

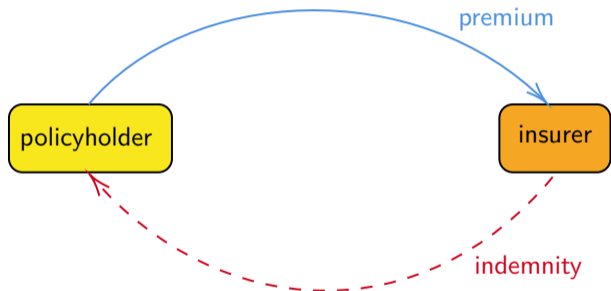
# Fairness and discrimination of predictive models, with an application in insurance

Agathe Fernandes-Machado, François Hu, Philipp Ratz & **Arthur Charpentier**

SDAI'24

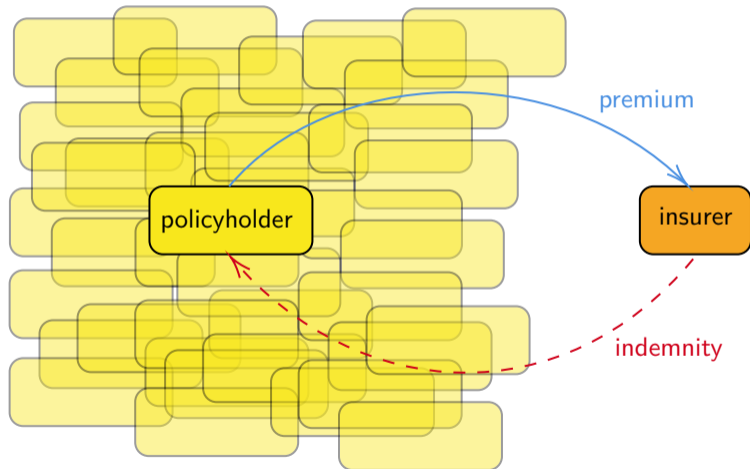
# Insurance (and “Actuarial Fairness”)

- Insurance is a **risk transfer** (from a policyholder to an insurance company)



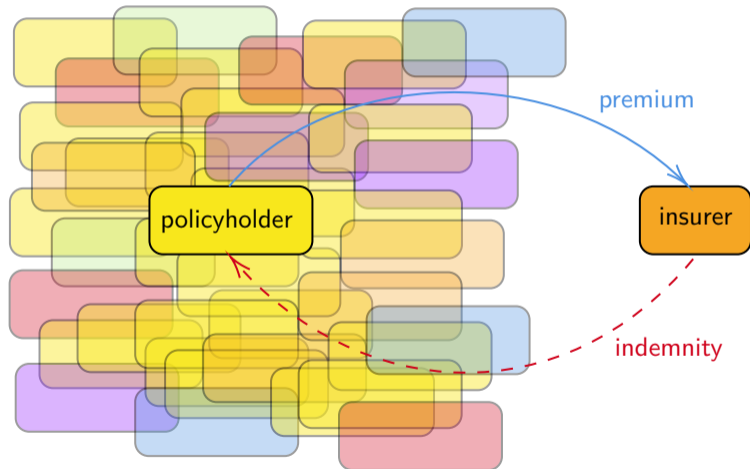
# Insurance (and “Actuarial Fairness”)

› *“Insurance is the contribution of the many to the misfortune of the few”*



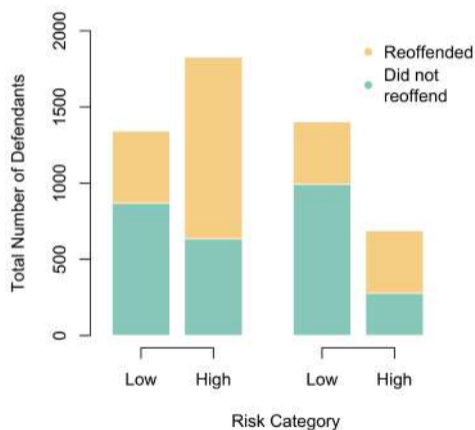
# Insurance (and “Actuarial Fairness”)

› *“Insurance is the contribution of the many to the misfortune of the few”*



# Motivation (1. Propublica, Actuarial Justice)

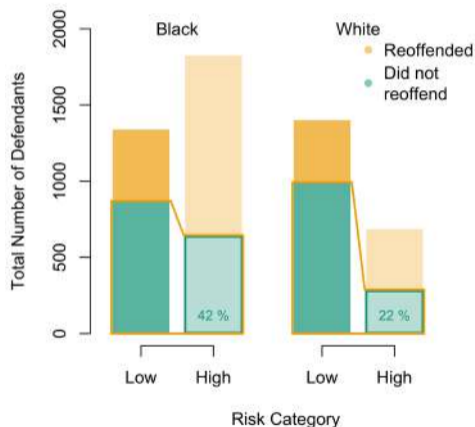
- › Concept of “**actuarial justice**” as coined in **Feeley and Simon (1994)**
- › **Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)**, **Perry (2013)**



- 🔗 <https://github.com/propublica/compas-analysis>
- › **Angwin et al. (2016)** Machine Bias  
**Dressel and Farid (2018)**

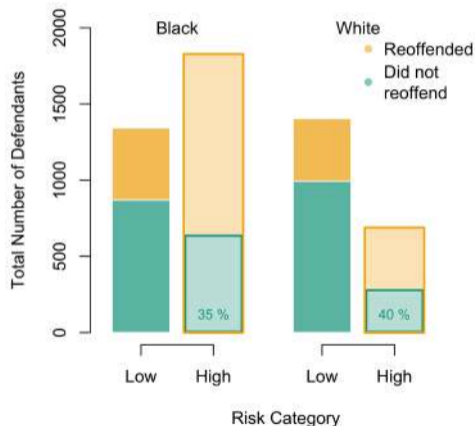
# Motivation (1. Propublica, Actuarial Justice)

- From Feller et al. (2016),
  - ▶ for White people, among those who did not re-offend, 22% were wrongly classified,
  - ▶ for Black people, among those who did not re-offend, 42% were wrongly classified,
  - ▶ problem, since 42%  $\gg$  22%



# Motivation (1. Propublica, Actuarial Justice)

- From Dieterich et al. (2016),
  - ▶ for White people, among those who were classified as high risk, 40% did not re-offend,
  - ▶ for Black people, among those who were classified as high risk, 35% did not re-offend,
  - ▶ no problem, since  $40\% \approx 35\%$



## Motivation (2. Legal Aspects)

› EU Directive ([2004/113/EC](#)), 2004 version

– Article 5 (Actuarial factors) –

1. Member States shall ensure that in all new contracts concluded after 21 December 2007 at the latest, **the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits.**

2. Notwithstanding paragraph 1, Member States may decide before 21 December 2007 to permit proportionate differences in individuals' premiums and benefits where the use of sex is a determining factor in the assessment of risk based on relevant and accurate actuarial and statistical data. The Member States concerned shall inform the Commission and ensure that accurate data relevant to the use of sex as a determining actuarial factor are compiled, published and regularly updated.



## Motivation (2. Legal Aspects)

- › Au 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.



## Motivation (2. Legal Aspects)

- › September 27, 2023, the Colorado Division of Insurance exposed a new proposed regulation entitled **Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes**

### – Section 5 (Estimating Race and Ethnicity) –

Insurers shall estimate the race or ethnicity of all proposed insureds that have applied for coverage on or after the insurer's initial adoption of the use of ECDIS, or algorithms and predictive models that use ECDIS, including a third party acting on behalf of the insurer that used ECDIS, or algorithms and predictive models that used ECDIS, in the underwriting decision-making process, by utilizing: BIFSG and the insureds' or proposed insureds' name and geolocation (...)

