Fairness for Insurers and Actuaries

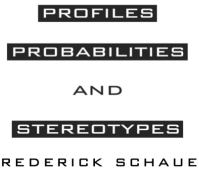
Arthur Charpentier
with Marie-Pier Côté, Olivier Côté, Agathe Fernandes-Machado
Ewen Gallic, François Hu & Philipp Ratz

ACPR/Télécom Paris webinar, April 2025

SCOR
FOUNDATION FOR SCIENCE

What is an "actuary"?

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


PROFILES

PROBABILITIES

AND

STEREOTYPES

FREDERICK SCHAUER

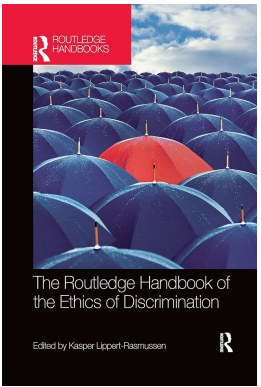
The Belknap Press of Harvard University Press
Cambridge, Massachusetts
London, England

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

Occasionally the actuarial generalizations of the insurance industry become controversial. One example is the use of generalizations about the comparative safety of different neighborhoods as a basis for setting the rates for homeowners' insurance or determining the willing-

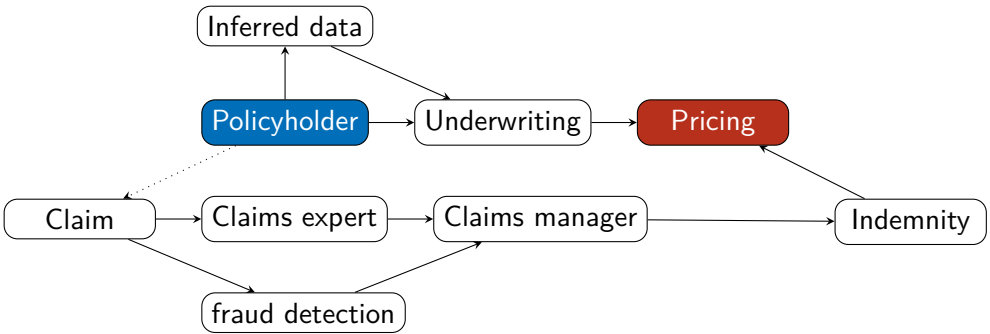
"At the core of insurance business lies discrimination".

"*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)

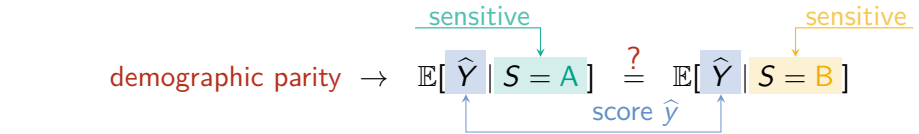


Decisions... decisions everywhere...

"*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*" [...] "*the subjective nature of a seemingly objective process*," Glenn (2000, 2003).



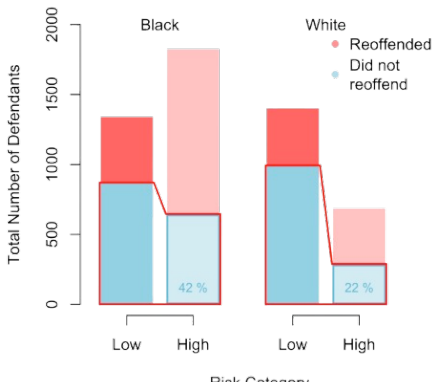
Several definitions of "fairness" or "non-discriminatory"



Isn't it a problem to have several definitions?

Consider a popular "**Actuarial Justice**" problem (Harcourt (2008)), **C**orrectional **O**ffender **M**anagement **P**rofiling for **A**lternative **S**anctions. From Feller et al. (2016),

- for White people, among those who did not re-offend (y), 22% were wrongly classified ($\widehat{y}$),
- for Black people, among those who did not re-offend, 42% were wrongly classified,
- **Problem**, since $42\% \gg 22\%$

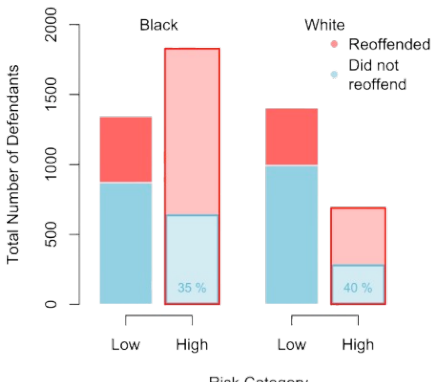


$$\mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{black}] = 42\% \stackrel{?}{=} \mathbb{P}[\widehat{Y} = \text{high} \mid Y = \text{no}, S = \text{white}] = 22\%,$$

Isn't it a problem to have several definitions?

