

Fairness and Ethic in Pricing

Arthur Charpentier

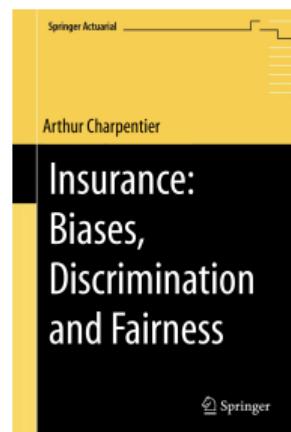
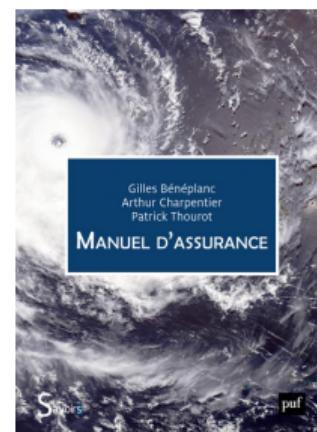
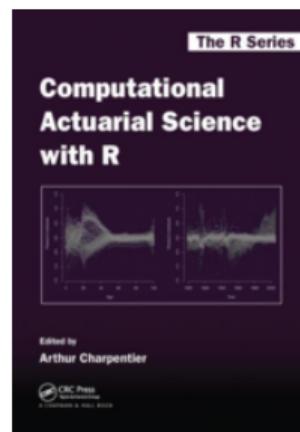
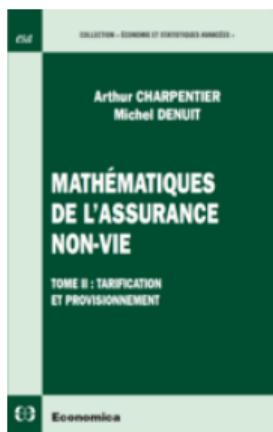
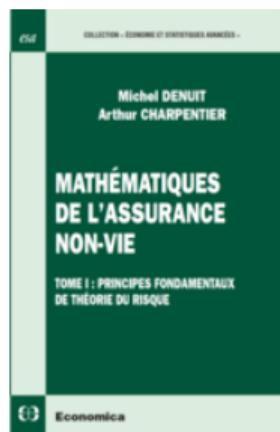
Pricing Seminar, September 2023



Bio (short)

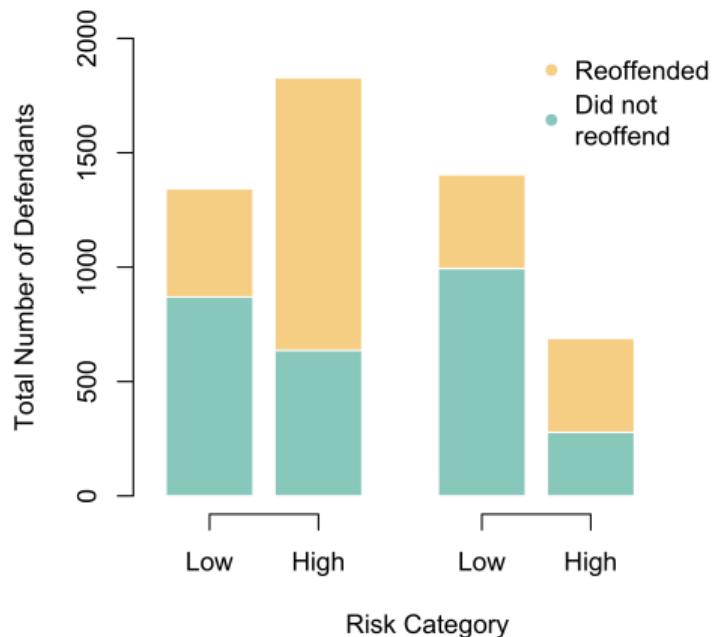
Professor at Université du Québec à Montréal, (<https://freakonometrics.github.io/>)

- › Denuit and Charpentier (2004, 2005) Mathématiques de l'Assurance Non-Vie,
- › Charpentier (2014) Computational Actuarial Science with R,
- › Bénéplanc et al. (2022) Manuel d'Assurance,
- › Charpentier (2023) Insurance: Biases, Discrimination and Fairness.



Motivation (1. Propublica, Actuarial Justice)

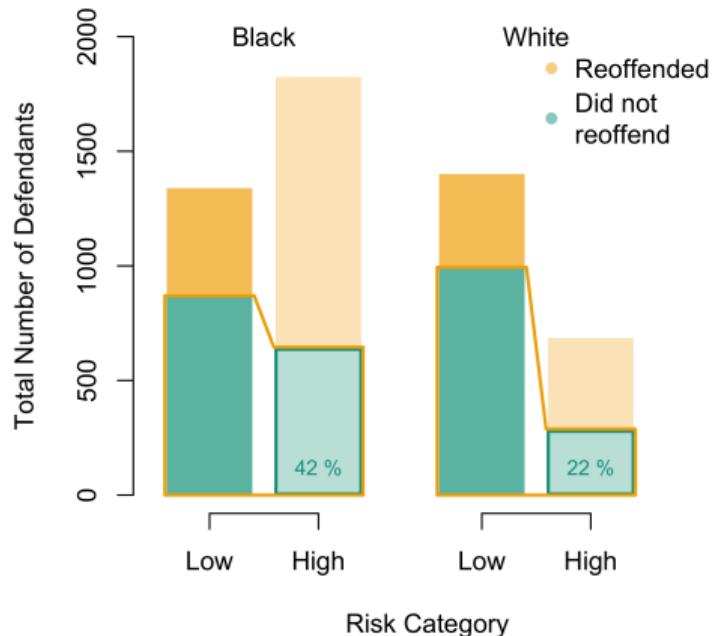
- Concept of "actuarial justice" as coined in Feeley and Simon (1994)
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), Perry (2013)



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

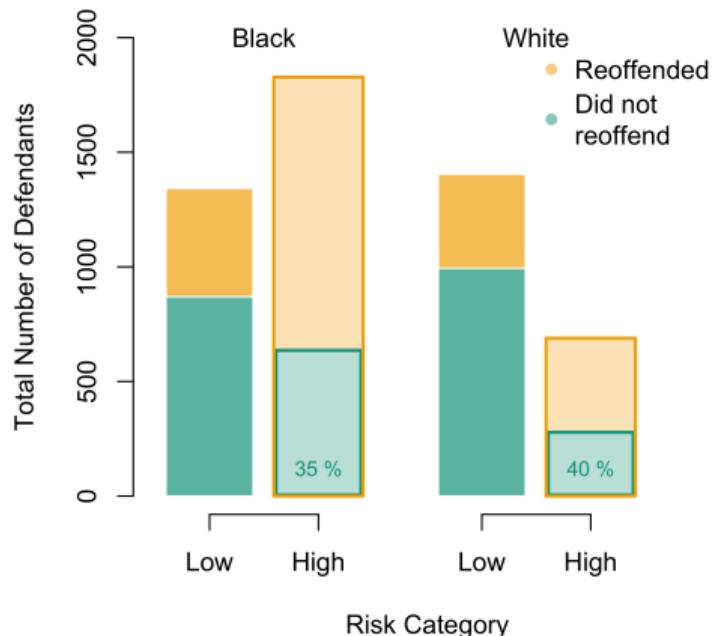
Motivation (1. Propublica, Actuarial Justice)

- From Feller et al. (2016),
 - for White people, among those who did not re-offend, 78% were properly classified, since 22% did re-offend,
 - for Black people, among those who did not re-offend, 58% were properly classified, since 42% did re-offend.