- › **Bayesian Improved First Name Surname Geocoding**, or “BIFSG”
- › **External Consumer Data and Information Source**, or “ECDIS”



## Motivation (2. Legal Aspects)

› EU Directive ([2010/41/EU](#)), 2010 version (on the application of the principle of equal treatment between men and women)

– Article 3 (Definition) –

(a) ‘**direct discrimination**’: where one person is treated less favourably on grounds of sex than another is, has been or would be, treated in a comparable situation;

(b) ‘**indirect discrimination**’: where an apparently neutral provision, criterion or practice would put persons of one sex at a particular disadvantage compared with persons of the other sex, unless that provision, criterion or practice is objectively justified by a legitimate aim, and the means of achieving that aim are appropriate and necessary;



## Motivation (2. Legal Aspects)

➤ In France, Loi n° 2008-496 du 27 mai 2008

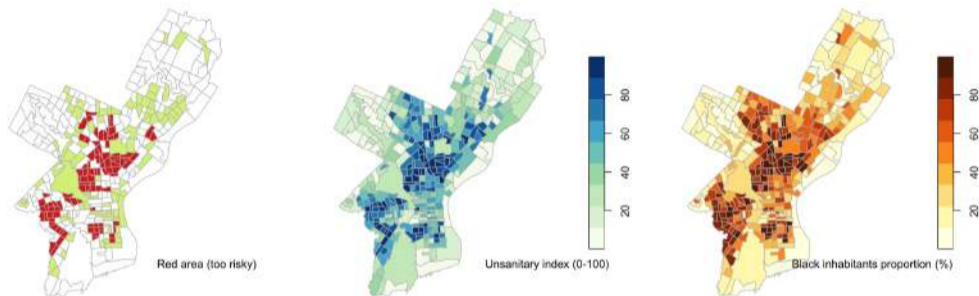
– Article 1 –

Constitue une **discrimination indirecte** une disposition, un critère ou une pratique neutre en apparence, mais susceptible d'entraîner, pour l'un des motifs mentionnés au premier alinéa, un désavantage particulier pour des personnes par rapport à d'autres personnes, à moins que cette disposition, ce critère ou cette pratique ne soit objectivement justifié par un but légitime et que les moyens pour réaliser ce but ne soient nécessaires et appropriés.

Extension of "Loi n° 72-546 du 1 juillet 1972", which removed the requirement for specific intent.



## Motivation (3. Redlining)



(Fictitious maps, inspired by a Home Owners' Loan Corporation map from 1937)

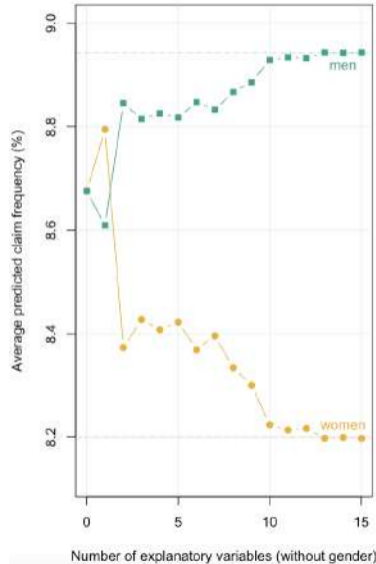
- ▶ Federal Home Loan Bank Board (FHLBB) "*residential security maps*" (for real-estate investments), [Crossney \(2016\)](#) and [Rhynhart \(2020\)](#)
- ▶ Unsanitary index and proportion of Black inhabitants
- ▶ Discrimination as an "**ill-posed problem**"?

## Motivation (4. Proxies)

- 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%

- Models simply tend to reproduce what was observed in the data (see “**is-ought**” problem, in Hume (1739)).



## Discrimination and Insurance

*“Machine learning won’t give you anything like gender neutrality ‘for free’ that you didn’t explicitly ask for,”* Kearns and Roth (2019)

*”What is unique about insurance is that even statistical discrimination which by definition is absent of 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)

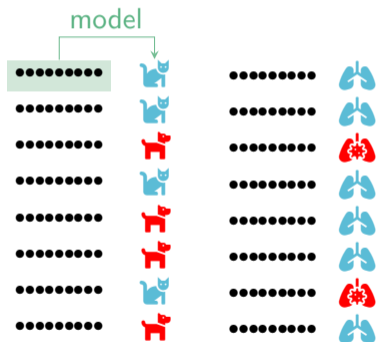
*“Technology is neither good nor bad; nor is it neutral,”* Kranzberg (1986)

# Classifiers (or why actuarial science $\neq$ computer science)

Classifiers on pictures,

→ 🐱 (cats) – 🐶 (dogs)

→ 🫁 (healthy) – 🦠 (sick)

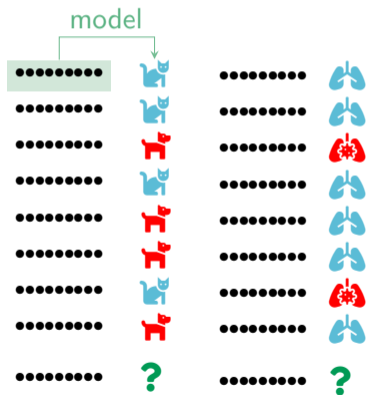


# Classifiers (or why actuarial science $\neq$ computer science)

Classifiers on pictures,



→ 🐱 (cats) – 🐶 (dogs)



→ 🫁 (healthy) – 🦠 (sick)

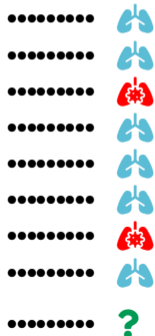
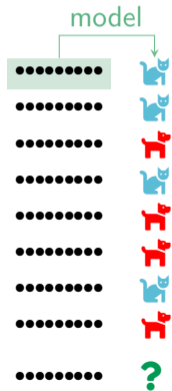


# Classifiers (or why actuarial science $\neq$ computer science)



Classifiers on pictures,



→  (cats) –  (dogs)



→  (healthy) –  (sick)

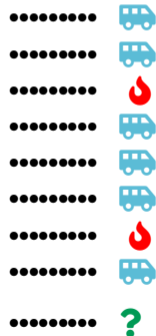


Classifiers, we need some “**probabilities**”

→  (sunny) –  (rainy)

→  (woman) –  (man)

→  (no claim) –  (accident)



## Fairness for Classifiers

$$\left\{ \begin{array}{l} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{array} \right.$$

# Fairness for Classifiers

$$\left\{ \begin{array}{l} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{array} \right.$$

class  $\in \{0, 1\}$

# Fairness for Classifiers

$$\left\{ \begin{array}{l} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{array} \right.$$

class  $\in \{0, 1\}$

score  $\in [0, 1] \subset \mathbb{R}$

# Fairness for Classifiers

$$\left\{ \begin{array}{l} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{array} \right.$$

Following Barocas et al. (2017), standard definitions are

A model  $m$  satisfies the **independence property** if  $m(\mathbf{X}, S) \perp\!\!\!\perp S$ , with respect to the distribution  $\mathbb{P}$  of the triplet  $(\mathbf{X}, S, Y)$  ← demographic parity

# Fairness for Classifiers

$$\left\{ \begin{array}{l} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{"sensitive variable"} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{array} \right.$$

Following Barocas et al. (2017), standard definitions are

A model  $m$  satisfies the **independence property** if  $m(\mathbf{X}, S) \perp\!\!\!\perp S$ , with respect to the distribution  $\mathbb{P}$  of the triplet  $(\mathbf{X}, S, Y)$  ← **demographic parity**

A model satisfies the **separation property** if  $m(\mathbf{X}, S) \perp\!\!\!\perp S \mid Y$ , with respect to the distribution  $\mathbb{P}$  of the triplet  $(\mathbf{X}, S, Y)$  ← **equalized odds**

# Fairness for Classifiers

$$\begin{cases} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d : \text{'explanatory' variables} \\ s \in \{A, B\} : \text{'sensitive variable'} \\ y \in \{0, 1\} : \text{classification problem} \\ \hat{y} \in \{0, 1\} : \text{prediction, classically } \hat{y} = \mathbf{1}(m(\mathbf{x}, s) > t) \end{cases}$$

class  $\in \{0, 1\}$

score  $\in [0, 1] \subset \mathbb{R}$

Following [Barocas et al. \(2017\)](#), standard definitions are

A model  $m$  satisfies the **independence property** if  $m(\mathbf{X}, S) \perp\!\!\!\perp S$ , with respect to the distribution  $\mathbb{P}$  of the triplet  $(\mathbf{X}, S, Y)$  ← **demographic parity**