Consider a popular "**Actuarial Justice**" problem (Harcourt (2008)), **COMPAS**. From Dieterich et al. (2016),

- for White people, among those who were classified as high risk ($\widehat{y}$), 40% did not re-offend (y),
- for Black people, among those who were classified as high risk ($\widehat{y}$), 35% did not re-offend (y),
- **No problem**, since $35 \approx 40\%$



$$\mathbb{P}[\, Y = \text{no} \mid \widehat{Y} = \text{high}, S = \text{black} \,] = 35\% \overset{?}{=} \mathbb{P}[\, Y = \text{no} \mid \widehat{Y} = \text{high}, S = \text{white} \,] = 40\%.$$

Is it always possible to have a sensitive-free model (with respect to ...)?

For **decisions** ($\widehat{y} \in \{0, 1\}$, e.g., "obtain a loan"),

decision $\widehat{y}$

$$\text{demographic parity} \rightarrow \mathbb{P}[\widehat{Y} = 1 \mid S = \text{A}] \stackrel{?}{=} \mathbb{P}[\widehat{Y} = 1 \mid S = \text{B}]$$

those decisions are usually based on **scores**, and **thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\widehat{m}(\boldsymbol{X}, S) > t \mid S = \text{A}] \stackrel{?}{=} \mathbb{E}[\widehat{m}(\boldsymbol{X}, S) > t \mid S = \text{B}]$$

score $\widehat{m}$

One can achieve demographic parity, simply selecting **different thresholds**

$$\text{demographic parity} \rightarrow \mathbb{E}[\widehat{m}(\boldsymbol{X}, S) > t_{\text{A}} \mid S = \text{A}] \stackrel{?}{=} \mathbb{E}[\widehat{m}(\boldsymbol{X}, S) > t_{\text{B}} \mid S = \text{B}]$$

(with that strategy, usually impossible to achieve equalized odds)

Is it always possible to have a sensitive-free model (with respect to ...)?

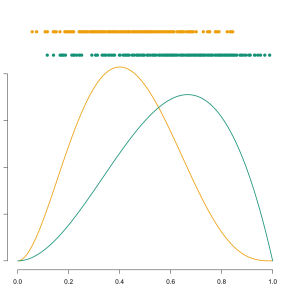
For **decisions** ($\widehat{y} \in \{0, 1\}$, e.g., "obtain a loan"), we considered

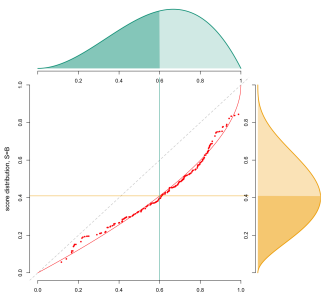
$$\text{demographic parity} \quad \rightarrow \quad \mathbb{E}[\,\widehat{Y} \,|\, S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{Y} \,|\, S = \mathrm{B}\,]$$

and we can consider the analogous for **scores** (possibly used to assess premiums),

$$\text{demographic parity} \quad \rightarrow \quad \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S) \,|\, S = \mathrm{A}\,] \overset{?}{=} \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S) \,|\, S = \mathrm{B}\,]$$

