# From Fair Models to Fair Insurance Systems

## What we have learned — and what we are still missing

Arthur Charpentier

Atelier sur l'Équité et la Discrimination en Assurance (AEDA) – October 2026

# What have we learnt?

**Define fairness**
choose what should be equal

**Measure disparity**
compare group distributions

**Explain disparity**
separate sources of inequality

**Pre-processing**
transform inputs

**In-processing**
constrain learning

**Post-processing**
transform predictions

**Fairness requires a target, a measure, and possibly an intervention.**

# Insurance starts by making differences

## Actuarial problem

Estimate risk and classify policyholders:

$$\mu(x, s) = \mathbb{E}[Y \mid X = x, S = s].$$

Better information can improve prediction, pricing and prevention.

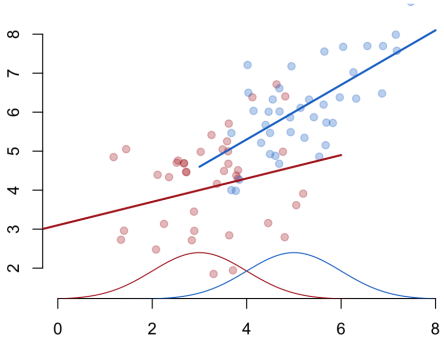
## Fairness problem

Some distinctions are legally or socially restricted; others can act as proxies.

risk differentiation $\neq$ legitimate differentiation.

- The difficulty is not that insurance discriminates, it is to decide which distinctions should remain in the price.
- Background: Charpentier (2024) for a general survey

# Fairness by awarness ("actuarial fairness")




- best estimate (on full information)

$$\begin{cases} \mathbb{E}[Y \mid X, s=1] = \alpha_1 + \beta_1 x \\ \mathbb{E}[Y \mid X, s=0] = \alpha_0 + \beta_0 x \end{cases}$$

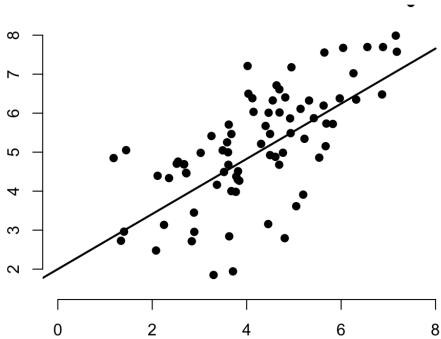
obtained solving

$$(\alpha_1, \beta_1) = \underset{a,b}{\arg\min} \sum_{i:s_i=1} (y_i - (a+bx_i))^2$$

$$(\alpha_0, \beta_0) = \underset{a,b}{\arg\min} \sum_{i:s_i=0} (y_i - (a+bx_i))^2$$

See e.g., Denuit et al. (2019); Charpentier (2024)

# Fairness by **un**awarness




See e.g., Charpentier (2024)

- "blind" with respect to $s$
- best estimate (on $x$ only)

$$\mathbb{E}[Y \mid X] = \alpha + \beta x$$

obtained solving

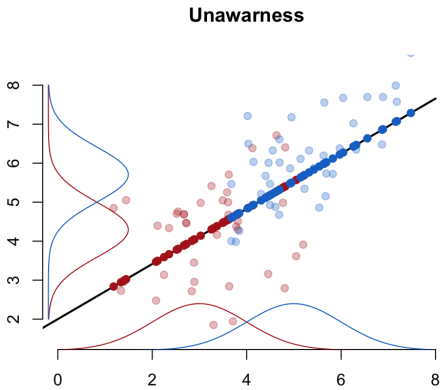
$$(\alpha, \beta) = \arg\min_{a,b} \sum_i \big(y_i - (a + bx_i)\big)^2$$

- if $X \mid S = s \sim \mathcal{N}(\mu_s, \sigma_s^2)$,

$$b = \frac{\pi_0\beta_0\sigma_0^2 + \pi_1\beta_1\sigma_1^2 + \pi_0\pi_1(\mu_1 - \mu_0)(m_1 - m_0)}{\pi_0\sigma_0^2 + \pi_1\sigma_1^2 + \pi_0\pi_1(\mu_1 - \mu_0)^2},$$

$a = \bar{m} - b\bar{\mu}$, where $m_s = \alpha_s + \beta_s\mu_s$,
$\bar{\mu} = \pi_0\mu_0 + \pi_1\mu_1$, $\bar{m} = \pi_0 m_0 + \pi_1 m_1$.

# Fairness by **un**awarness




- Hence, in general

$$\underbrace{(\widehat{Y} \mid S=0)}_{\sim\mathcal{N}\left(a+b\mu_0,\, b^2\sigma_0^2\right)} \overset{\mathcal{L}}{\neq} \underbrace{(\widehat{Y} \mid S=1)}_{\sim\mathcal{N}\left(a+b\mu_1,\, b^2\sigma_1^2\right)}$$

(no strong demographic parity)

$$\mathbb{E}[\widehat{Y} \mid S=0] \neq \mathbb{E}[\widehat{Y} \mid S=1]$$

(not even weak demographic parity)

Idea $\mathbb{P}[S \mid X] \neq \mathbb{P}[S]$, Lindholm et al. (2022) (propensity score).

