol{x}, u)$.
- $e_i = h_{i|\boldsymbol{x}}(u_i)$ represents the conditional quantile of $E_i$ at probability level $u_i$.
- Its **counterfactual counterpart** $e_i^\star$ is the conditional quantile (conditioned on $\boldsymbol{x}^*$) at the same level $u_i$.




$$\begin{cases} C = h_c(U_C) \\ X = h_x(C, U_X) \\ E = h_e(C, X, U_E), \end{cases}$$




$$\begin{cases} C = c \quad (\text{or do}(C = c)) \\ X_c^\star = h_x(c, U_X) \\ E_c^\star = h_e(c, X_c^\star, U_E), \end{cases}$$

where $u \mapsto h_c(\cdot, u)$, $u \mapsto h_x(\cdot, u)$ and $u \mapsto h_e(\cdot, u)$ are strictly increasing in $u$, $U_C$, $U_X$ and $U_E$ are independent, supposed to be uniform on $[0, 1]$.

# Topological Ordering (1/4)

**Step 1**: Assuming a causal graph $\mathcal{G}$.
**Step 2**: Derive the **topological ordering** from the DAG:

- **Knothe-Rosenblatt rearrangement** (Bonnotte, 2013), inspired by the Rosenblatt chain rule: provides the "monotone lower triangular map" ("marginally optimal" Villani, 2003)

$$T_{\underline{kr}}(x_1,\cdots,x_d)=\begin{pmatrix} T^\star_{\underline{1}}(x_1)\\ T^\star_{\underline{2}}(x_2|x_1)\\ \vdots\\ T^\star_{\underline{d-1}}(x_{d-1}|x_1,\cdots,x_{d-2})\\ T^\star_{\underline{d}}(x_d|x_1,\cdots,x_{d-1})\end{pmatrix}.$$

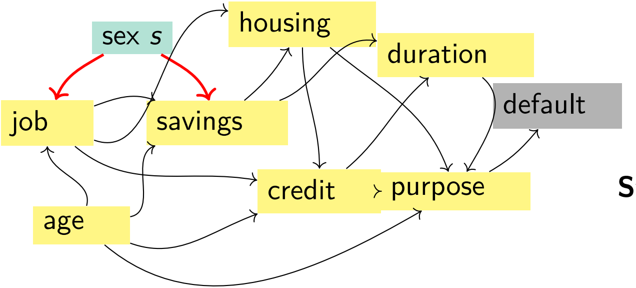
$\rightarrow$ Sequentially mapping $\mathbf{X}|S=0$ to $\mathbf{X}|S=1$ by conditioning on each preceding node in the topological order.

# Topological Ordering (2/4)

- **Sequential Transport** extends the Knothe-Rosenblatt map to transport individuals from $\mathbf{X}|S=0$ to $\mathbf{X}|S=1$, while respecting any assumed underlying causal graph.
- The sequential conditional transport on graph $\mathcal{G}$ writes:

$$T^\star_{\mathcal{G}}(x_1,\cdots,x_d)=\begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|\ parents(x_2)) \\ \vdots \\ T^\star_{d-1}(x_{d-1}|\ parents(x_{d-1})) \\ T^\star_d(x_d|\ parents(x_d)) \end{pmatrix}.$$

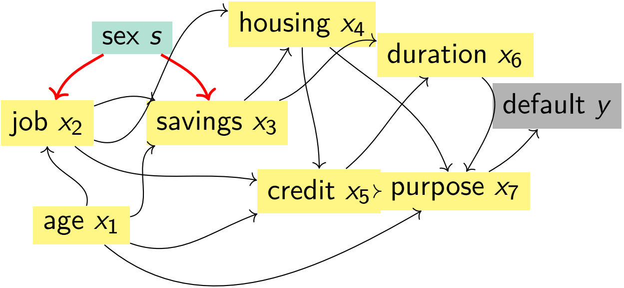
# Topological Ordering (3/4)




Causal graph in the German Credit dataset from Watson et al. (2021).

