Mathematics Meets Morality: Fairness Through a Mathematical Lens

Ewen Gallic

joint with Arthur Charpentier and Agathe Fernandes Machado

Les séminaires universitaires en mathématique à Montréal, 10 Janvier 2025

An Introductory Example: Risk Prediction with a Poisson Model

- 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.

An Introductory Example: Risk Prediction with a Poisson Model

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

Let us 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.

An Introductory Example: Risk Prediction with a Poisson Model

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:

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

$\times e^{\beta_3}$ ceteris paribus

If β_3 is small, $e^{\beta_3} \approx 1 + \beta_3$. Thus, if $\beta_3 = 0.2$, it corresponds to $+20\%$ for men.

An Introductory Example: Risk Prediction with a Poisson Model

- In the previous example, the estimates indicate that men are at higher risks than women.
- 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?

A Sketch of Insurance Business

Assume the following overly simplistic situation (adapted form Landes, 2014):

- A pool of insured made of 10 people: **5 women** and **5 men**.
- **Equal individual probability** of having an accident in the upcoming year of 10%.
- In the event of an accident, the insurance will pay the insured $1,000.

Actuarial Fairness

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

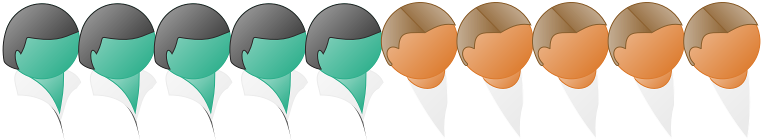
X. Landes (✉)
University of Copenhagen, Copenhagen, Denmark
e-mail: xavier.landes@gmail.com

[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

Expected Annual Loss



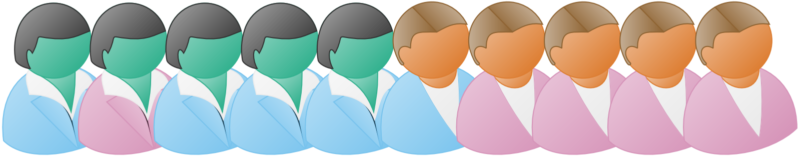
10% 10% 10% 10% 10% 10% 10% 10% 10% 10%

In such a situation, the **expected global loss** is: $10 \times .1 \times \$1,000 = \$1,000$.

- Since all the individuals have **equal risks**, they should be charged with a \$100 premium each.
- For questions on fair allocation in Game Theory, see, e.g., Nash et al. (1950); Shapley (1953); Harsanyi (1959)

Unequal Risks

Now, assume that among the 10 insured, 5 of them engage in **riskier driving behaviors** (speeding, aggressive overtaking) which doubles their **probability of having an accident**.



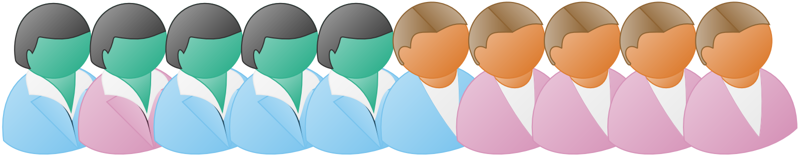
10% 20% 10% 10% 10% 10% 20% 20% 20% 20%

The **expected annual loss** becomes: $(5 \times .1 + 5 \times .2) \times \$1,000 = \$1,500$

Unequal Risks

- If premiums remain at $100, the insurance company will not be able to indemnify the unlucky drivers: the collected resources will be insufficient to cover the losses.
- How should the premiums be adjusted?
- If the premium is increased by $50 for each insured:
  - Actuarially sound, account for the general increase in risk exposure in the population.
  - But, **actuarially unfair**: low-risk drivers subsidize high-risks.
  - May entail **moral hazard** and **adverse selection** (Akerlof, 1978)

Actuarially Fair Prices



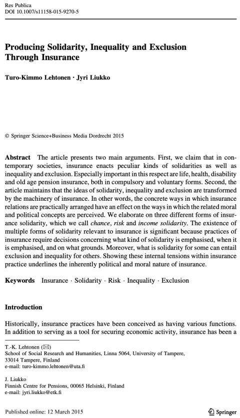
10% 20% 10% 10% 10% 10% 20% 20% 20% 20%

- The premium paid by **low-risk** drivers may remain unchanged ($100) but be doubled ($200) for **high-risk** drivers:
  - Actuarially sound solution: risk-based, allows the insurer to cover the expected annual loss.
  - Actuarially fair solution:
    - individuals with similar risk levels pay similar amounts (**horizontal equity**),
    - those with higher risks pay correspondingly higher premiums (**vertical equity**).