score $\widehat{y}$





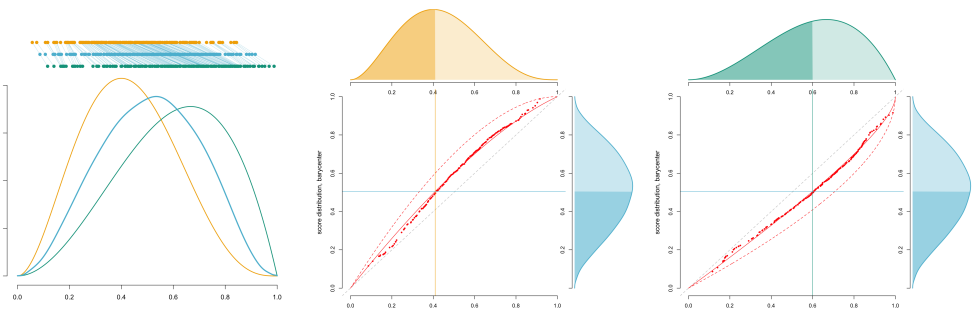
- individual in group A
  with a score $\widehat{y}(\mathrm{A}) = 60\%$
  corresponding to quantile α
  (here 0.5)
- in group B, the same
  quantile α
  corresponds to $\widehat{y}(\mathrm{B}) = 40\%$

Is it always possible to have a sensitive-free model (with respect to ...)?

- To get a fair model (**neutral with respect to s**), consider an average between the two models,

score in group A with quantile α — score in group B with quantile α

$$\widehat{y}^\star = \mathbb{P}[S = \mathsf{A}] \cdot \widehat{y}(\mathsf{A}) + \mathbb{P}[S = \mathsf{B}] \cdot \widehat{y}(\mathsf{B})$$



"In order to treat some persons equally, we must treat them differently"

- 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)

See philosophical discussions about **affirmative action**, e.g., Rubenfeld (1997); Pojman (1998); Anderson (2004)

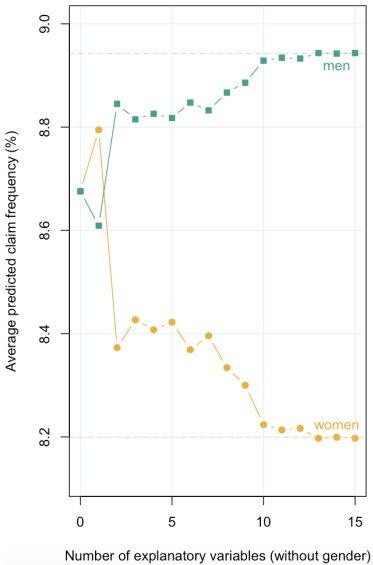
Discrimination in the data, or in the model?

On a French motor dataset, average claim frequencies are 8.94% (men) and 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 in the data, or in the model?

David Hume's "**is-ought**" problem, in Hume (1739)

what **is** observed, what is **statistically normal**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ where } \mathbb{P} \text{ is the historical probability}$$

$\neq$ what **should be**, what we expect from an **ethical norm**

$$\pi(\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}^\star}[Y|\boldsymbol{X} = \boldsymbol{x}] \text{ where } \mathbb{P}^\star \text{ is some "fair" probability}$$

"*keep in mind that machine learning can only be used to memorize patterns that are present in your training data. You can only recognize what you've seen before. Using machine learning trained on past data to predict the future is making the assumption that the future will behave like the past*," Chollet (2021)

Classical **clausula rebus sic stantibus** ("with things thus standing") in predictive modeling (statistics and machine learning)

Multiple sensitive attributes, "robbing Peter to pay Paul"?

sensitive attribute 1

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_1 = \mathrm{A}\,] \neq \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_1 = \mathrm{B}\,]$$

$$\mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_2 = \mathrm{C}\,] \approx \mathbb{E}[\,\widehat{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_2 = \mathrm{D}\,]$$

sensitive attribute 2

Distort model $\widehat{m}$ to achieve fairness with respect to S_1 $\longrightarrow$ model $\widetilde{m}$

sensitive attribute 1

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_1 = \mathrm{A}\,] = \mathbb{E}[\,\widetilde{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_1 = \mathrm{B}\,]$$

$$\mathbb{E}[\,\widetilde{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_2 = \mathrm{C}\,] \neq \mathbb{E}[\,\widetilde{m}(\boldsymbol{X}, S_1, S_2) \,|\, S_2 = \mathrm{D}\,]$$

sensitive attribute 2

What if we neither observe nor collect sensitive personal information (s) ?