**Step 1**: Asusming a causal graph $\mathcal{G}$.

# Topological Ordering (4/4)

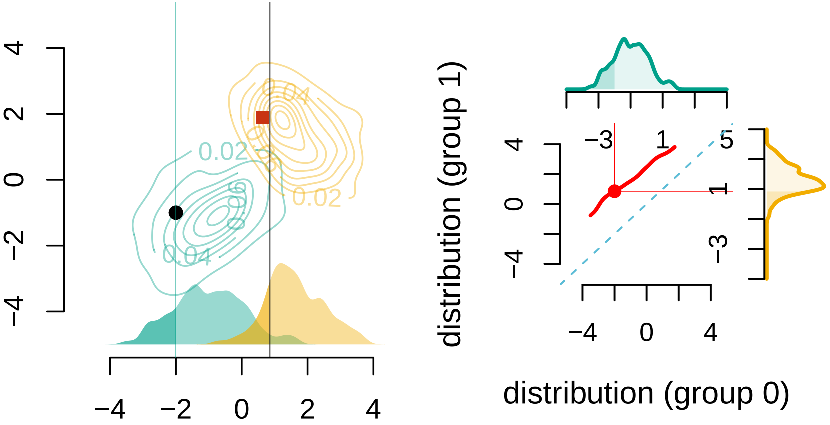



Causal graph in the German Credit dataset from Watson et al. (2021).

- **Step 2**: sequential conditional transport based on a topological ordering:

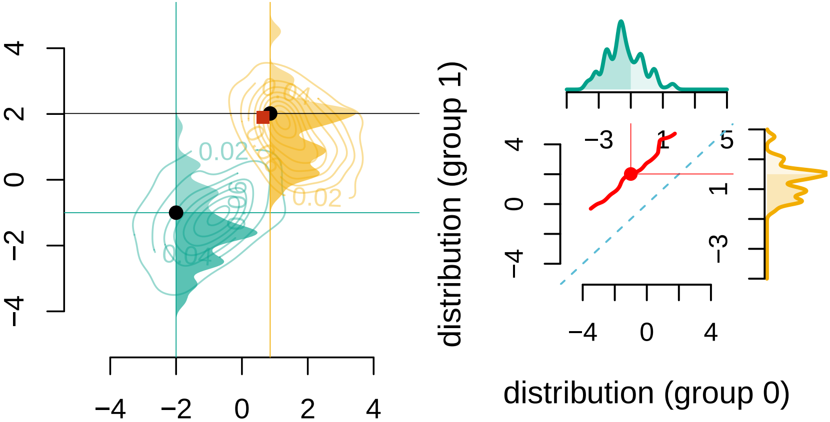
$$T^\star_{\mathcal{G}}(x_1, \cdots, x_7) = \begin{pmatrix} T^\star_1(x_1) \\ T^\star_2(x_2|x_1) \\ T^\star_3(x_3|x_1, x_2) \\ T^\star_4(x_4|x_2, x_3) \\ T^\star_5(x_5|x_1, x_2, x_4) \\ T^\star_6(x_6|x_3, x_5) \\ T^\star_7(x_7|x_1, x_4, x_5, x_6) \end{pmatrix}.$$

# Transport $x_1 \mid s$ From Group 0 to Group 1




Sequential Transport (simulated data). Red square: multivariate OT. **transport $x_1 \mid s$**.

# Transport $x_2 \mid x_1, s$ From Group 0 to Group 1




Sequential Transport (simulated data). Red square: multivariate OT. **transport** $x_2 \mid x_1, s$

# Code

This can be easily done with our R functions from our small package:

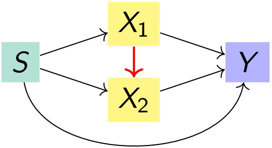
```
remotes::install_github(
  repo = "fer-agathe/sequential_transport", subdir = "seqtransfairness")
library(seqtransfairness)
sim_dat <- simul_dataset() # Simulate data
variables <- c("S", "X1", "X2", "Y")
adj <- matrix(
  # S  X1 X2 Y
  c(0, 1, 1, 1,# S
    0, 0, 1, 1,# X1
    0, 0, 0, 1,# X2
    0, 0, 0, 0  # Y
  ),
  ncol = length(variables), byrow = TRUE
  dimnames = rep(list(variables), 2))
# Sequential transport according to the causal graph
transported <- seq_trans(data = sim_dat, adj = adj, s = "S", S_0 = 0, y = "Y")
predict(transported) # Transp. values from S=0 to S=1, using the causal graph.
```

# Interpretable Counterfactual Fairness

Now, assume a logistic regression model was fitted on the simulated data and returned scores according to:

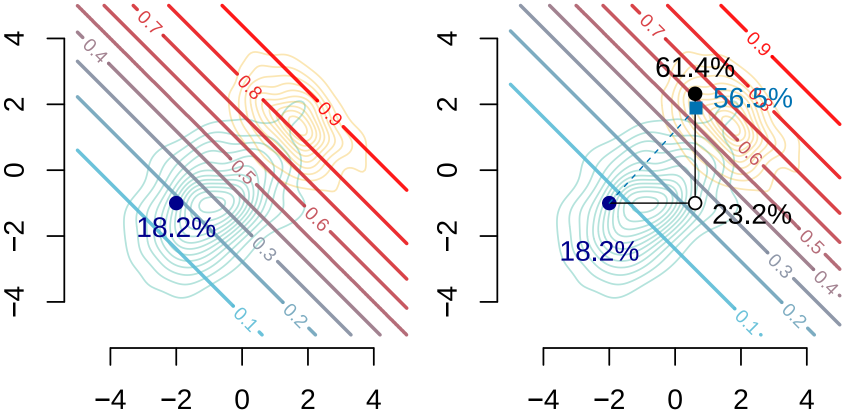
$$m(x_1, x_2, s) = \left(1 + \exp\left[-\left((x_1 + x_2)/2 + \mathbf{1}(s = 1)\right)\right]\right)^{-1}.$$

| | | | |
|---|---|---|---|
| Observation: | $(\mathbf{s{=}0}, x_1 = -2, x_2 = -1)$ | | |
| **Prediction** | : | $m(0, -2, -1)$ | $= 18.24\%.$ |
| **Pred. with Seq. T** | : | $m(s = 1, x_1^\star, x_2^\star)$ | $= 61.4\%$ |
| **Pred with OT** | : | $m(s = 1, x_1^\star, x_2^\star)$ | $= 56.5\%$ |




Assumed causal structure.

# Counterfactual assuming $X_2$ is caused by $X_1$



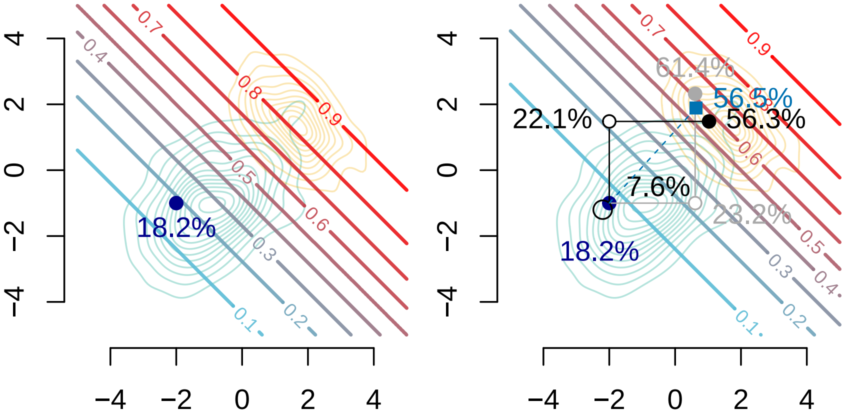

Predictions by $m$ of: the **observation** using factual (left), counterfactual (right): **counterfactual by Seq. T.** (assuming $X_1 \rightarrow X_2$) and **optimal. transport**.

# Decomposition of the *mutatis mutandis* difference

The ***mutatis mutandis* difference can be decomposed:**

$$
\begin{aligned}
& m(s=1, x_1^\star, x_2^\star) - m(s=0, x_1, x_2) = +43.16\% \ (\textit{\textbf{mutatis mutandis}}\ \textbf{diff.})\\
= \ & m(s=1, x_1, x_2) - m(s=0, x_1, x_2) \quad : -10.66\% \ (\textit{\textbf{cet. par. diff.}})\\
+ \ & m(s=1, x_1^\star, x_2) - m(s=1, x_1, x_2) \quad : +15.63\% \ (\textbf{change in } x_1)\\
+ \ & m(s=1, x_1^\star, x_2^\star) - m(s=1, x_1^\star, x_2) \quad : +38.18\% \ (\textbf{change in } x_2 | x_1^\star).
\end{aligned}
$$

# Counterfactual assuming $X_1$ is caused by $X_2$




Predictions by $m$ of: the **observation** using factual (left), counterfactual (right): **counterfactual by Seq. T.** (assuming $X_2 \rightarrow X_1$) and **optimal. transport** .