Risk Classification

In that previous toy example, the insurer needs to correctly **evaluate risk** and **risk classification**.

"*the ratio between risk and premiums should be exactly the same for all members of the pool. Those with lower risk also pay less. Behind this idea, there is the* ***technical capacity to calculate levels of risk for categories of insureds****. If taken to its extreme, risk classification could mean that each insured could constitute his or her own separate risk class. Still, in most forms of insurance, whether private or social,* ***premium levels are allocated to large groups of people, or risk classes****.*" (Lehtonen and Liukko, 2015)



Res Publica
DOI 10.1007/s11158-015-9270-5

Producing Solidarity, Inequality and Exclusion Through Insurance

Turo-Kimmo Lehtonen · Jyri Liukko

Abstract The article presents two main arguments. First, we claim that in contemporary societies, insurance enacts peculiar kinds of solidarities as well as inequality and exclusion. Especially important in this respect are life, health, disability and old age pension insurance, both in compulsory and voluntary forms. Second, the article maintains that the ideas of solidarity, inequality and exclusion are transformed by the machinery of insurance. In other words, the concrete ways in which insurance relations are practically arranged have an effect on the ways in which the related moral and political concepts are perceived. We elaborate on three different forms of insurance solidarity, which we call *chance*, *risk* and *income solidarity*. The existence of multiple forms of solidarity relevant to insurance is significant because practices of insurance require decisions concerning what kind of solidarity is emphasised, when it is emphasised, and on what grounds. Moreover, what is solidarity for some can entail exclusion and inequality for others. Showing these internal tensions within insurance practice underlines the inherently political and moral nature of insurance.

Keywords Insurance · Solidarity · Risk · Inequality · Exclusion

Introduction

Historically, insurance practices have been conceived as having various functions. In addition to serving as a tool for securing economic activity, insurance has been a

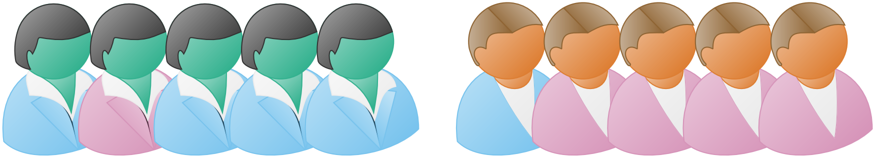
T.-K. Lehtonen (✉)
School of Social Research and Humanities, Linna 5064, University of Tampere, 33014 Tampere, Finland
e-mail: turo-kimmo.lehtonen@uta.fi

J. Liukko
Finnish Centre for Pensions, 00065 Helsinki, Finland
e-mail: jyri.liukko@etk.fi

Published online: 12 March 2015 Springer

Risk Classification

Now, let us assume that the insurer no longer observe risk, and uses **gender** to estimate the risk of the insured. They obtain the following estimates:



10%	20%	10%	10%	10%	10%	20%	20%	20%	20%

$\underbrace{\hphantom{10\%\ 20\%\ 10\%\ 10\%\ 10\%}}_{\text{Estimated risk: } 12\%}$ $\underbrace{\hphantom{10\%\ 20\%\ 20\%\ 20\%\ 20\%}}_{\text{Estimated risk: } 18\%}$

Risk Classification

- The pool of insured is thus segmented into two groups:
  - **Women**: low-risk, with a \$120 premium,
  - **Men**: high-risk, with \$180 premium.
- While this solution allows the insurer to cover the expected annual loss, it is **no longer actuarially fair**: individuals in each segment do not pay a premium according to their risk.
- Note that if the insurer knew the individual risks and still decided to charge women a \$120 premium and men a \$180 premium, this would correspond to an equalization of pool members' risk premiums, also termed **risk solidarity** in Lehtonen and Liukko (2015).