A model satisfies the **separation property** if  $m(\mathbf{X}, S) \perp\!\!\!\perp S \mid Y$ , with respect to the distribution  $\mathbb{P}$  of the triplet  $(\mathbf{X}, S, Y)$  ← **equalized odds**

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

# Fairness for Classifiers

classical definition of “demographic parity” for a classifier

$$\mathbb{P}[\hat{Y} = 1 \mid S = A] \stackrel{?}{=} \mathbb{P}[\hat{Y} = 1 \mid S = B]$$

# Fairness for Classifiers

classical definition of “demographic parity” for a classifier

$$\mathbb{P}[\hat{Y} = 1 \mid \overset{\text{sensitive}}{\color{teal}S = A}] \stackrel{?}{=} \mathbb{P}[\hat{Y} = 1 \mid \overset{\text{sensitive}}{\color{orange}S = B}]$$

# Fairness for Classifiers

classical definition of “demographic parity” for a classifier

$$\mathbb{P}[\hat{Y} = 1 \mid S = A] \stackrel{?}{=} \mathbb{P}[\hat{Y} = 1 \mid S = B]$$

The diagram illustrates the classical definition of demographic parity for a classifier. It shows the equality of two conditional probabilities:  $\mathbb{P}[\hat{Y} = 1 \mid S = A]$  and  $\mathbb{P}[\hat{Y} = 1 \mid S = B]$ . The term  $\hat{Y} = 1$  is highlighted in a light brown box and labeled "class prediction" with a red arrow pointing to it from below. The term  $S = A$  is highlighted in a light green box and labeled "sensitive" with a green arrow pointing to it from above. The term  $S = B$  is highlighted in a light yellow box and labeled "sensitive" with a yellow arrow pointing to it from above. A red question mark is placed between the two probability expressions, indicating the equality to be tested.

# Fairness for Classifiers

classical definition of “demographic parity” for a classifier

$$\mathbb{E}[\hat{Y} \mid S = A] \stackrel{?}{=} \mathbb{E}[\hat{Y} \mid S = B]$$

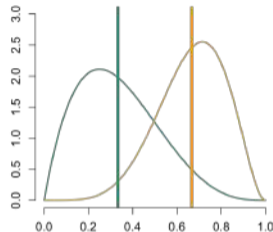
The diagram illustrates the classical definition of demographic parity for a classifier. It shows the equality of the expected class prediction for two different sensitive groups, A and B. The left side of the equation is  $\mathbb{E}[\hat{Y} \mid S = A]$ , where  $\hat{Y}$  is highlighted in a light orange box and labeled "class prediction" with a red arrow. The conditional part  $S = A$  is highlighted in a light teal box and labeled "sensitive" with a teal arrow. The right side of the equation is  $\mathbb{E}[\hat{Y} \mid S = B]$ , where  $\hat{Y}$  is highlighted in a light orange box and labeled "class prediction" with a red arrow. The conditional part  $S = B$  is highlighted in a light yellow box and labeled "sensitive" with a yellow arrow. A red question mark is placed above the equals sign.

# Fairness for Classifiers

(**weak**) definition of “demographic parity” for a classifier

$$\mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$$

The diagram includes several annotations: a green arrow labeled "sensitive" points to the variable  $S$  in the left-hand side of the equation; a yellow arrow labeled "sensitive" points to the variable  $S$  in the right-hand side; a red arrow labeled "score" points to the function  $m(\mathbf{X}, S)$  in both sides; and a red question mark is placed above the equals sign.



# Fairness for Classifiers

(weak) definition of “demographic parity” for a classifier

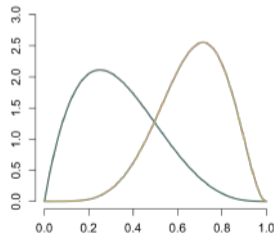
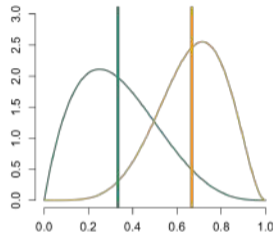
$$\mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$$

The diagram includes several annotations: a green arrow labeled “sensitive” points from the text “sensitive” to the variable  $S$  in the left-hand side of the equation; a yellow arrow labeled “sensitive” points from the text “sensitive” to the variable  $S$  in the right-hand side; a red arrow labeled “score” points from the text “score” to the function  $m(\mathbf{X}, S)$  in both sides.

(strong) definition of “demographic parity” for a classifier

$$\mathbb{P}[m(\mathbf{X}, S) \in \mathcal{I} \mid S = A] \stackrel{?}{=} \mathbb{P}[m(\mathbf{X}, S) \in \mathcal{I} \mid S = B]$$

$$\forall \mathcal{I} \subset [0, 1],$$



# Fairness for Classifiers

(weak) definition of “demographic parity” for a classifier

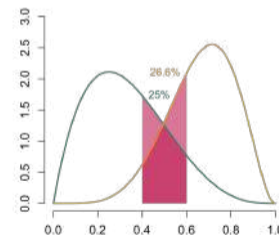
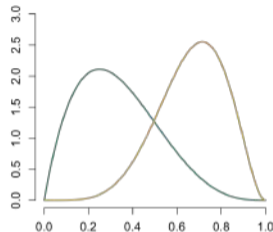
$$\mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$$

Diagram annotations: A green arrow labeled “sensitive” points from the text “sensitive” to the variable  $S$  in the left-hand side of the equation. A yellow arrow labeled “sensitive” points from the text “sensitive” to the variable  $S$  in the right-hand side. A red arrow labeled “score” points from the text “score” to the function  $m(\mathbf{X}, S)$  in both sides.

(strong) definition of “demographic parity” for a classifier

$$\mathbb{P}[m(\mathbf{X}, S) \in \mathcal{I} \mid S = A] \stackrel{?}{=} \mathbb{P}[m(\mathbf{X}, S) \in \mathcal{I} \mid S = B]$$

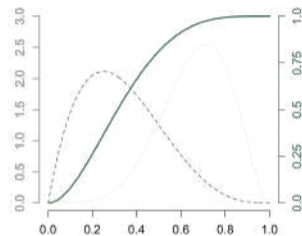
$\forall \mathcal{I} \subset [0, 1]$ , e.g. [40%; 60%].



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P} \left[ \overset{\text{score}}{m(\mathbf{X}, S)} \leq u \mid \overset{\text{sensitive}}{S = A} \right]$$

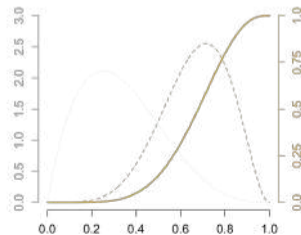
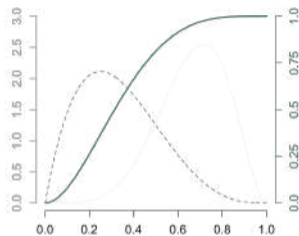
let  $F_A$  denote the cumulative distribution function of scores in group  $A$



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[\overset{\text{score}}{m(\mathbf{X}, S)} \leq u \mid S = \overset{\text{sensitive}}{A}]$$
$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

and  $F_B$  denote the cumulative distribution function of scores in group B





# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

and  $F_B$  denote the cumulative distribution function of scores in group B

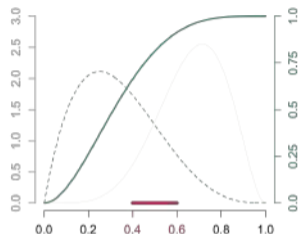
Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A$$

score

sensitive

levels of the score



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

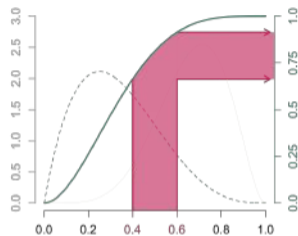
$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

and  $F_B$  denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A \text{ then } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = A$$