# Counterfactuals for Categorical Data

# What About Transporting Categorical Data?

- So far, to build **counterfactuals**, we have mentioned a quantile interpretation when the characteristics to transport, $\mathbf{x}$ is **univariate**.
- In **higher dimensions**:
  - Quantile interpretation (Hallin et al., 2021; Hallin and Konen, 2024)
  - *Mutatis mutandis* with DAGs (Plečko and Meinshausen, 2020; Plečko et al., 2024)
  - or with OT (Black et al., 2020; Charpentier et al., 2023; De Lara et al., 2021)
  - or with sequential transport (as previously shown).
- How can we handle **categorical data**? *What would have been the **marital status** of this woman, had she been a man?*
- In Fernandes Machado et al. (2025), we suggest a method based on **transporting** the values of categorical data represented in the **simplex**.

# In a Nutshell

Consider a **categorical feature** $\mathbf{x}_j \in \{x_{j,1}, \ldots, x_{j,d_j}\}$ ($d_k$ categories)
Which can also be denoted, $\mathbf{x}_j \in [\![d_j]\!]$, with $[\![d_j]\!] = \{1, \cdots, d_j\}$.

Our suggested methodology, in two steps:

1. Learn a mapping from $\mathcal{X}_{-x}$ (all other features) to $\mathcal{S}_d$ (and not the usual $[\![d_j]\!]$), using a **probabilistic classifier**.
2. Build counterfactuals for the data in $\mathcal{S}_d$, using **optimal transport**,

Counterfactuals for Categorical Data

## Step 1: Categorical Data to Compositional Data

# Categorical Data to Compositional Data

To predict the labels of $x$, a probabilistic classifier learns a mapping:

$$T : \mathcal{X}_{-x} \to \mathcal{S}_d.$$

For a multinomial logistic regression model, with a **softmax loss function**:

$$\widehat{T}(\mathbf{x}) = \mathcal{C}(1, e^{\mathbf{x}^\top \widehat{\beta}_2}, \cdots, e^{\mathbf{x}^\top \widehat{\beta}_d}) \in \mathcal{S}_d,$$

where $\widehat{\beta}_2, \ldots, \widehat{\beta}_d$ are the **estimated coefficients for each category** (first category taken as reference), and where $\mathcal{C} : \mathbb{R}_+^d \to \mathcal{S}_d$ is the closure operator:

$$\mathcal{C}(\mathbf{x}) = \frac{\mathbf{x}}{\mathbf{x}^\top 1},$$

with $1^\top$ a row vector of ones.

Counterfactuals for Categorical Data

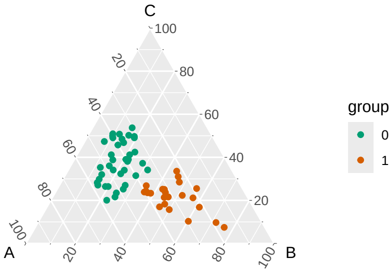
## Step 2: Gaussian Case in the Euclidean Representation

# Normal Distribution on the Simplex

Let $\mathbf{X}_0$ and $\mathbf{X}_1$ be random vectors taking values in $\mathcal{S}_3$, both following a "**normal distribution on the simplex**".

## Definition

$\mathbf{X} \in \mathcal{S}_d$ follow a "normal distribution on the simplex" if, for some isomorphism $h$, the vector of orthonormal coordinates $\mathbf{Z} = h(\mathbf{X})$ follows a multivariate normal distribution in $\mathbb{R}^{d-1}$.




Toy data, $n = 61$ points in $\mathcal{S}_3$.

# Optimal Mapping

- We consider, e.g., the center log ration transform ($h =$clr):

$$\text{clr}(\mathbf{x}) = \left[\log \frac{x_1}{\overline{\mathbf{x}}_g}, \cdots, \log \frac{x_D}{\overline{\mathbf{x}}_g}\right],$$