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****.*"

Policymakers Point of View: Colorado

- The Colorado Division of Insurance issued a regulation (effective November 14, 2023) titled: "Governance and risk management framework requirements for life insurers' use of external consumer data and information sources, algorithms, and predictive models". Section 5-A. writes:

  Life insurers that use ECDIS [External Consumer Data and Information Source], as well as ***algorithms and predictive models*** *that use ECDIS in any insurance practice, must establish a risk-based governance and risk management framework that facilitates and supports policies, procedures, systems, and controls designed to determine* ***whether the use of such ECDIS****, algorithms, and predictive models potentially* ***result in unfair discrimination with respect to race and remediate unfair discrimination****, if detected.*

Policymakers Point of View: Definition of (Un)fair Discrimination

- Colorado Revised Statutes (10-3-1104.9):

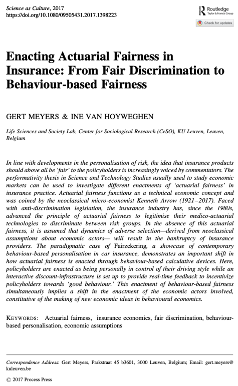
"*'**Unfairly discriminate**' and '**unfair discrimination**' include the use of one or more external consumer data and information sources, as well as algorithms or predictive models using external consumer data and information sources, that have a **correlation to** race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression, and **that use results in a disproportionately negative outcome** for such classification or classifications, which negative outcome exceeds the reasonable correlation to the underlying insurance practice, including losses and costs for underwriting.*"

Is Risk Classification Fair?

One might ask if discriminating based on gender in our toy example is fair or not.

- On the one hand, **governments** enacted legislation **prohibiting insurance discrimination** based on some **protected characteristics**
- On the other hand, **insurers** argue they need to know people's risk in advance.

"*governments must recognise that there is a difference between unfair discrimination and insurers differentiating prices according to risk,*" (Swiss Re, 2015 through Meyers and Van Hoyweghen, 2017)



Science as Culture, 2017
https://doi.org/10.1080/09505431.2017.1398223

Routledge
Taylor & Francis Group

Check for updates

Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-based Fairness

GERT MEYERS & INE VAN HOYWEGHEN

Life Sciences and Society Lab, Center for Sociological Research (CeSO), KU Leuven, Leuven, Belgium

In line with developments in the personalisation of risk, the idea that insurance products should above all be 'fair' to the policyholders is increasingly voiced by commentators. The performativity thesis in Science and Technology Studies usually used to study economic markets can be used to investigate different enactments of 'actuarial fairness' in insurance practice. Actuarial fairness functions as a technical economic concept and was coined by the neoclassical micro-economist Kenneth Arrow (1921–2017). Faced with anti-discrimination legislation, the insurance industry has, since the 1980s, advanced the principle of actuarial fairness to legitimise their medico-actuarial technologies to discriminate between risk groups. In the absence of this actuarial fairness, it is assumed that dynamics of adverse selection—derived from neoclassical assumptions about economic actors— will result in the bankruptcy of insurance providers. The paradigmatic case of Fairzekering, *a showcase of contemporary behaviour-based personalisation in car insurance, demonstrates an important shift in how actuarial fairness is enacted through behaviour-based calculative devices. Here, policyholders are enacted as being personally in control of their driving style while an interactive discount-infrastructure is set up to provide real-time feedback to incentivize policyholders towards 'good behaviour.' This enactment of behaviour-based fairness simultaneously implies a shift in the enactment of the economic actors involved, constitutive of the making of new economic ideas in behavioural economics.*

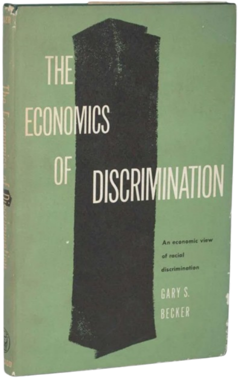
KEYWORDS: Actuarial fairness, insurance economics, fair discrimination, behaviour-based personalisation, economic assumptions

Correspondence Address: Gert Meyers, Parkstraat 45 b3601, 3000 Leuven, Belgium; Email: gert.meyers@kuleuven.be

Fair Discrimination in Insurance: an Oxymoron

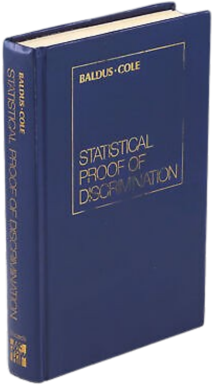
"*what is unique about insurance is that even statistical discrimination (the act by which an insurer uses a characteristic of an insured or potential insured as a statistic for the risk it poses to an insurer), which by definition is absent any malicious intentions, poses significant moral and legal 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)

Some Definitions: Discrimination



- 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.

Some Definitions: Statistical Discrimination



- **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.

Some Definitions: 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).