↑  
quantile



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

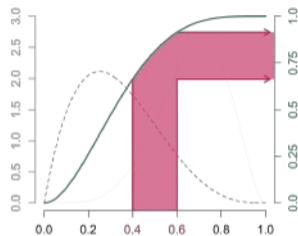
and  $F_B$  denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A \text{ then } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = A$$

quantile

probabilities associated to the score



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

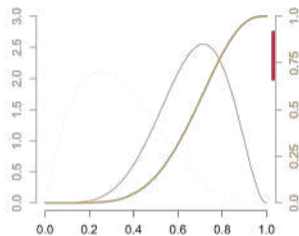
and  $F_B$  denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A \text{ then } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = A$$

then, in group B

$$\text{if } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = B$$



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

and  $F_B$  denote the cumulative distribution function of scores in group B

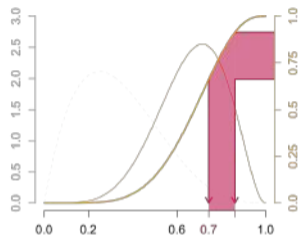
Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A \text{ then } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = A$$

then, in group B

$$\text{if } \text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = B \text{ then } m(\mathbf{X}, S) \in [0.743; 0.861] \mid S = B$$

score
sensitive



# Fairness for Classifiers using Optimal Transport

$$F_A(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = A]$$

$$F_B(u) = \mathbb{P}[m(\mathbf{X}, S) \leq u \mid S = B]$$

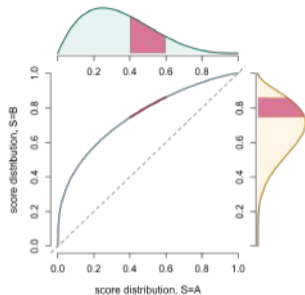
and  $F_B$  denote the cumulative distribution function of scores in group B

Consider individuals in group A such that

$$m(\mathbf{X}, S) \in [0.4; 0.6] \mid S = A$$

then, in group B ← optimal transport mapping  $T^*$

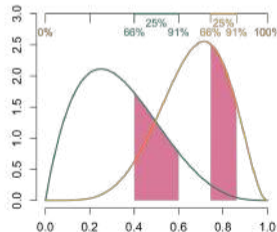
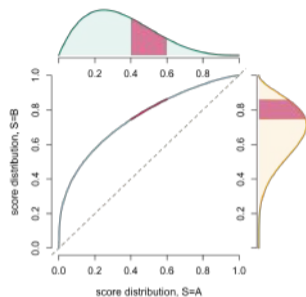
if  $\text{ranks}(m(\mathbf{X}, S)) \in [66.3\%; 91.3\%] \mid S = B$  then  $m(\mathbf{X}, S) \in [0.743; 0.861] \mid S = B$



# Formalizing Optimal Transport

Consider the following  $[0, 1] \rightarrow [0, 1]$  mapping

$$T^*(x) = F_B^{-1} \circ F_A(x)$$

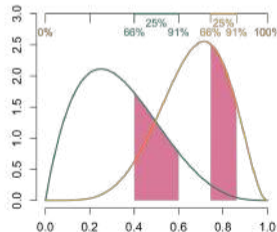
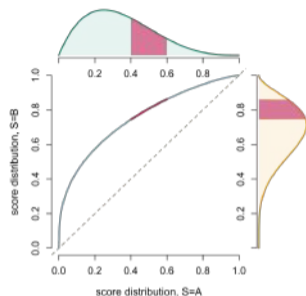


# Formalizing Optimal Transport

Consider the following  $[0, 1] \rightarrow [0, 1]$  mapping

$$T^*(x) = F_B^{-1} \circ F_A(x)$$

probability  $p$  associated with score  $x$  in group A



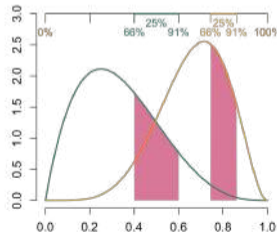
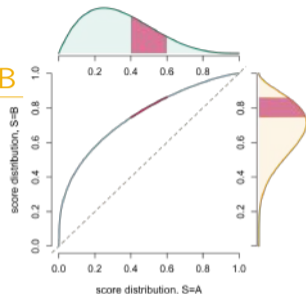
# Formalizing Optimal Transport

Consider the following  $[0, 1] \rightarrow [0, 1]$  mapping

$$T^*(x) = F_B^{-1} \circ F_A(x)$$

quantile of level  $p$  in group B

probability  $p$  associated with  $x$  in group A



# Fairness for Classifiers

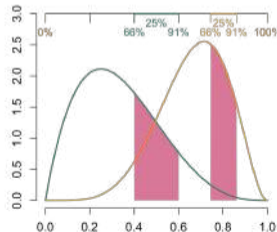
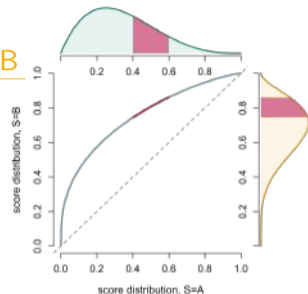
Consider the following  $[0, 1] \rightarrow [0, 1]$  mapping

optimal transport mapping

$$T^*(x) = F_B^{-1} \circ F_A(x)$$

quantile of level  $p$  in group B

probability  $p$  associated with  $x$  in group A



# Formalizing Optimal Transport

Consider the following  $[0, 1] \rightarrow [0, 1]$  mapping

$$T^*(x) = F_B^{-1} \circ F_A(x)$$

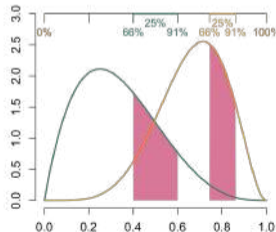
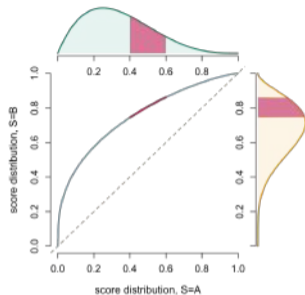
$$T^* = \operatorname{argmin}_{T:[0,1] \rightarrow [0,1]} \int_0^1 (T(x) - x)^2 dF_A(x)$$

i.e.  $\operatorname{argmin}_{T:[0,1] \rightarrow [0,1]} \mathbb{E}[(T(X) - X)^2]$  where  $X \sim F_A$ ,

$Y$  with  $Y \sim F_B$

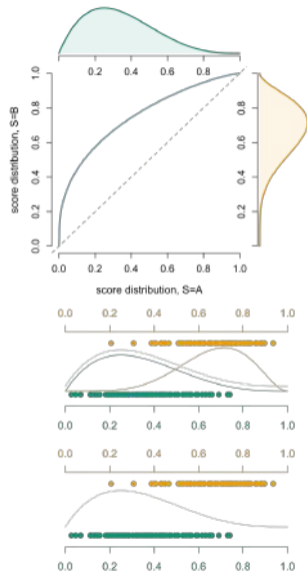
corresponding to **Monge (1781)** problem,  
revisited by **Kantorovich (1942)**.

(the minimum value is called **Wasserstein distance**)



# Optimal Transport with a Finite Sample (another interpretation)

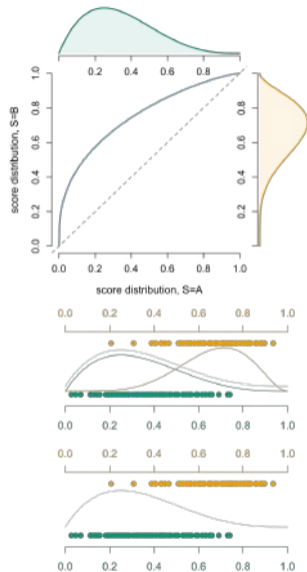
Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$



# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$

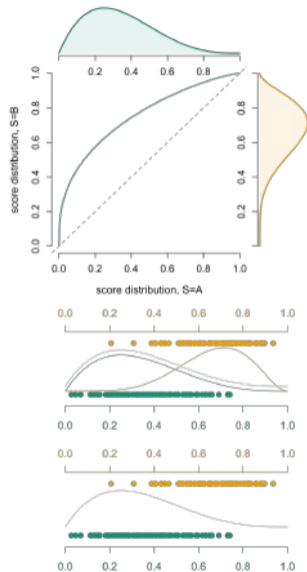


# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

$$m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$

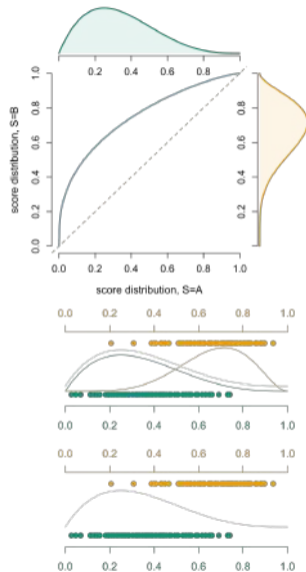


# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

$$m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$

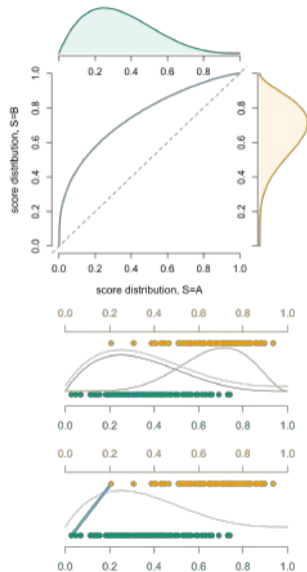
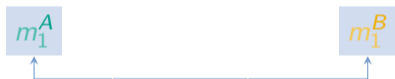


# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

$$m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$

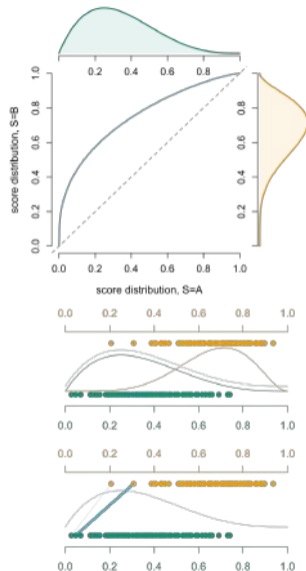


# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

$$m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$

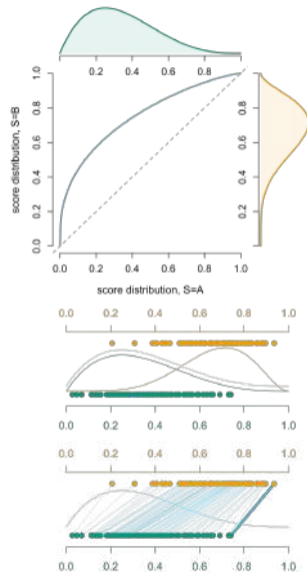
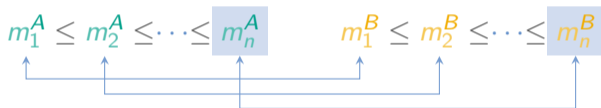


# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A$$

$$m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(\mathbf{x}_i, s_i = A))$  and  $(m(\mathbf{x}_i, s_i = B))$



# Optimal Transport with a Finite Sample (another interpretation)

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A \quad m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

Consider two samples,  $(m(x_i, s_i = A))$  and  $(m(x_i, s_i = B))$

$$m_1^A \leq m_2^A \leq \dots \leq m_n^A \quad \text{and} \quad m_1^B \leq m_2^B \leq \dots \leq m_n^B$$

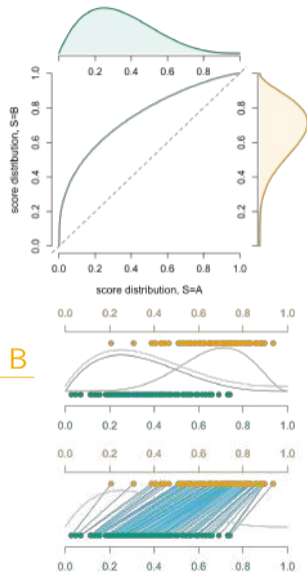
$m$  is not fair with respect to  $s$  if  $T^*(x) \neq x$ , or  $m_i^A \neq m_i^B$

optimal transport mapping

quantile of level  $p$  in group B

$$T^*(x) = F_B^{-1} \circ F_A(x) \neq x$$

probability  $p$  associated with  $u$  in group A

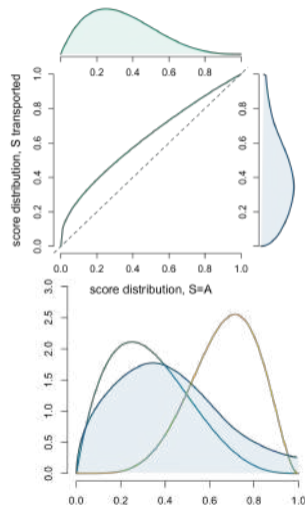
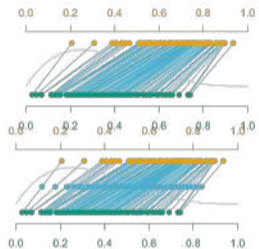


# Mitigating Discrimination with Wasserstein Barycenters

**Mitigation** is about finding some  $m^*$  “in-between” (Demographic Parity)

For individual  $i$ , why not

$$m_i^* = \frac{1}{2} m_i^A + \frac{1}{2} m_i^B$$



# Mitigating Discrimination with Wasserstein Barycenters

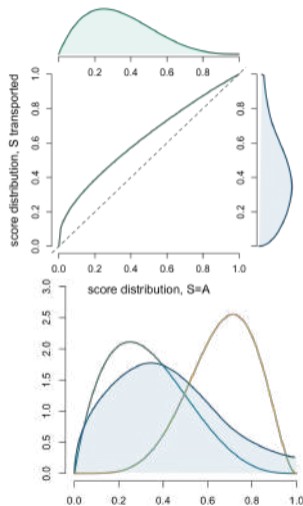
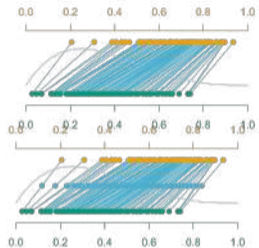
Mitigation is about finding some  $m^*$  “in-between” (Demographic Parity)

For individual  $i$ , why not

$$m_i^* = \frac{1}{2} m_i^A + \frac{1}{2} m_i^B$$

corresponding to

$$m^*(x, A) = \frac{1}{2} m^A(x) + \frac{1}{2} T^*(m^A(x))$$



# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some  $m^*$  “in-between” (Demographic Parity)

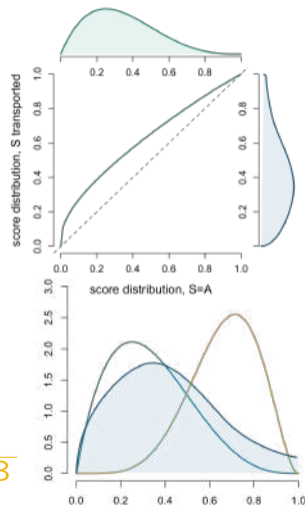
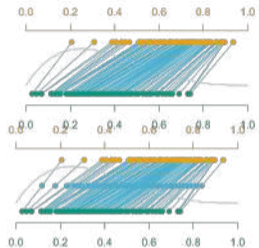
For individual  $i$ , why not

$$m_i^* = \frac{1}{2} m_i^A + \frac{1}{2} m_i^B$$

corresponding to

$$m^*(x, A) = \underbrace{\frac{1}{2}}_{\mathbb{P}[S=A]} m^A(x) + \underbrace{\frac{1}{2}}_{\mathbb{P}[S=B]} T^*(m^A(x))$$

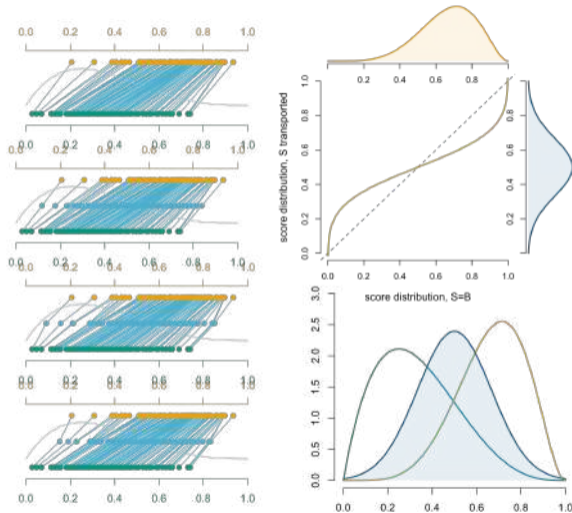
associated score in group B



# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some  $m^*$   
“in-between” (Demographic Parity)