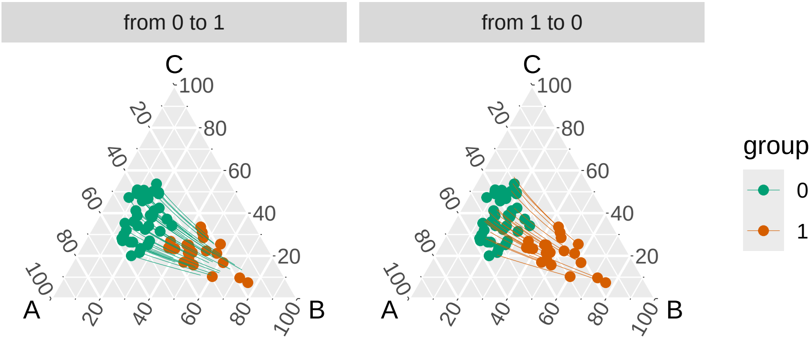
  where $\overline{\mathbf{x}}_g$ denotes the geometric mean of $\mathbf{x}$.

- Hence, we have $\mathbf{Z}_0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $\mathbf{Z}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$.
- The **optimal mapping** writes:

$$\mathbf{z}_1 = T^\star(\mathbf{z}_0) = \boldsymbol{\mu}_1 + \boldsymbol{A}(\mathbf{z}_0 - \boldsymbol{\mu}_0),$$

  where $\boldsymbol{A}$ is a symmetric positive matrix that satisfies $\boldsymbol{A}\boldsymbol{\Sigma}_0\boldsymbol{A} = \boldsymbol{\Sigma}_1$, which has a unique solution: $\boldsymbol{A} = \boldsymbol{\Sigma}_0^{-1/2}\left(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2}\right)^{1/2}\boldsymbol{\Sigma}_0^{-1/2}$.

# Gaussian Transport




Counterfactuals using the clr transformation and Gaussian optimal transports, $\mu_0 \mapsto \mu_1$ (left), and $\mu_1 \mapsto \mu_0$ (right)

Counterfactuals for Categorical Data

## Step 2: **Optimal Transport for Measured on the Simplex**

# Optimal Transport on the Simplex

- Instead of using an isomorphism to represent the data in the Euclidean space an then apply OT, we can apply OT for measures on $\mathcal{S}_d$ using a proper **cost function**.
- In the unit simplex, the Monge-Kantorovitch optimal transport problem can be expressed using the following cost function Pal and Wong (2020):

$$c(\mathbf{x}, \mathbf{y}) = \log\left(\frac{1}{d}\sum_{i=1}^{d}\frac{y_i}{x_i}\right) - \frac{1}{d}\sum_{i=1}^{d}\log\left(\frac{y_i}{x_i}\right)$$

# Optimal Transport on the Simplex

The discrete version of the Monge-Kantorovich problem writes:

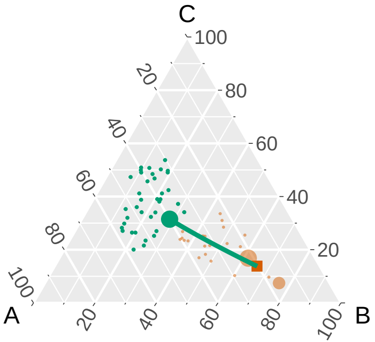
$$\min_{\mathrm{P}\in U(n_0,n_1)} \left\{ \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \underbrace{\mathrm{P}_{i,j}}_{\text{weights}} \underbrace{\mathrm{C}_{i,j}}_{n_0 \times n_1 \text{ cost matrix } \mathbf{C}_{ij} = c(\mathbf{x}_i, \mathbf{x}_j)} \right\}$$

with $U(n_0, n_1)$ the set of $n_0 \times n_1$ matrices (convex transportation polytope):

$$U(n_0, n_1) = \left\{ \mathrm{P} : \mathrm{P}\mathbf{1}_{n_1} = \mathbf{1}_{n_0} \text{ and } \mathrm{P}^\top \mathbf{1}_{n_0} = \frac{n_0}{n_1}\mathbf{1}_{n_1} \right\},$$

$n_0$, $n_1$: number of observations in group 0 and in group 1.

# Counterfactual




Empirical counterfactual of $\mathbf{x}_{0,3}$ (orange square) and path to the counterfactual obtained with Gaussian optimal transport on the simplex (shown with the line).

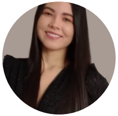
# Conclusion

- Without addressing algorithmic fairness issues: having fair model is illusive.
- Addressing fairness using a sequential approach provides an explainable method.
- We suggest using optimal transport on the simplex to build counterfactuals for categorical data.