Actuarial Fairness and Accuracy

- 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)

Actuarial Fairness and Accuracy

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

average loss / empirical losses

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

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.

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)

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)

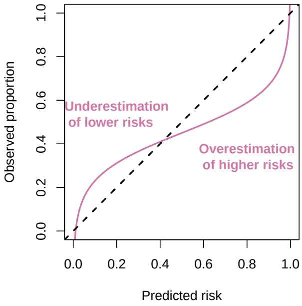
- If 40 out of 100 in this group are found to have the disease, the risk is **underestimated**.
- If 10 out of 100 in this group are found to have the disease, the risk is **overestimated**.

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 $\hat{y} = 20\%$



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)

Quantifying Unfairness

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

How Can Fairness be Quantified?

As mentioned earlier, algorithmic fairness can be defined in multiple ways (see Veale and Binns, 2017 for a brief overview, or Charpentier, 2024).

- Most metrics focus on differences in treatment between **protected** and **non-protected** groups.
- Here, we focus on three metrics: **demographic parity**, **equalized odds**, and **calibration**.
- Individual fairness will be briefly mentioned later.

Group Fairness: Demographic Parity

A model m satisfies the **independence property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ (Dwork et al., 2012).

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

Group Fairness: Equalized Odds

A model m satisfies the **separation property** if $m(\boldsymbol{X}, S) \perp\!\!\!\perp S \mid Y$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ (Hardt et al., 2016).

$$\textbf{Equalized Odds} \rightarrow \mathbb{E}[\underbrace{\overbrace{\hat{Y}}}_{\text{score } \hat{y}} \mid \overbrace{Y = y}^{\text{outcome } y}, S = A] \overset{?}{=} \mathbb{E}[\hat{Y} \mid Y = y, S = B], \forall y$$

Group Fairness: Calibration

A model m satisfies the **sufficiency property** if $Y \perp\!\!\!\perp S \mid m(\boldsymbol{X}, S)$, with respect to the distribution $\mathbb{P}$ of the triplet $(\boldsymbol{X}, S, Y)$ (Chouldechova, 2017).

outcome y

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

score $\hat{y}$

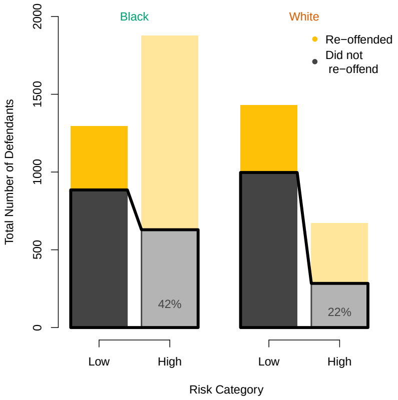
Illustration With the COMPAS Dataset

- 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**.

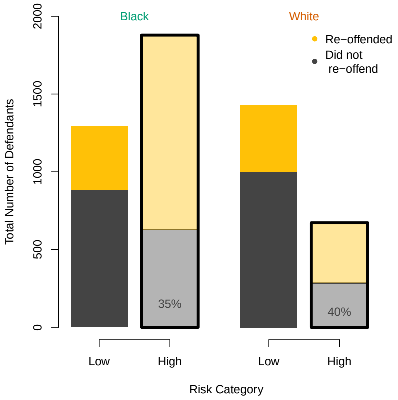


$$\mathbb{P}[\hat{Y} = \text{High} \mid Y = \text{no}, S = \text{Black}] = 42\% \overset{?}{=} \mathbb{P}[\hat{Y} = \text{High} \mid Y = \text{no}, S = \text{White}] = 22\%$$

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% ≈ 40%: **fair**.



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

Mitigation

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** (not addressed here): modify algorithms to embed fairness constraints during training.
3. **Postprocessing**: alter models after training to correct unfair outcomes.

Group Fairness: 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** τ:

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

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.

For a Scoring Classifier

When facing a **score** rather than a binary decision:

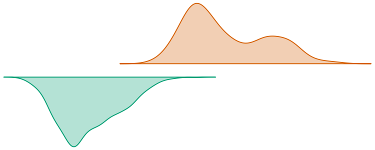
$$\textbf{Demographic Parity} \rightarrow \mathbb{P}[\underbrace{\hat{m}(\boldsymbol{X}, S)}_{\text{score } \hat{y}} \mid S = A] \overset{?}{=} \mathbb{P}[\underbrace{\hat{m}(\boldsymbol{X}, S)}_{\text{score } \hat{y}} \mid S = B]$$

We can look at the **quantile level** of that score in the **protected group** and replace it with the quantile at that level in the **unprotected group**.