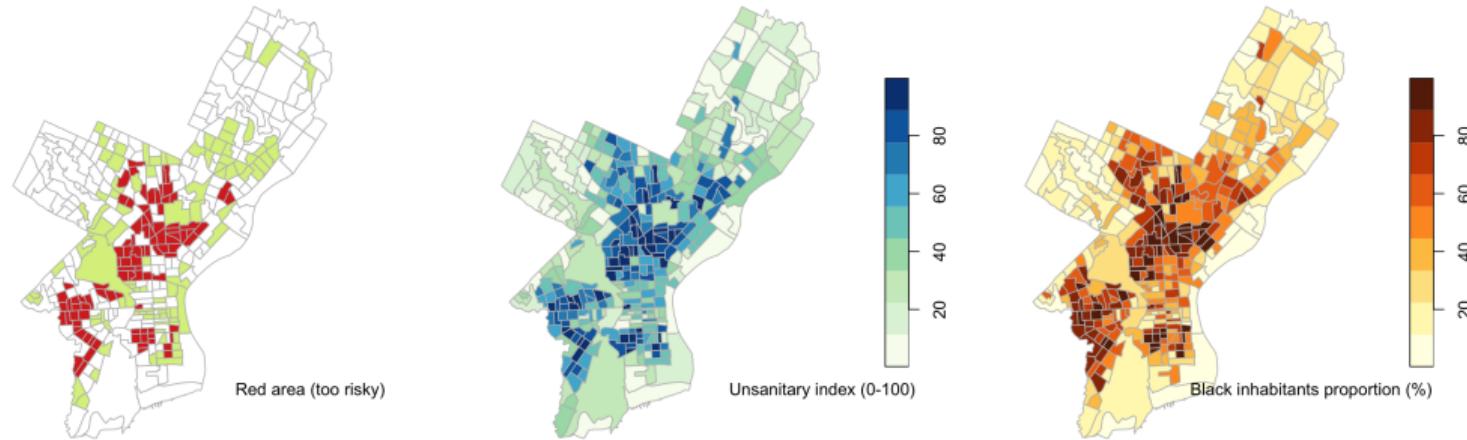


Motivation (1. Propublica, Actuarial Justice)

- 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.



Motivation (2. 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 and Insurance

"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)

"Machine learning won't give you anything like gender neutrality 'for free' that you didn't explicitly ask for", Kearns and Roth (2019)

Discrimination and Protected Attributes ?

California

Allowed (with applicable limitations): driving experience, marital status, address/zip code

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, residential status, insurance history

Notes & Clarifications: California's insurance commissioner banned gender as of January 2019. Occupation and education are permitted for use in group plans (i.e. for alumni associations and other membership programs).

Georgia

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, and employment status

Notes & Clarifications: none

Hawaii

Allowed (with applicable limitations): address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Illinois

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, education, occupation, employment status, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): none

Notes & Clarifications: none

Massachusetts

Allowed (with applicable limitations): years of driving experience, address/zip code, insurance history

Prohibited (or effectively prohibited): gender, age, credit history, education, occupation, employment status, marital status, residential status

Notes & Clarifications: none

Michigan

Allowed (with applicable limitations): gender (group-rated policies), age, years of driving experience, credit history, education, occupation, employment status, marital status (group-rated policies), residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): gender (non-group policies), marital status (non-group policies)

Notes & Clarifications: Gender and marital status are permitted only in rate-making for group plans (i.e. for alumni associations and other membership programs). **UPDATE: Michigan lawmakers approved a major insurance reform bill** in May 2019 that will ban insurers in the state from using gender, marital status, address/zipcode, residential status, education and occupation in rate setting. The ban will be enforced starting in July 2020. Insurers will be permitted to use "territory" as approved by the state regulators instead of zip code.

New York

Allowed (with applicable limitations): gender, age, years of driving experience, credit history, marital status, residential status, address/zip code, insurance history

Prohibited (or effectively prohibited): occupation, education, employment status

Notes & Clarifications: none

via **The Zebra (2022)**

Fairness (Demographic Parity) for Classifiers

- Some notations,

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

- One could consider multiple sensitive attributes, see [Hu et al. \(2023b\)](#)

Fairness (Demographic Parity) for Classifiers

- Defining "Demographic Parity", Corbett-Davies et al. (2017) or Agarwal (2021)

Weak Demographic Parity,

Decision function \hat{y} satisfies weak demographic parity if $\hat{Y} \perp\!\!\!\perp S$, i.e.

$$\mathbb{E}[\hat{Y}|S = A] = \mathbb{E}[\hat{Y}|S = B],$$

or

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t)|S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t)|S = B].$$

One can easily obtain weak Demographic Parity using different thresholds

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t_A)|S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) > t_B)|S = B].$$

freakonometrics

freakonometrics.hypotheses.org

– Arthur Charpentier, 2023

10 / 35

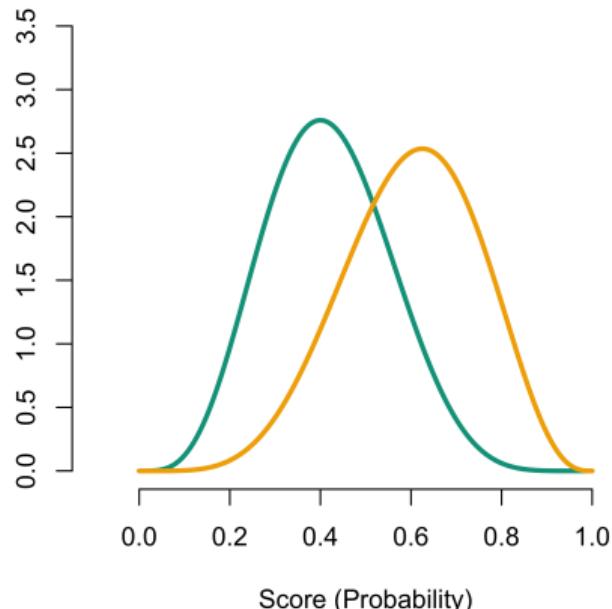
Fairness (Demographic Parity) for Scores