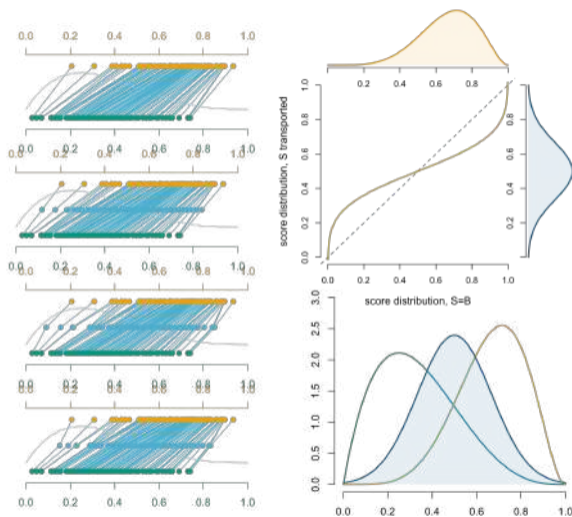
other “averages” could be considered



# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some  $m^*$  “in-between” (Demographic Parity)

other “averages” could be considered that one (“**Wasserstein barycenter**”) is actually optimal in terms of (empirical) risk

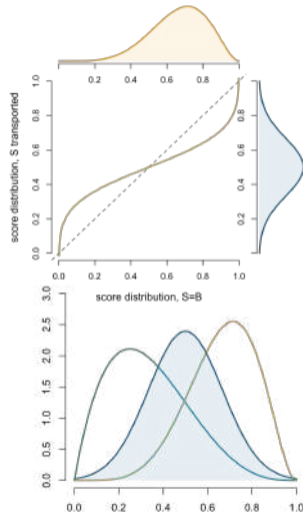
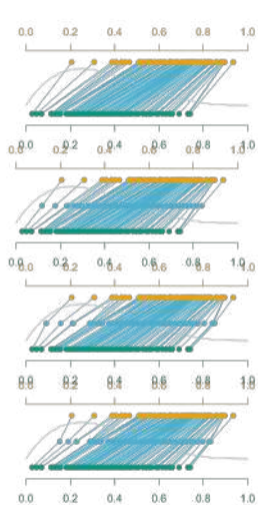


# Mitigating Discrimination with Wasserstein Barycenters

Mitigation is about finding some  $m^*$  “in-between” (Demographic Parity)

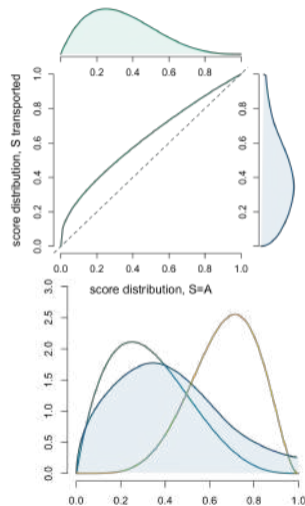
other “averages” could be considered that one (“**Wasserstein barycenter**”) is actually optimal in terms of (empirical) risk

Given a model  $m$  (regression, boosting, random forest, neural nets, etc) we can easily derive a “**fair model**”



# Mitigating Discrimination with Wasserstein Barycenters

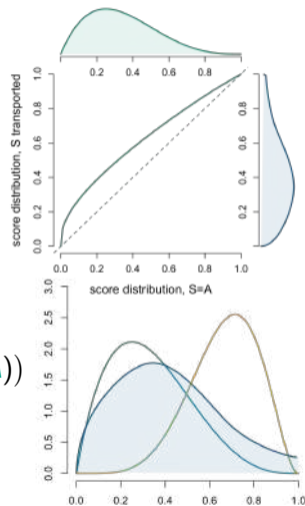
$$\left\{ \begin{array}{l} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{array} \right.$$



# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

$$\mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A))$$

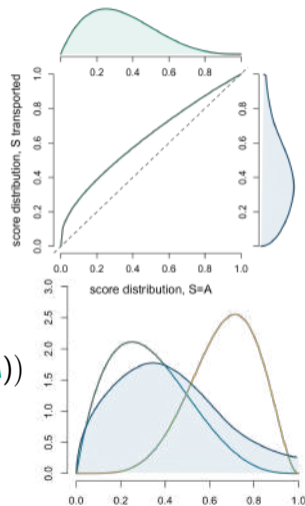


# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

$$\mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A))$$

↑ weights ↑



# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

score in group A

$$\mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A))$$

weights



# Mitigating Discrimination with Wasserstein Barycenters

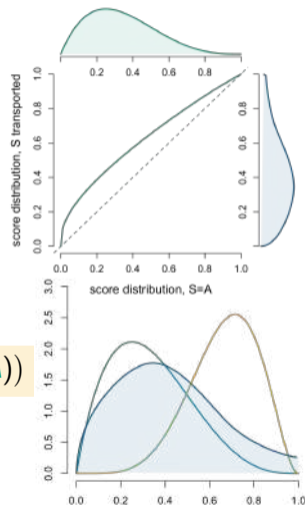
$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

score in group A

$$\mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A))$$

weights

$p = F_A(m(\mathbf{x}, s = A))$



# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

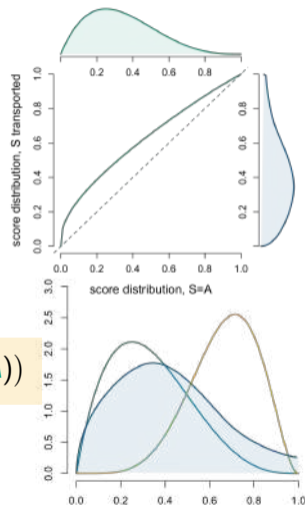
score in group A

$$\mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A))$$

weights

$p = F_A(m(\mathbf{x}, s = A))$

quantile score in group B associated with probability  $p$



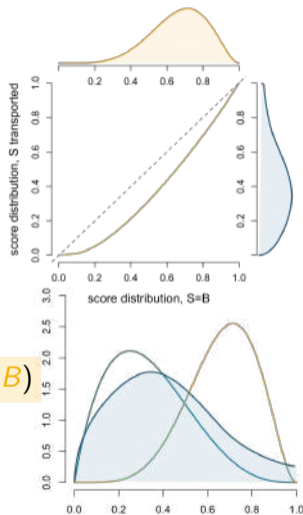
# Mitigating Discrimination with Wasserstein Barycenters

$$\begin{cases} m^*(\mathbf{x}, s = A) = \mathbb{P}[S = A] \cdot m(\mathbf{x}, s = A) \\ \quad + \mathbb{P}[S = B] \cdot F_B^{-1} \circ F_A(m(\mathbf{x}, s = A)) \\ m^*(\mathbf{x}, s = B) = \mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = B)) \\ \quad + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B). \end{cases}$$

$$p = F_B(m(\mathbf{x}, s = B))$$

$$\mathbb{P}[S = A] \cdot F_A^{-1} \circ F_B(m(\mathbf{x}, s = A)) + \mathbb{P}[S = B] \cdot m(\mathbf{x}, s = B)$$

quantile score in group A associated with probability  $p$



## Back to the COMPAS Example

$$\begin{cases} S : \text{race (binary), black \& white} \\ Y : \text{re-offense (binary), no \& yes} \\ \hat{Y} : \text{classifier (risk category), low \& high} \end{cases}$$

(standard) demographic parity would be translated as

$$\mathbb{P}[\hat{Y} = \text{high} \mid S = \text{black}] = 58\% \stackrel{?}{=} \mathbb{P}[\hat{Y} = \text{high} \mid S = \text{white}] = 33\%,$$

The diagram illustrates the concept of demographic parity. It shows the probability of a high risk prediction ( $\hat{Y} = \text{high}$ ) for two racial groups: black and white. The probability for black individuals is 58%, and for white individuals, it is 33%. A red question mark indicates that these probabilities are not equal, which is a violation of demographic parity. Annotations include 'sensitive' above 'black' and 'white', and 'predictions' below the conditional part of the equation, with arrows pointing to the respective terms.