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. Use of **BIFSG** (Bayesian Improved First Name Surname and Geocoding), from Elliott et al. (2009). Consider 12 people living near Atlanta, GA (Fulton & Gwinnett counties),

	last	first	county	city	zipcode	whi	bla	his	asi
2	RADLEY	OLIVIA	Fulton	Fairburn	30213	14	83	1	0
3	BOORSE	KEISHA	Fulton	Atlanta	30331	97	0	3	0
4	MAZ	SAVANNAH	Gwinnett	Norcross	30093	5	6	76	13
5	GAULE	NATASHIA	Gwinnett	Snellville	30078	67	19	14	0
6	MCMELLEN	ISMAEL	Gwinnett	Lilburn	30047	73	15	6	3
7	WASHINGTON	BRYN	Gwinnett	Norcross	30093	0	95	3	0

(ongoing *Predicting Unobserved Multi-Class sensitive Attributes : Enhancing Calibration with Nested Dichotomies for Fairness* with A.M. Patrón Piñerez, A. Fernandes Machado, & E. Gallic)

Discrimination, with different perspectives

- Regulatory perspective, "**group fairness**" (discussed previously)
- Policyholders perspective, "**individual fairness**"

A decision satisfies individual 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*."

- also named "**counterfactual fairness**" in Kusner et al. (2017), and should be related to classical causal inference problem, (conditional) average treatment effect (the "treatement" being the sensitive attribute),

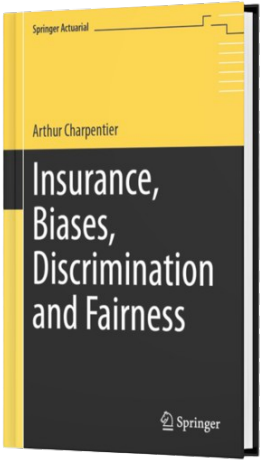
  "*other things being equal*" ?**ceteris paribus** ? See "revolving variable" in Kilbertus et al. (2017). Consider a men ($s = \mathrm{A}$) with height $x = 6'3$ (or 190 cm). If that person had been a women ($s = \mathrm{B}$) would she have height $x = 6'3$?

  (hint: no, consider similar quantiles, as discussed previously, see Charpentier et al. (2023a))

Conclusion (?)

- dealing with discrimination in insurance is tricky since actuarial pricing is deeply related to the idea of focusing on groups, and not individuals
- if we do not address properly those questions, there is no way we can get fair models
- not collecting and not using protected attributes is clearly not a good strategy
- there are still important questions that should be addressed by regulators, that should provide guidelines: portfolio fairness or market fairness? (what is $\mathbb{P}$?)

To go further, **Charpentier (2024) Insurance, Biases, Discrimination and Fairness. Springer**.



References

Anderson, T. H. (2004). *The pursuit of fairness: A history of affirmative action*. Oxford University Press.

Avraham, R. (2017). Discrimination and insurance. In Lippert-Rasmussen, K., editor, *Handbook of the Ethics of Discrimination*, pages 335–347. Routledge.

Charpentier, A. (2024). *Insurance: biases, discrimination and fairness*. Springer Verlag.

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. In *3rd Workshop on Bias and Fairness in AI, International Workshop of ECML PKDD*.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Côté, O., Côté, M.-P., and Charpentier, A. (2024). A fair price to pay: exploiting causal graphs for fairness in insurance. *forthcoming*.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(7.4):1.

References

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

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.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024a). From uncertainty to precision: Enhancing binary classifier performance through calibration. *arXiv preprint arXiv:2402.07790*.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024b). Post-calibration techniques: Balancing calibration and score distribution alignment. *Thirty-Eighth Annual Conference on Neural Information Processing Systems*.

Fernandes Machado, A., Charpentier, A., Flachaire, E., Gallic, E., and Hu, F. (2024c). Probabilistic scores of classifiers, calibration is not enough. *arXiv preprint arXiv:2408.03421*.

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.

References

Harcourt, B. E. (2008). *Against prediction*. University of Chicago Press.

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.

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

Knowlton, R. E. (1978). Regents of the university of california v. bakke. *Arkansas Law Review*, 32:499.

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.

Pojman, L. P. (1998). The case against affirmative action. *International Journal of Applied Philosophy*, 12(1):97–115.

References

Rubenfeld, J. (1997). Affirmative action. *Yale Law Journal*, 107:427.

Sabbagh, D. (2007). *Equality and transparency: A strategic perspective on affirmative action in American law*. Springer.

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

Turner, R. (2015). The way to stop discrimination on the basis of race. *Stanford Journal of Civil Rights & Civil Liberties*, 11:45.