Strong Demographic Parity,

$$\mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) \in E) | S = A] = \mathbb{E}[\mathbf{1}(m(\mathbf{X}, S) \in E) | S = B],$$

for any $E \subset [0, 1]$, or $\mathbb{P}_A[E] = \mathbb{P}_B[E]$,

$$\begin{cases} \mathbb{P}_A[E] = \mathbb{P}[m(\mathbf{X}, S) \in E | S = A] \\ \mathbb{P}_B[E] = \mathbb{P}[m(\mathbf{X}, S) \in E | S = B] \end{cases}$$



Fairness (Demographic Parity) for Scores

- Use some "distance" between \mathbb{P}_A and \mathbb{P}_B
(TV, KL, or Wasserstein)

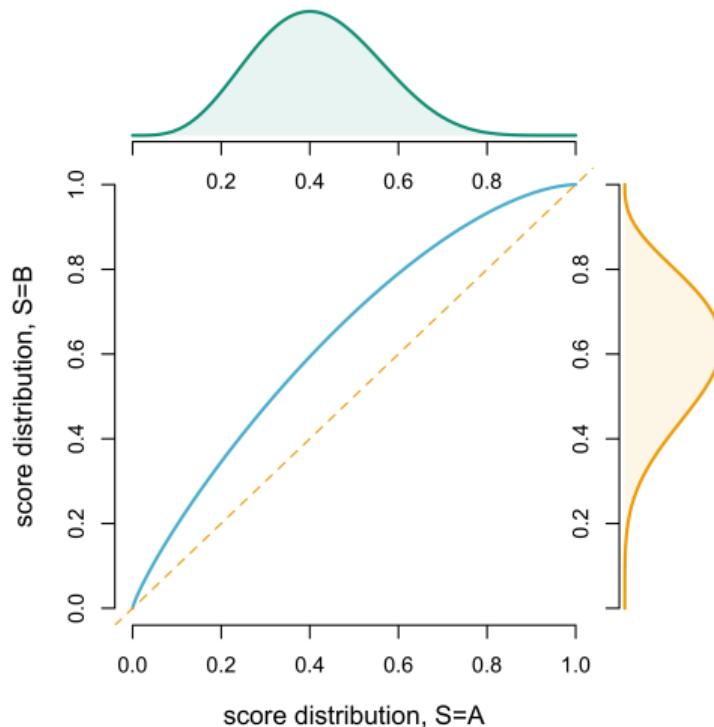
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \left\{ \mathbb{E}[\ell(X, Y)], (X, Y) \sim \pi \right\}$$

or

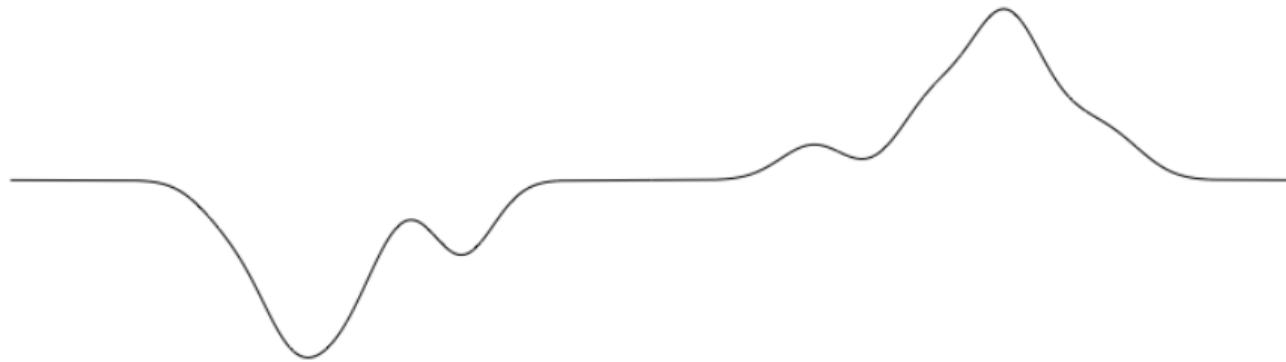
$$\inf_{\pi \in \Pi(\mathbb{P}_A, \mathbb{P}_B)} \left\{ \int \ell(x, y) \pi(dx, dy) \right\}.$$

or using a transport mapping \mathcal{T}

$$\inf_{\mathcal{T}: \mathbb{P}_A = \mathbb{P}_B} \left\{ \int \ell(x, \mathcal{T}(x)) d\mathbb{P}_A(x) \right\}.$$



Fairness and Optimal Transport



Monge (1781), Mémoire sur la théorie des déblais et des remblais

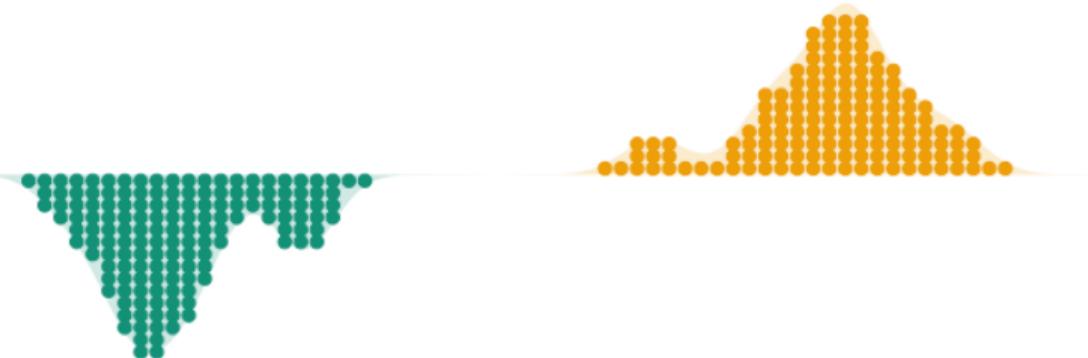
We want to **transport** optimally sand from a **hole** (with shape $-d\mathbb{P}_A$) to a **pile** (with shape $d\mathbb{P}_B$). "*Rien ne se perd, rien ne se crée, tout se transporte*": $\int d\mathbb{P}_A = \int d\mathbb{P}_B$.