# Back to the COMPAS Example, from Discrimination to Calibration

$$\begin{cases} S : \text{race (binary), black \& white} \\ Y : \text{re-offense (binary), no \& yes} \\ \hat{Y} : \text{classifier (risk category), low \& high} \end{cases}$$

$$\mathbb{P}[\hat{Y} = \text{high} | Y = \text{no}, S = \text{black}] = 42\% \stackrel{?}{=} \mathbb{P}[\hat{Y} = \text{high} | Y = \text{no}, S = \text{white}] = 22\%,$$

*sensitive* (teal arrow pointing to  $S = \text{black}$ )

*sensitive* (orange arrow pointing to  $S = \text{white}$ )

false positive rate (orange arrow pointing to both sides)

$$\mathbb{P}[Y = \text{no} | \hat{Y} = \text{high}, S = \text{black}] = 35\% \stackrel{?}{=} \mathbb{P}[Y = \text{no} | \hat{Y} = \text{high}, S = \text{white}] = 40\%.$$

false discovery rate (blue arrow pointing to both sides)

# From Discrimination to Calibration

demographic parity  $\rightarrow \mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$

The diagram shows the equation  $\mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$ . The term  $m(\mathbf{X}, S)$  is highlighted in a light brown box. The term  $S = A$  is highlighted in a teal box, with a teal arrow pointing to it from the word "sensitive" written above it. The term  $S = B$  is highlighted in a yellow box, with a yellow arrow pointing to it from the word "sensitive" written above it. A red arrow labeled "score" points from the teal box to the yellow box. A red question mark is placed above the equals sign.

# From Discrimination to Calibration

demographic parity  $\rightarrow \mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$

*sensitive* *sensitive*  
score

equalized odds  $\rightarrow \mathbb{E}[m(\mathbf{X}, S) \mid Y = y, S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid Y = y, S = B], \forall y$

*sensitive* *sensitive*  
score

# From Discrimination to Calibration

demographic parity  $\rightarrow \mathbb{E}[m(\mathbf{X}, S) \mid S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid S = B]$

*sensitive* (teal) points to  $S = A$ . *score* (red) points to  $m(\mathbf{X}, S)$ . *sensitive* (yellow) points to  $S = B$ . *score* (red) points from  $m(\mathbf{X}, S)$  in the left expression to  $m(\mathbf{X}, S)$  in the right expression. A red question mark is above the equals sign.

equalized odds  $\rightarrow \mathbb{E}[m(\mathbf{X}, S) \mid Y = y, S = A] \stackrel{?}{=} \mathbb{E}[m(\mathbf{X}, S) \mid Y = y, S = B], \forall y$

*sensitive* (teal) points to  $S = A$ . *score* (red) points to  $Y = y$ . *score* (red) points to  $m(\mathbf{X}, S)$ . *sensitive* (yellow) points to  $S = B$ . *score* (red) points from  $m(\mathbf{X}, S)$  in the left expression to  $m(\mathbf{X}, S)$  in the right expression. A red question mark is above the equals sign.

calibration  $\rightarrow \mathbb{E}[Y \mid m(\mathbf{X}, S) = u, S = A] \stackrel{?}{=} \mathbb{E}[Y \mid m(\mathbf{X}, S) = u, S = B], \forall u$

*sensitive* (teal) points to  $S = A$ . *score* (red) points to  $Y$ . *score* (red) points to  $m(\mathbf{X}, S) = u$ . *sensitive* (yellow) points to  $S = B$ . *score* (red) points from  $m(\mathbf{X}, S) = u$  in the left expression to  $m(\mathbf{X}, S) = u$  in the right expression. A red question mark is above the equals sign.

## From Discrimination to Calibration (an Epistemological Detour)

$$\underbrace{\frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \in \mathcal{A})}_{\text{(empirical) frequency}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{P}(\{X \in \mathcal{A}\})}_{\text{probability}} = \mathbb{P}[X \in \mathcal{A}], \text{ as } n \rightarrow \infty.$$

*“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 for us at all,”* von Mises (1928, 1939)

## From Discrimination to Calibration (an Epistemological Detour)

This frequentist approach is unable to make sense of the probability of a “single singular event”.

$$\underbrace{\frac{\sum_{i=1}^n Y_i \cdot \mathbf{1}(X_i \in \mathcal{A})}{\sum_{i=1}^n \mathbf{1}(X_i \in \mathcal{A})}}_{\text{(empirical) average}} \xrightarrow{\text{a.s.}} \underbrace{\mathbb{E}(Y|X \in \mathcal{A})}_{\text{expected value}}, \text{ as } n \rightarrow \infty.$$

Property  $\mathbb{E}[Y | m(\mathbf{X}, S) = u] = u, \forall u \in [0, 1]$  corresponds to “**calibration**”.

*“Out of all the times you said there was a 40 percent chance of rain, how often did rain actually occur? If, over the long run, it really did rain about 40 percent of the time, that means your forecasts were well calibrated,”* Silver (2012)

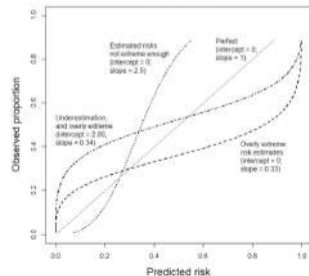
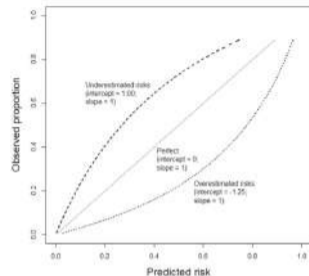
# From Discrimination to Calibration (an Epistemological Detour)

As explained in [Van Calster et al. \(2019\)](#), *“among patients with an estimated risk of 20%, we expect 20 in 100 to have or to develop the event,”*

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

Most machine learning models can be poorly calibrated, [Denuit et al. \(2021\)](#), [Machado et al. \(2024\)](#).

(picture source: [Van Calster et al. \(2019\)](#))



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



Designed to Deceive: Do These People Look Real to You?  
by Kashmir Hill and Jeremy White Nov. 21, 2020,

**The New York Times**

Use of GAN (Generative adversarial network) to generate fake pictures (StyleGAN2 package, implemented in TensorFlow)



## From Discrimination to Calibration



# From Discrimination to Calibration



$S = \text{female}$



$S = \text{male}$

## From Discrimination to Calibration



female (0.984)  
male (0.016)



female (0.983)  
male (0.017)



female (0.982)  
male (0.018)



**female (0.960)**  
male (0.040)



**female (0.009)**  
male (0.991)



female (0.013)  
male (0.987)



female (0.014)  
male (0.986)



female (0.015)  
male (0.985)

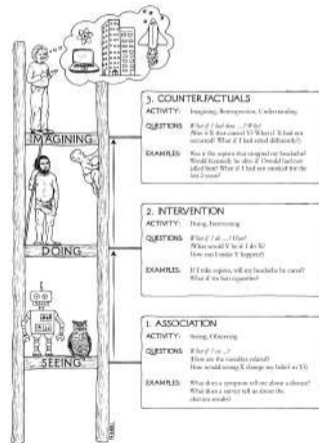
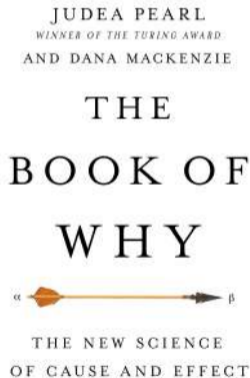
## Individual Fairness

We have **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)

# Individual Fairness

We have **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)

“Ladder of causation” from Pearl et al. (2009), Pearl and Mackenzie (2018)

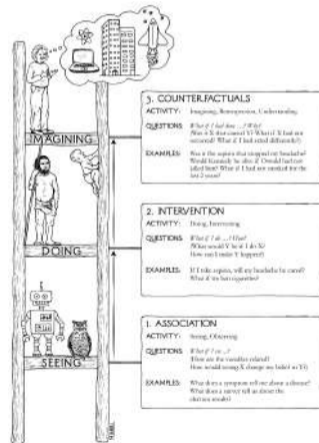
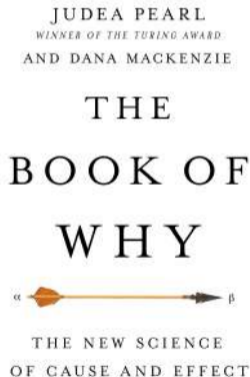