Agathe
Fernandes Machado

Arthur
Charpentier

Marouane
Il Idrissi



Ana María
Patrón Piñerez

# References I

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network flows: Theory, algorithms, and applications*. Prentice Hall.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *The American Economic Review*, 53(5):941–973.

Avraham, R. (2017). *Discrimination and Insurance*, page 335–347. Routledge.

Baldus, D. C. and Cole, J. W. (1980). Statistical proof of discrimination. *(No Title)*.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. Adaptive Computation and Machine Learning series. MIT Press.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago press.

Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121.

Bonnotte, N. (2013). From Knothe's rearrangement to Brenier's optimal transport map. *SIAM Journal on Mathematical Analysis*, 45(1):64–87.

# References II

Campbell, C. and Smith, D. (2023). Distinguishing between direct and indirect discrimination. *The Modern Law Review*, 86(2):307–330.

Charpentier, A., Flachaire, E., and Gallic, E. (2023). Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal Transport Statistics for Economics and Related Topics*, pages 45–89. Springer.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Commission, E. (2011). Press release – eu rules on gender-neutral pricing in insurance industry enter into force. https://ec.europa.eu/commission/presscorner/detail/en/ip_12_1430. Accessed: 2010-09-30.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

De Lara, L., González-Sanz, A., Asher, N., and Loubes, J.-M. (2021). Transport-based counterfactual models.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4):1–36.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226. ACM.

# References III

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a). Post-calibration techniques: Balancing calibration and score distribution alignment. In *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b). Probabilistic scores of classifiers, calibration is not enough.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2024c). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025). Optimal transport on categorical data for counterfactuals using compositional data and dirichlet transport.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

Fourcade, M. and Healy, K. (2024). *The ordinal society*. Harvard University Press-T.

Frezal, S. and Barry, L. (2019). Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, 167(1):127–136.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459.

# References IV

Hallin, M., Del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165.

Hallin, M. and Konen, D. (2024). Multivariate quantiles: Geometric and measure-transportation-based contours. In *Applications of Optimal Transport to Economics and Related Topics*, pages 61–78. Springer.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323.

Harsanyi, J. C. (1959). *17. A Bargaining Model for the Cooperative n-Person Game*, page 325–356. Princeton University Press.

Hellman, D. (2008). When is discrimination wrong?

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017a). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

# References V

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017b). Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 656–666, Red Hook, NY, USA. Curran Associates Inc.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. NIPS.

Lamont, O. A. and Thaler, R. H. (2003). Anomalies: The law of one price in financial markets. *Journal of Economic Perspectives*, 17(4):191–202.

Landes, X. (2014). How fair is actuarial fairness? *Journal of Business Ethics*, 128(3):519–533.

Larson, Jeff ans Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704.

Nash, J. F. (1950). The bargaining problem. *Econometrica*, 18(2):155–162.

Pal, S. and Wong, T.-K. L. (2020). Multiplicative schrödinger problem and the dirichlet transport. *Probability Theory and Related Fields*, 178(1):613–654.

# References VI

Pearl, J. (2000). Comment. *Journal of the American Statistical Association*, 95(450):428–431.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD08, page 560–568. ACM.

Plečko, D., Bennett, N., and Meinshausen, N. (2024). fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing.

Scalia, A. (1979). The disease as cure: In order to get beyond racism, we must first take account of race. *Wash. ULQ*, page 147.

Schauer, F. (2006). *Profiles, probabilities, and stereotypes*. Harvard University Press.

# References VII

Shapley, L. S. (1953). A value for n-person games. *Contribution to the Theory of Games*, 2.

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stan. JCR & CL*, 11:45.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., and Steyerberg, E. W. (2019). Calibration: the achilles heel of predictive analytics. *BMC Medicine*, 17(1).

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Society.

Villani, C. (2009). *Optimal Transport*. Springer Berlin Heidelberg.

von Mises, R. (1957). *Probability, Statistics and Truth*. George Allend and Unwin Ltd. Second revised English Edition prepared by Hilda Geiringer.

Walters, M. A. (1981). Risk classification standards. In *Proceedings of the Casualty Actuarial Society*, volume 68, pages 1–18.

Watson, D. S., Gultchin, L., Taly, A., and Floridi, L. (2021). Local explanations via necessity and sufficiency: unifying theory and practice. In de Campos, C. and Maathuis, M. H., editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1382–1392. PMLR.