Fairness and Optimal Transport



$$\inf_{\mathcal{T}: \mathcal{T}_{\#}\mathbb{P}_A = \mathbb{P}_B} \left\{ \int (\textcolor{teal}{x} - \mathcal{T}(x))^k d\mathbb{P}_A(x) \right\} . = \left(\int_0^1 \left| \mathcal{F}_0^{-1}(u) - \mathcal{F}_1^{-1}(u) \right|^k du \right)^{1/k},$$

Fairness and Optimal Transport



$$\inf_{\mathcal{T}: \mathcal{T}_{\#}\mathbb{P}_A = \mathbb{P}_B} \left\{ \int (\mathbf{x} - \mathcal{T}(x))^k d\mathbb{P}_A(\mathbf{x}) \right\} = \left(\int_0^1 |F_A^{-1}(u) - F_B^{-1}(u)|^k du \right)^{1/k},$$

Fairness and Optimal Transport



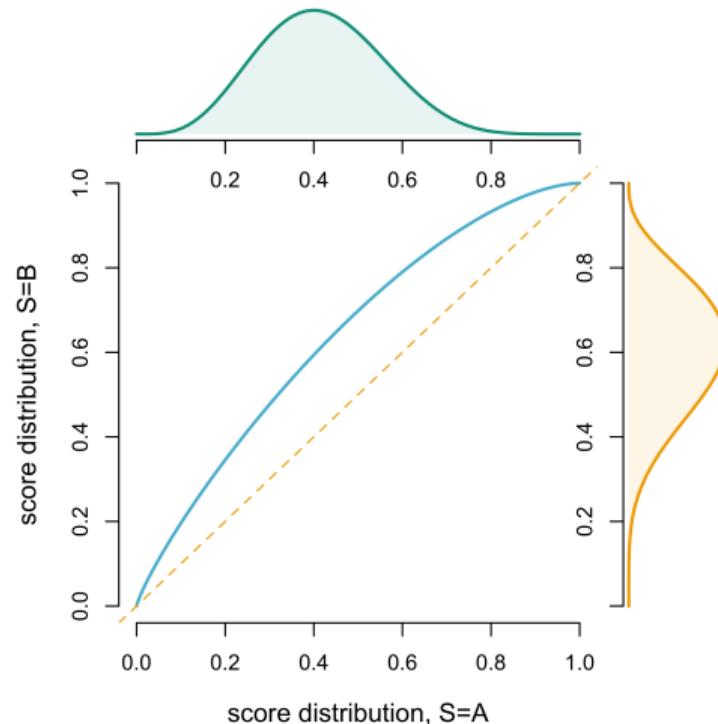
Optimal transport plan is here $\mathcal{T}^* : x \mapsto y = F_B^{-1} \circ F_A(x)$ (increasing function)

Counterfactual Fairness (and Optimal Transport)

- Used to quantify unfairness,

m satisfies **Strong Demographic Parity**
if $W_2 = 0$,

$$W_2 = \left(\int_0^1 \left(F_A^{-1}(u) - F_B^{-1}(u) \right)^2 du \right)^{1/2}.$$



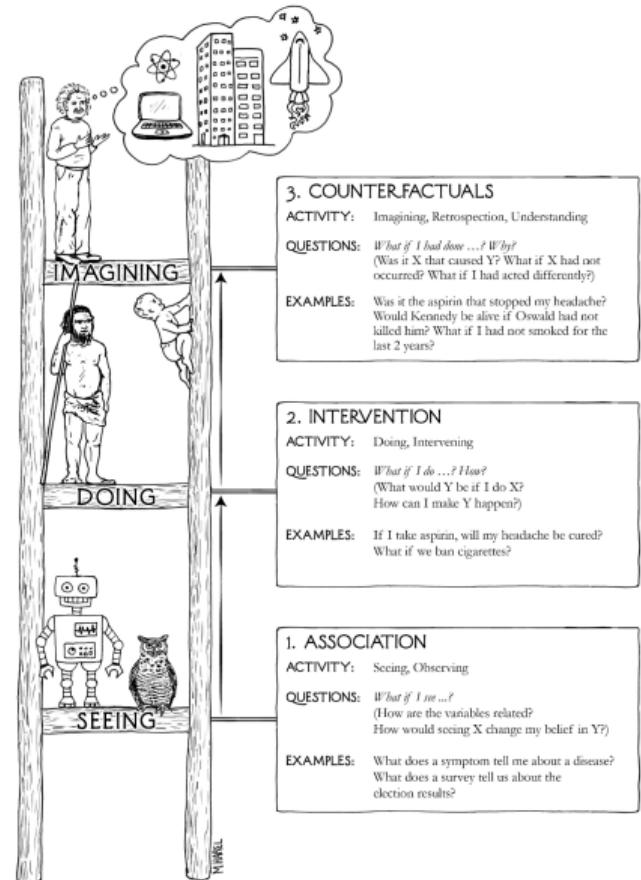
Counterfactual Fairness (and Optimal Transport)

“Ladder of causation” from Pearl et al. (2009)

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

Picture source: Pearl and Mackenzie (2018)

What would be the impact of a treatment T on a variable of interest Y ?



Counterfactual Fairness (and Optimal Transport)

- › Define individual or counterfactual fairness, Castelnovo et al. (2022)
"Individual fairness is embodied in the following principle: similar individuals should be given similar decisions. This principle deals with the comparison of single individuals rather than focusing on groups of people sharing some characteristics."
- › Following Kusner et al. (2017)

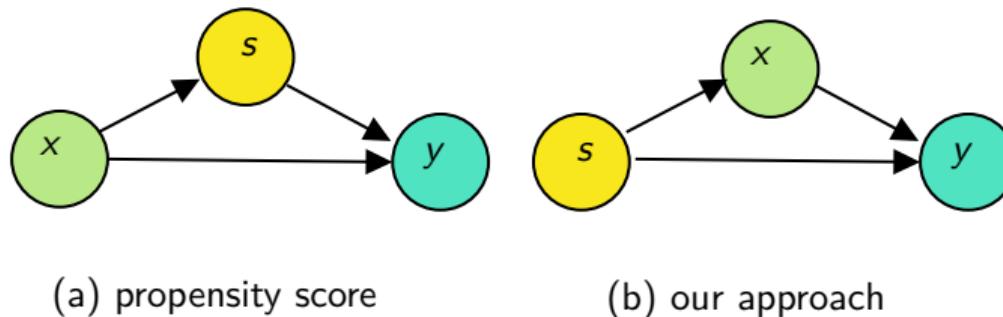
A decision is **counterfactually fair** if the prediction in the real world is the same as the prediction in the counterfactual world

$$\mathbb{E}[Y_{S \leftarrow A}^* | \mathbf{X} = \mathbf{x}] = \mathbb{E}[Y_{S \leftarrow B}^* | \mathbf{X} = \mathbf{x}], \quad \forall \mathbf{x},$$

where $Y_{S \leftarrow A}^*$ and $Y_{S \leftarrow B}^*$ denote "potential outcomes".