# Individual Fairness

We have **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)

“Ladder of causation” from Pearl et al. (2009), Pearl and Mackenzie (2018)

➤ 1. **Association**  
(Seeing, “*what if I see...*”)

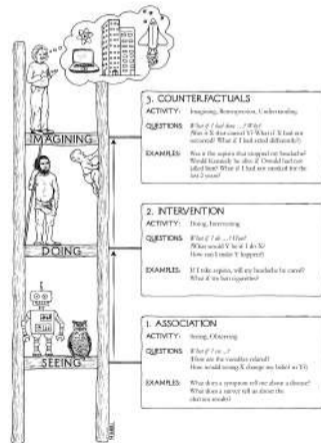
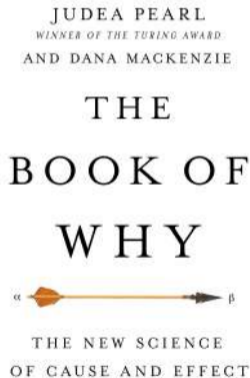


# Individual Fairness

We have **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)

“Ladder of causation” from Pearl et al. (2009), Pearl and Mackenzie (2018)

- 2. **Intervention**  
(Doing, “*what if I do...*”)
- 1. **Association**  
(Seeing, “*what if I see...*”)

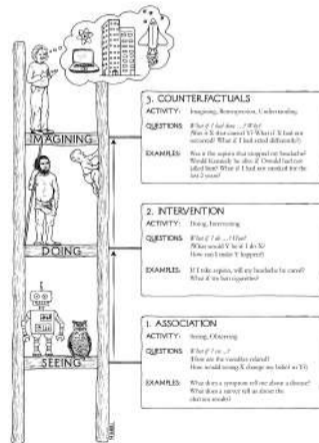
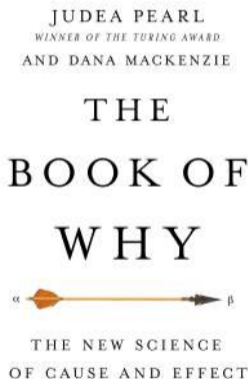


# Individual Fairness

We have **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)

“Ladder of causation” from Pearl et al. (2009), Pearl and Mackenzie (2018)

- 3. **Counterfactuals**  
(Imagining, “*what if I had done...*”)
- 2. **Intervention**  
(Doing, “*what if I do...*”)
- 1. **Association**  
(Seeing, “*what if I see...*”)



# What About Interpretation ?

*“Humans think in stories rather than facts, numbers or equations - and the simpler the story, the better,”* Harari (2018)

For Glenn (2000), insurer’s risk selection process has two sides:

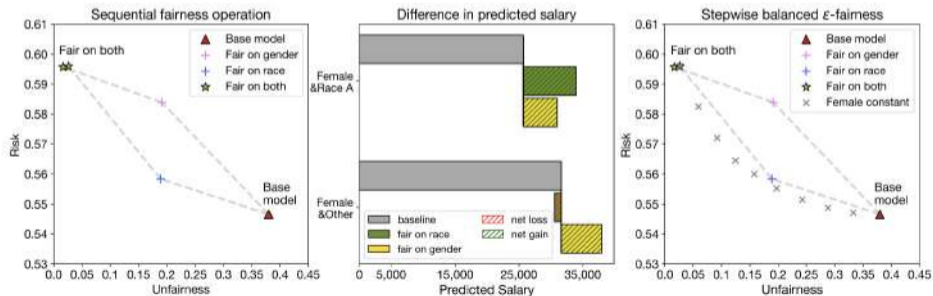
- › the one presented to regulators and policyholders (numbers, statistics and objectivity),
- › the other presented to underwriters (stories, character and subjective judgment).

The rhetoric of insurance exclusion – numbers, objectivity and statistics – forms what Brian Glenn calls *“the myth of the actuary,”* *“a powerful rhetorical situation in which decisions appear to be based on objectively determined criteria when they are also largely based on subjective ones”* or *“the subjective nature of a seemingly objective process”*.

# What About Interpretation ?

*“The fact that the selection of risk factors is subjective and contingent upon narratives of risk and responsibility has in the past played a far larger role than whether or not someone with a wood stove is charged higher premiums.”*

Going further, *“virtually every aspect of the insurance industry is predicated on stories first and then numbers,”* Glenn (2003)



## Mitigating Discrimination ? (brief conclusion)

If it is mandatory to mitigate, there are robust techniques that can guarantee fairness


Supreme Court Justice Harry Blackmun stated, in 1978,

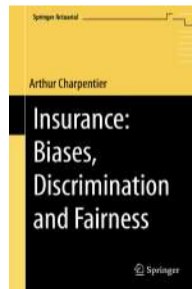
*“In order to get beyond racism, we must first take account of race. There is no other way. And in order to treat some persons equally, we must treat them differently,”* Knowlton (1978), cited in Lippert-Rasmussen (2020)

In 2007, John G. Roberts of the U.S. Supreme Court submits

*“The way to stop discrimination on the basis of race is to stop discriminating on the basis of race,”* Sabbagh (2007) and Turner (2015)

To go further,

Charpentier (2024) Insurance: Biases, Discrimination and Fairness. 



# Mitigating Discrimination ? (brief conclusion)



UNIVERSITÉ  
LAVAL

Institut Intelligence et données (IID)

ULaval nouvelles



[À propos](#) [Expertises](#) [Services](#) [Projets](#) [Actualités](#) [Évènements](#)

## Journée sur l'équité et la discrimination en assurance 2024

16 mai 2024, 8h30 à 17h | En mode hybride  
En direct de l'Université Laval, pavillon Alexandre-Vachon + Zoom

Réservez la date!

*Ouverture des inscriptions à venir.*

## Workshop on Fairness and Discrimination in Insurance 2024

May 16th 2024, 8:30AM - 5PM | Hybrid Event  
Live from Université Laval, pavillon Alexandre-Vachon + Zoom

Save the date!

*Registrations opening soon.*

## Workshop on Fairness and Discrimination in Insurance

*Journée sur l'équité et la discrimination en assurance*

May 16th 2024,  
8:30AM - 5PM  
Hybrid Event

16 mai 2024,  
8h30 à 17h  
Événement hybride



## References

- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*, May 23.
- Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.
- Barocas, S., Hardt, M., and Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, 1:2017.
- Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.
- Crossney, K. B. (2016). Redlining. <https://philadelphiaencyclopedia.org/essays/redlining/>.
- Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics & Economics*.
- Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.
- Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.
- Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.

## References

- Feller, A., Pierson, E., Corbett-Davies, S., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. *The Washington Post*, October 17.
- Glenn, B. J. (2000). The shifting rhetoric of insurance denial. *Law and Society Review*, pages 779–808.
- Glenn, B. J. (2003). Postmodernism: the basis of insurance. *Risk Management and Insurance Review*, 6(2):131–143.
- Harari, Y. N. (2018). *21 Lessons for the 21st Century*. Random House.
- Hu, F., Ratz, P., and Charpentier, A. (2023). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases – ECML PKDD*.
- Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Annual AAAI Conference on Artificial Intelligence*.
- Hume, D. (1739). *A Treatise of Human Nature*. Cambridge University Press Archive.
- Kantorovich, L. V. (1942). On the translocation of masses. In *Doklady Akademii Nauk USSR*, volume 37, pages 199–201.
- Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

# References

- Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.
- Kranzberg, M. (1986). Technology and history:" kranzberg's laws". *Technology and culture*, 27(3):544–560.
- Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.
- Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.
- Machado, A. F., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024). From uncertainty to precision: Enhancing binary classifier performance through calibration. *arXiv preprint arXiv:2402.07790*.
- Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.
- Pearl, J. et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.
- Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.
- Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

# References

- Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office of the Controller*.
- Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.
- Silver, N. (2012). *The signal and the noise: Why so many predictions fail-but some don't*. Penguin.
- Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 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):1–7.
- von Mises, R. (1928). *Wahrscheinlichkeit Statistik und Wahrheit*. Springer.
- von Mises, R. (1939). *Probability, statistics and truth*. Macmillan.