This strategy corresponds to **transporting** the score from the **protected group** to the **unprotected one**.

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 μ_0 (resp. μ_1) on a metric space $\mathcal{X}_0$ (resp. $\mathcal{X}_1$). The goal is to move every elementary mass from μ_0 to μ_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 μ_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 μ_0 is atomless, then there exists T such that $\mu_1 = T_\# \mu_0$.

Definition: Monge problem, Monge (1781)

If we further assume μ_0 and μ_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 μ_0 and μ_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 $\mathbb{R}$.

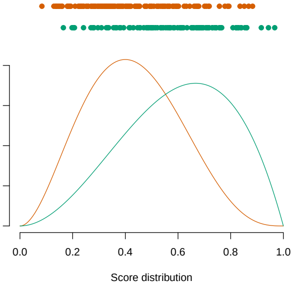
As shown in Santambrogio (2015), the optimal Monge map $T^\star$ for some strictly convex cost c such that $T^\star_{\#}\mu_0 = \mu_1$ is:

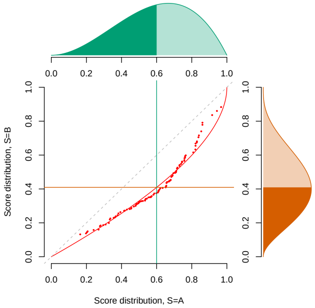
cdf associated with μ_0

$$T^\star = \boxed{F_1^{-1}} \circ \boxed{F_0},$$

generalized inverse (quantile function)

Mitigation for a Scoring Classifier (2)



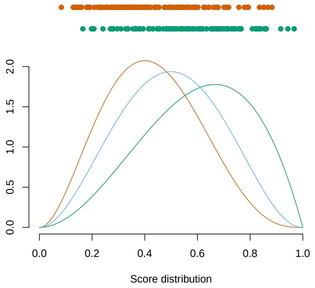


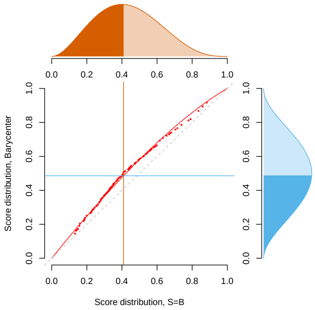
- Individual in group **A** with score $\hat{y}(A) = 60\%$: corresponding to quantile α (here $\hat{F}_{\hat{Y}|S=A}(.6) = .47$)
- In group **B**, this corresponds to $\hat{y}(B) = .41 = \hat{F}^{-1}_{\hat{Y}|S=B}(.47)$.

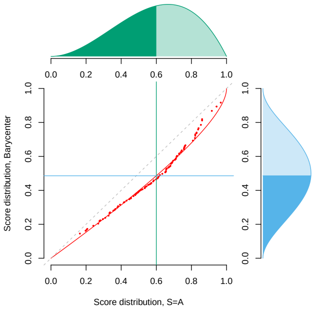
Mitigating Discrimination with (Wasserstein) Barycenters

To get a fair model w.r.t. the **sensitive attribute**, we can consider an average:

$$\hat{y}^{\star} = \mathbb{P}[S=A] \cdot \hat{y}(A) + \mathbb{P}[S=B] \cdot \hat{F}^{-1}_{\hat{Y}|S=B}\left[\hat{F}_{\hat{Y}|S=A}(\hat{y}(A))\right]$$







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?*" (De Lara et al., 2021; Charpentier et al., 2023; Fernandes Machado et al., 2024c)

Counterfactual Fairness in Brief: Links with Causal Inference

	Sex	Name	Treatment t_i	Weight (Outcome) y_i	$y^\star_{i,T\leftarrow A}$	$y^\star_{i,T\leftarrow B}$	TE	Height 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):

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

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:

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

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:

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

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., 2017).
- The ***mutatis mutandis*** version of the CATE writes:

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

- In this version, $\boldsymbol{X} \mid \mathbf{A}$ is transported to $\boldsymbol{X} \mid \mathbf{B}$ (see Plečko and Meinshausen, 2020; Plečko et al., 2024; De Lara et al., 2021; Charpentier et al., 2023), according to an assumed **causal structure**.