- › since we use the same \mathbf{x} it is a **ceteris paribus counterfactual**.
(is the counterfactual of a man with height 190 cm a woman with height 190 cm ?)

Counterfactual Fairness (and Optimal Transport)



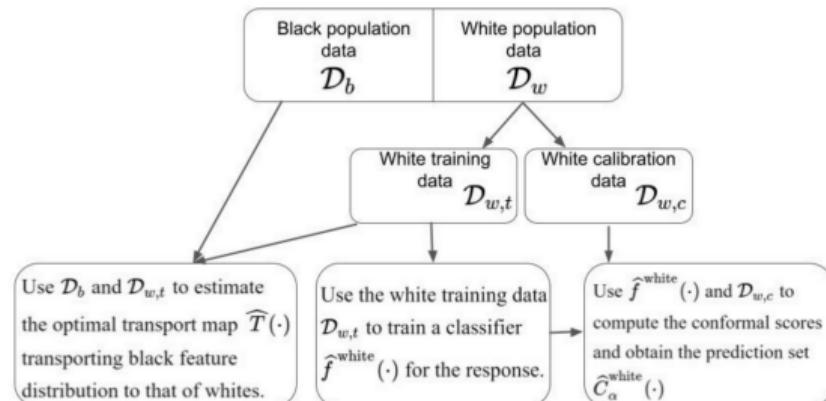
- Charpentier et al. (2023a) defined **mutatis mutandis** counterfactual fairness,

$$\mathbb{E}[Y_{S \leftarrow A}^* | \mathbf{X} = \mathbf{x}] = \mathbb{E}[Y_{S \leftarrow B}^* | \mathbf{X} = \mathcal{T}^*(\mathbf{x})], \quad \forall \mathbf{x}.$$

(probability to get surgery when delivering a baby for Black / non-Black mother)

Back to Actuarial Justice

➤ See Berk et al. (2021)



Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets*

Richard A. Berk
University of Pennsylvania
Arun Kumar Kuchibhotla
Carnegie Mellon University
Eric Tchetgen Tchetgen
University of Pennsylvania

Abstract

In the United States and elsewhere, risk assessment algorithms are being used to help inform criminal justice decision-makers. A common intent is to forecast an offender's "future dangerousness." Such algorithms have been correctly criticized for potential unfairness, and there is an active cottage industry trying to make repairs. In this paper, we use counterfactual reasoning to consider the prospects for improved fairness when members of a disadvantaged class are treated by a risk algorithm as if they are members of an advantaged class. We combine a machine learning classifier trained in a novel manner with an optimal transport adjustment for the relevant joint probability distributions, which together provide a constructive response to claims of bias-in-bias-out. A key distinction is made between fairness claims that are empirically testable and fairness claims that are not. We then use confusion tables and conformal prediction sets to evaluate achieved fairness for estimated risk. Our data are a random sample of 300,000 offenders at their arraignments for a large metropolitan area in the United States during which decisions to release or detain are made. We show that substantial improvement in fairness can be achieved consistent with a Pareto improvement for legally protected classes.

*Cary Coglianese and Sandra Mayson provided many insightful suggestions for legal conceptions of fairness and the prospect for criminal justice reform. Emanuele Candès offered several very instructive insights when commenting on this work at the Stanford/Berkeley Online Causal Inference Seminar. We also received very helpful feedback from a group of researchers at MIT and Harvard who work on causal inference. In that regard, a special thanks go to Devavrat Shah. Thanks also go to three thoughtful reviewers.

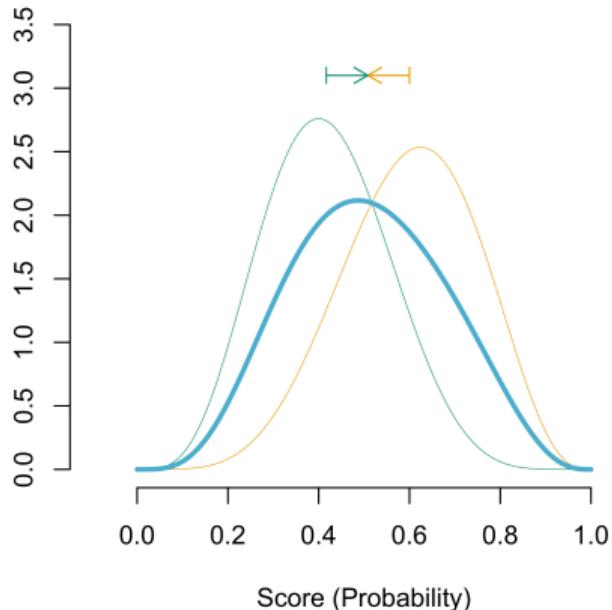
Mitigation with Wasserstein Barycenter

- › If $W_2 \neq 0$ can we mitigate discrimination ?
- › Use of Wasserstein Barycenter
see Charpentier et al. (2023b)
- › In Euclidean spaces

$$\mathbf{z}^* = \operatorname{argmin}_{\mathbf{z}} \left\{ \sum_{i=1}^n \omega_i d(\mathbf{z}, \mathbf{z}_i)^2 \right\},$$

- › For probability measures

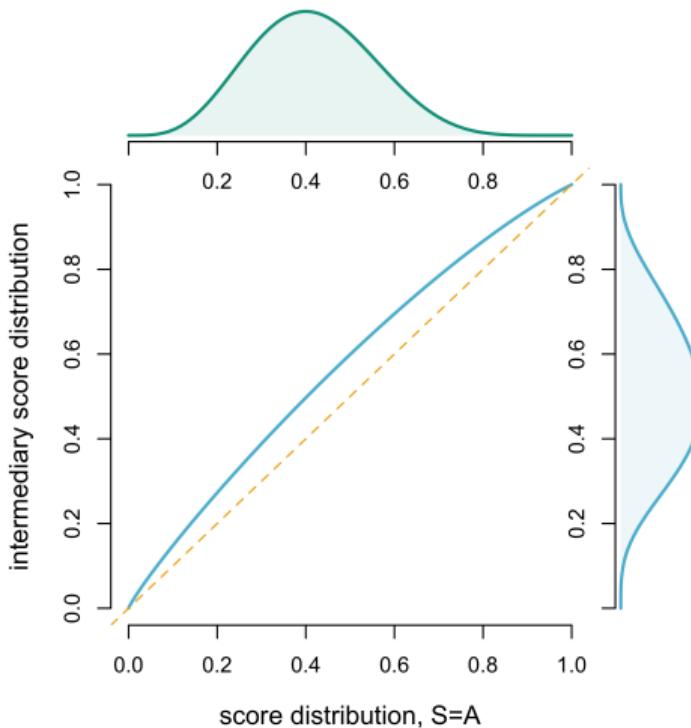
$$\mathbb{P}^* = \operatorname{argmin}_{\mathbb{Q}} \left\{ \sum_{i=1}^n \omega_i d(\mathbb{Q}, \mathbb{P}_i)^2 \right\},$$



Mitigation with Wasserstein Barycenter

Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$,
the “fair barycenter score” is

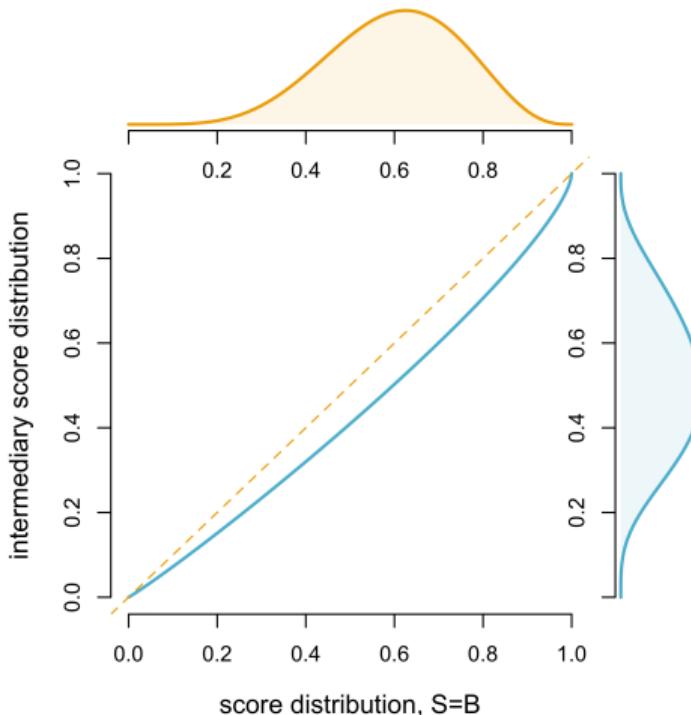
$$\begin{aligned} & m^*(\mathbf{x}, s = \text{A}) \\ = & \mathbb{P}[S = \text{A}] \cdot m(\mathbf{x}, s = \text{A}) \\ + & \mathbb{P}[S = \text{B}] \cdot F_{\text{B}}^{-1} \circ F_{\text{A}}(m(\mathbf{x}, s = \text{A})) \end{aligned}$$



Mitigation with Wasserstein Barycenter

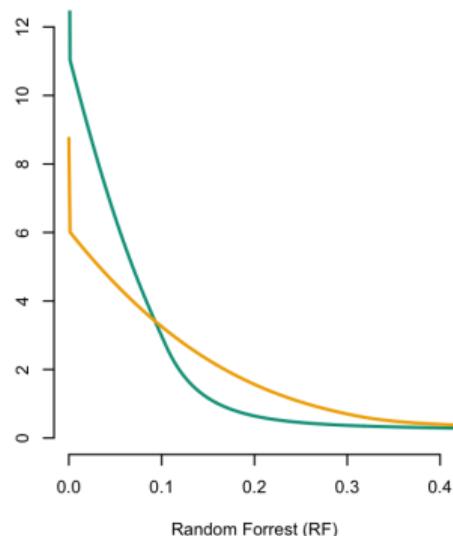
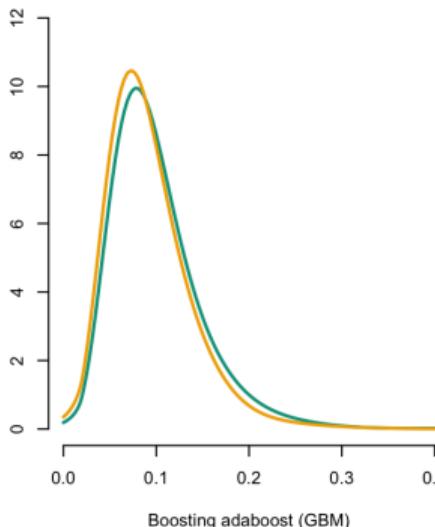
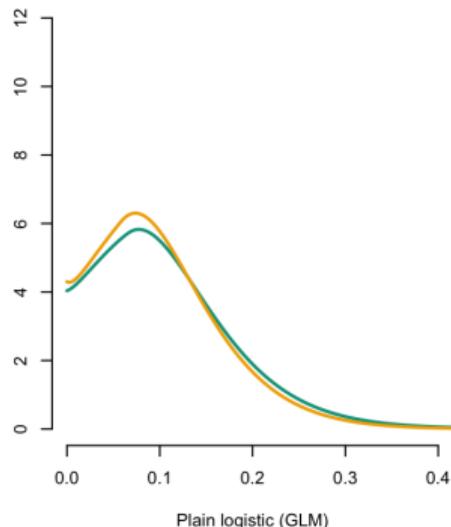
Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$,
the “fair barycenter score” is

$$\begin{aligned} & m^*(\mathbf{x}, s = \text{B}) \\ = & \mathbb{P}[S = \text{A}] \cdot F_{\text{A}}^{-1} \circ F_{\text{B}}(m(\mathbf{x}, s = \text{B})) \\ + & \mathbb{P}[S = \text{B}] \cdot m(\mathbf{x}, s = \text{B}) \end{aligned}$$