Fairness Without the Sensitive Attribute

Three Situations

We can consider three situations where the **sensitive attribute** is not fed to the model:

1. The variable is deliberately excluded from the model: **fairness through unawareness** $\rightarrow$ usually a bad idea (see the following example).
2. The sensitive attribute is **not observable**: we can try to infer it in a separate model: e.g., "Bayesian Improved Surname Geocoding" (BISG) algorithm (Elliott et al., 2009; Imai and Khanna, 2016).
3. Opting out: people decide to voluntarily prevent some of their characteristics to be used: may result in strong biases (not explored in this talk).

Why not Removing the Variable?

- Why not removing the **sensitive attribute** (e.g., **race**) and make the model **blind to it**?
  - If other variables in the model are correlated with it (**proxy variables**), the model may still exhibit disparities with respect to the sensitive attribute.
  - And in the context of "**big data**," it is easy to get proxies for the **sensitive** attributes.

y	urban	age	race
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮

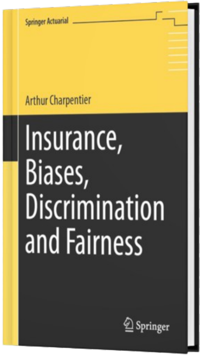
y	urban	age	zip	lastname	credit
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮

Illustration

Let illustrate this with an example from Charpentier (2024).

- On a French motor dataset, average claim frequencies are 8.94% (**men**), 8.20% (**women**).
- Consider some logistic regression to estimate annual claim frequency, on k explanatory variables **excluding gender**.

	Men	**Women**
$k = 0$	8.68%	8.68%
$k = 2$	8.85%	8.37%
$k = 8$	8.87%	8.33%
$k = 15$	8.94%	8.20%
empirical	**8.94%**	**8.20%**



Fairness With Uncollected Attribute

- Sometimes, the information about a sensitive attribute is not known by the modeler (often for legitimate reasons, such as privacy).
  - Race is often **infrequently or incompletely collected** by insurers (Haley et al., 2022).
- However, to assess the fairness of a model w.r.t. some sensitive attribute, access to that sensitive attribute is required:

"*What we can't measure, we can't understand.*" Andrus et al. (2021)

- **Bayesian methods for predicting race** have emerged (Elliott et al., 2009; Imai et al., 2022; Baeder et al., 2024), using surname, first name, and geolocation data from an aggregate source (the USA Census data).

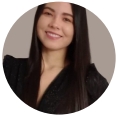
Conclusion

- Certain forms of discrimination, even if they have predictive value, are not socially acceptable.
- Protected attributes evolve with societal changes.
- Without addressing algorithmic fairness issues: having fair model is illusive.
- Not collecting and not using protected attributes is clearly not a good strategy.
- This field still requires substantial further research!

Agathe
Fernandes Machado

Arthur
Charpentier

Marouane
Il Idrissi



Ana María
Patrón Piñerez

References I

Akerlof, G. A. (1978). The market for “lemons”: Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pages 235–251. Elsevier.

Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can’t measure, we can’t understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 249–260, New York, NY, USA. Association for Computing Machinery.

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.

Baeder, L., Erica, B., Brinkmann, P., Long, J., Stracke, C., Togba-Doya, K., Usan, G., Weaver, N., and Woldeyes, M. (2024). Statistical methods for imputing race and ethnicity. Technical report.

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

References II

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.

Charpentier, A. (2024). *Insurance, Biases, Discrimination and Fairness*. Springer Verlag.

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*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163. PMID: 28632438.

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.

References III

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.

Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83.

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.

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

References IV

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.

Haley, J. M., Dubay, L., Garrett, B., Caraveo, C. A., Schuman, I., Johnson, K., Hammersla, J., Klein, J., Bhatt, J., Rabinowitz, D., et al. (2022). Collection of race and ethnicity data for use by health plans to advance health equity. *Working Paper*.

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?

Imai, K. and Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2):263–272.

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully bayesian improved surname geocoding and name supplements. *Science Advances*, 8(49):eadc9824.

References V

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. (2017). 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.

References VI

Lehtonen, T.-K. and Liukko, J. (2015). Producing solidarity, inequality and exclusion through insurance. *Res Publica*, 21(2):155–169.

Meyers, G. and Van Hoyweghen, I. (2017). Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438.

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

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

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

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.

References VII

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

Swiss Re (2015). Life insurance risk selection: Required differentiation or unfair discrimination? Technical report, Swiss Re.

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).

Veale, M. and Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530.

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.