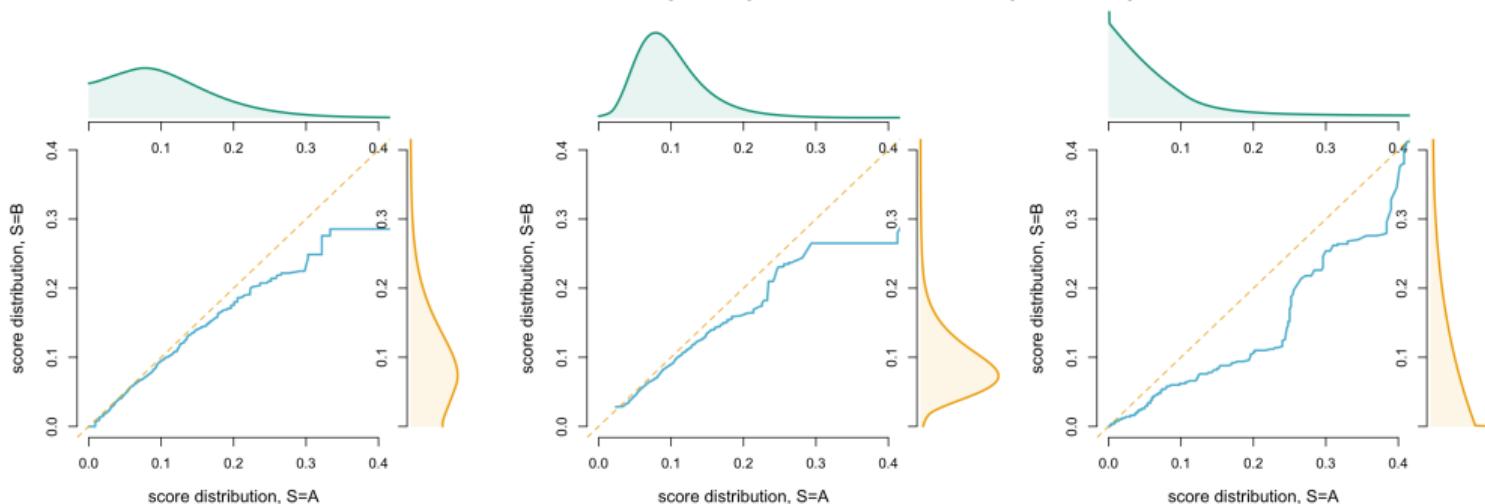
Mitigation with Wasserstein Barycenter

- › If the two models are balanced, m^* is also balanced.
- › Annual claim occurrence (motor insurance, Charpentier et al. (2023b))
- › Three models (plain GLM, GBM, Random Forest)



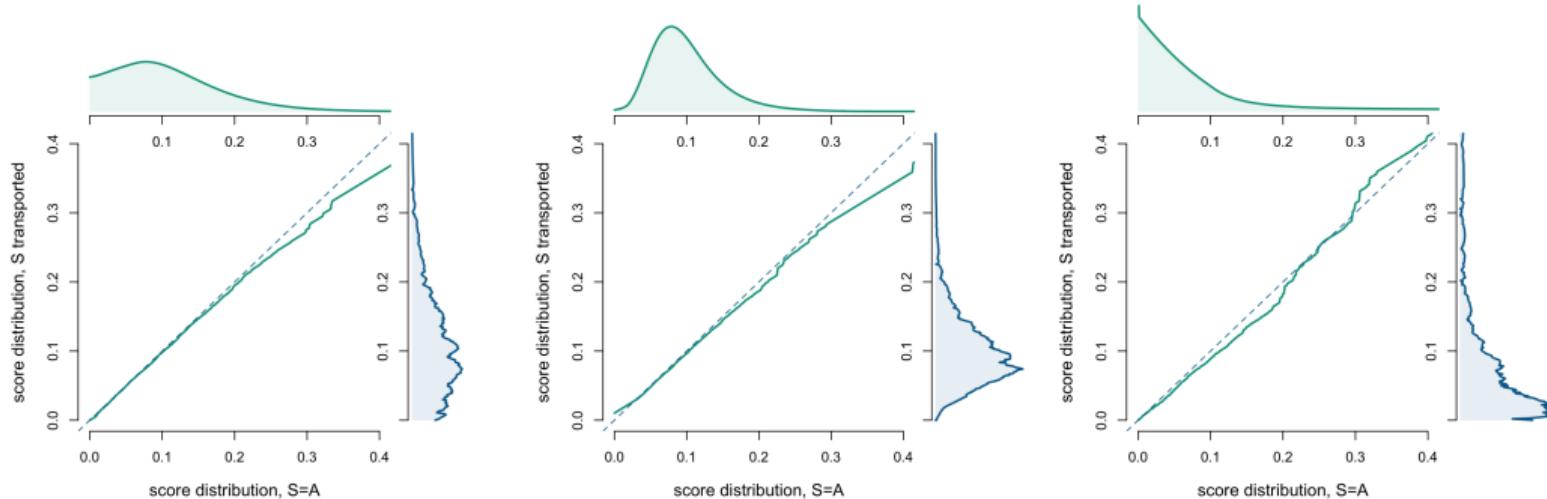
Mitigation with Wasserstein Barycenter

- › Predictions are different for men ($= A$) and women ($S = B$)



- › since $W_2 \neq 0$ consider post processing mitigation

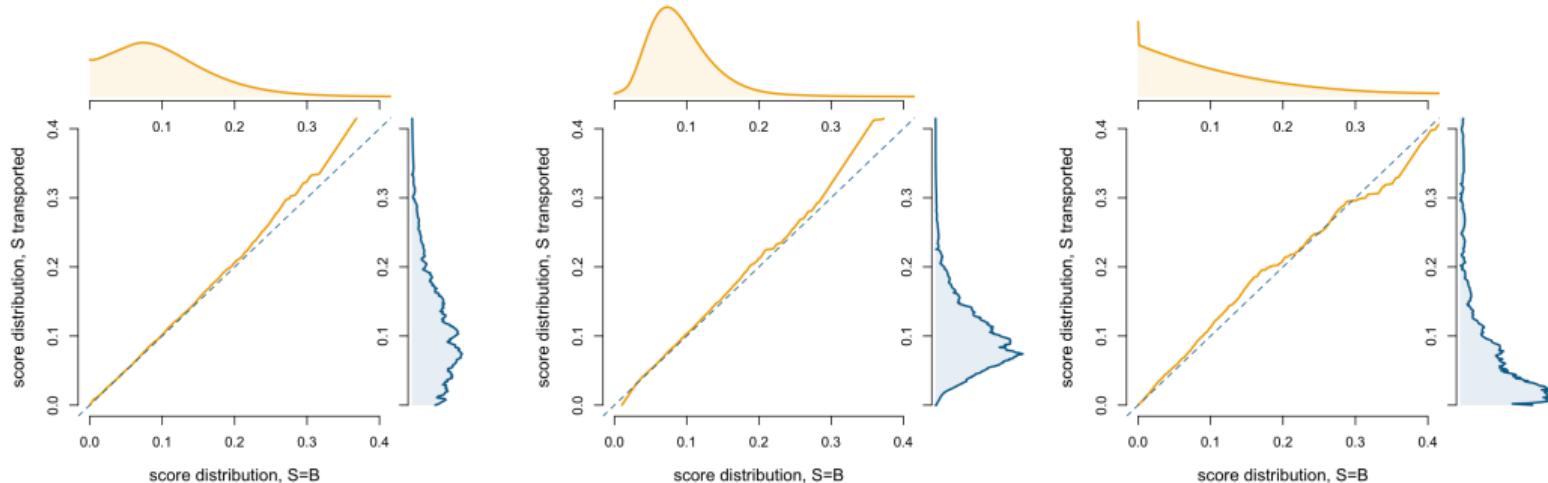
Mitigation with Wasserstein Barycenter



- Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$, the “fair barycenter score” is

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

Mitigation with Wasserstein Barycenter

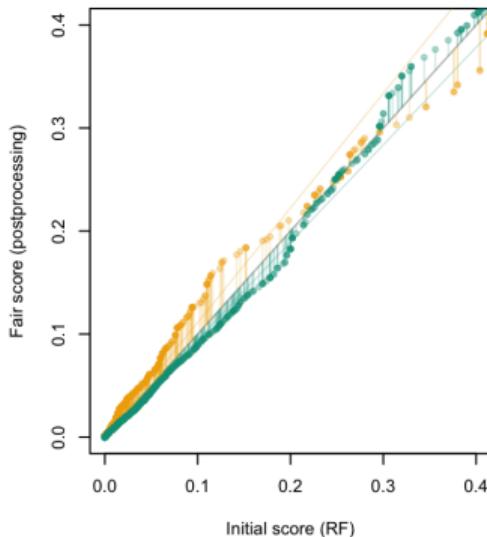
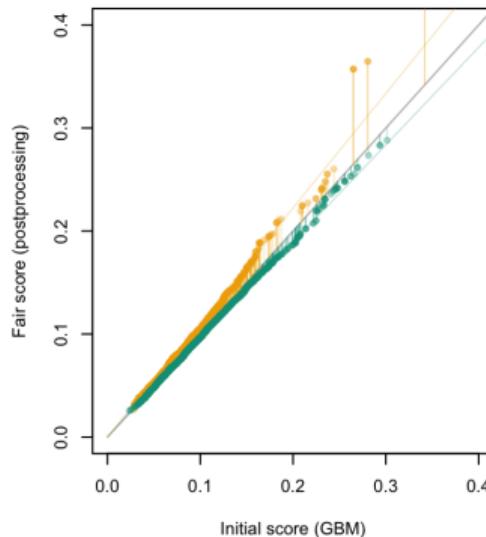
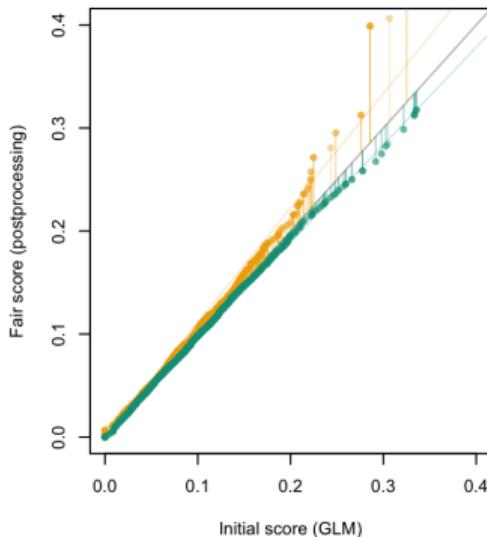


- Given scores $m(\mathbf{x}, s = \text{A})$ and $m(\mathbf{x}, s = \text{B})$, the “fair barycenter score” is

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

Mitigation with Wasserstein Barycenter

- › We can plot $\{(m(\mathbf{x}_i, \mathbb{A}), m^*(\mathbf{x}_i, \mathbb{A})\}$ and $\{(m(\mathbf{x}_i, \mathbb{B}), m^*(\mathbf{x}_i, \mathbb{B})\}$



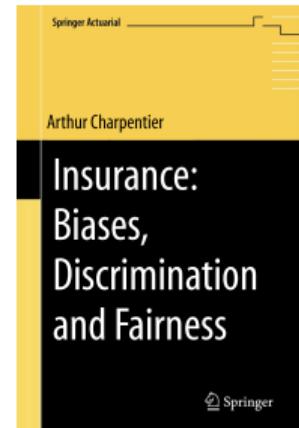
Mitigation with Wasserstein Barycenter

- Numerical values, for initial occurrence probability of 5%, 10% and 20%, we have

	A (men)				B (women)			
	$\times 0.94$	GLM	GBM	RF	$\times 1.11$	GLM	GBM	RF
$m(x) = 5\%$	4.73%	4.94%	4.80%	4.42%	5.56%	5.16%	5.25%	6.15%
$m(x) = 10\%$	9.46%	9.83%	9.66%	8.92%	11.12%	10.38%	10.49%	12.80%
$m(x) = 20\%$	18.91%	19.50%	18.68%	18.26%	22.25%	20.77%	21.63%	21.12%

Mitigation ? (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 (2023) Insurance: Biases, Discrimination and Fairness.



References

- Agarwal, S. (2021). Trade-offs between fairness and interpretability in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*.
- 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.
- Berk, R. A., Kuchibhotla, A. K., and Tchetgen, E. T. (2021). Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. *arXiv*, 2111.09211.
- Bénéplanc, G., Charpentier, A., and Thourot, P. (2022). *Manuel d'assurance*. Presses Universitaires de France.
- Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21.
- Charpentier, A. (2014). *Computational actuarial science with R*. CRC press.
- Charpentier, A. (2023). *Insurance: biases, discrimination and fairness*. Springer Verlag.

freakonometrics

freakonometrics.hypotheses.org

– Arthur Charpentier, 2023

32 / 35

References

- Charpentier, A., Flachaire, E., and Gallic, E. (2023a). Causal inference with optimal transport. In Thach, N. N., Kreinovich, V., Ha, D. T., and Trung, N. D., editors, *Optimal Transport Statistics for Economics and Related Topics*. Springer Verlag.
- Charpentier, A., Hu, F., and Ratz, P. (2023b). Mitigating discrimination in insurance with wasserstein barycenters. *BIAS, 3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. *arXiv*, 1701.08230.
- Crossney, K. B. (2016). Redlining. <https://philadelphiaencyclopedia.org/essays/redlining/>.
- Denuit, M. and Charpentier, A. (2004). *Mathématiques de l'assurance non-vie: Tome I Principes fondamentaux de théorie du risque*. Economica.
- Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie: Tome II Tarification et provisionnement*. Economica.
- 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.

References

- Feeley, M. and Simon, J. (1994). Actuarial justice: The emerging new criminal law. *The futures of criminology*, 173:174.
- 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.
- Hu, F., Ratz, P., and Charpentier, A. (2023a). 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. (2023b). A sequentially fair mechanism for multiple sensitive attributes. *ArXiv*, 2309.06627.
- Kearns, M. and Roth, A. (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.
- 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.

References

- 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.
- Lippert-Rasmussen, K. (2020). *Making sense of affirmative action*. Oxford University Press.
- 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.
- Rhynhart, R. (2020). Mapping the legacy of structural racism in philadelphia. *Philadelphia, Office pf the Controller*.
- Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.
- The Zebra (2022). Car insurance rating factors by state. <https://www.thezebra.com/>.
- Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.