# An old question: where does a disparity come from?

For two groups, (Kitagawa)–Oaxaca–Blinder separates an average outcome gap into differences in *characteristics* and differences in *coefficients*:

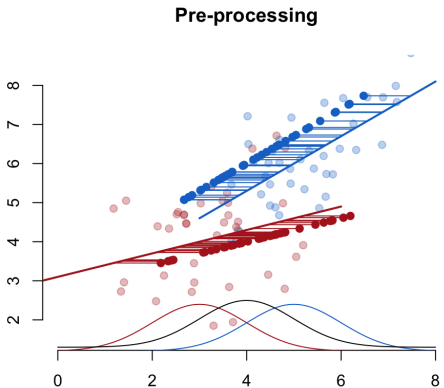
$$\bar{Y}_0 - \bar{Y}_1 = \underbrace{(\bar{X}_0 - \bar{X}_1)^\top \beta_1}_{\text{composition / characteristics}} + \underbrace{\bar{X}_0^\top (\beta_0 - \beta_1)}_{\text{returns / coefficients}} .$$

**Modern fairness work keeps this intuition, but changes the object:** from observed outcome gaps to disparities generated by a fitted model.
Kitagawa (1955); Oaxaca (1973); Blinder (1973).

| Direct | Indirect | Interaction | Structure |
|---|---|---|---|
| explicit contribution of $S$ | different $X$ profiles across groups | direct and indirect effects reinforce / offset | different within-group covariance / dispersion |

With GLMs, the nonlinear inverse link adds *curvature coupling* and *curvature amplification*. Linear decomposition: Tierny et al. (2026a); GLM extension and Oaxaca–Blinder connection: Tierny et al. (2026b).

# Route A — change the information used by the model




See e.g., Charpentier (2024)

- Shift the covariates in $x \longrightarrow x'$ so that

$$(X' \mid S = 0) \overset{\mathcal{L}}{=} (X' \mid S = 1)$$

- If $m_s = \alpha_s + \beta_s \mu_s$ and $\tau_s = \beta_s \sigma_s$,

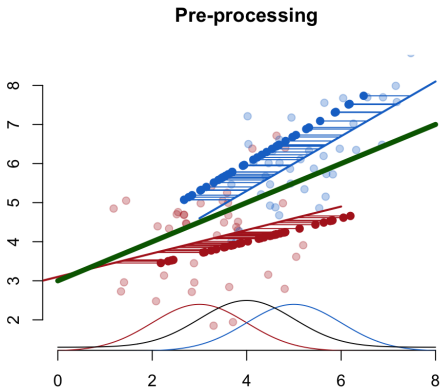
$$x' \longleftarrow \mu^\star + \frac{\sigma^\star}{\sigma_s}(x - \mu_s).$$

- Then

$$(X' \mid S = s) \sim \mathcal{N}(m^\star, \tau^\star), \ \forall s,$$

$$m^\star = \sum_s \pi_s m_s \text{ and } \tau^\star = \sum_s \pi_s \tau_s.$$

# Route A — change the information used by the model




- Fit a model on $(x', \hat{y})$

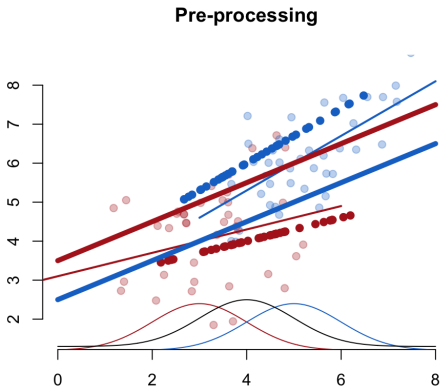
$$(\alpha_\star, \beta_\star) = \arg\min_{a,b} \sum_i \left(y_i - (a + bx_i')\right)^2$$

i.e.,

$$\underbrace{m^\star - \frac{\tau^\star}{\sigma^\star}\mu^\star}_{=\alpha_\star} + \underbrace{\frac{\tau^\star}{\sigma^\star}}_{=b_\star} x'.$$

See e.g., Charpentier (2024)

# Route A — change the information used by the model




- Then shift back $x' \longrightarrow x$

$$\widehat{y}_{s,\star}(x) = m^\star + \frac{\tau^\star}{\sigma_s}(x - \mu_s)$$

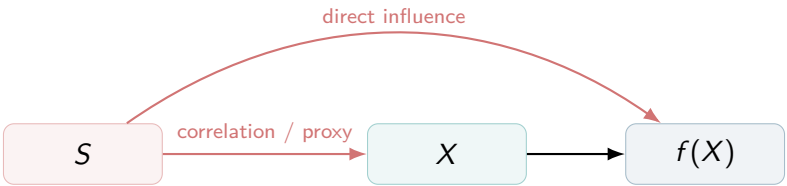
i.e,

$$\begin{cases} \widehat{y}_{s=1,\star}(x) = \alpha_1^\star + \beta_1^\star x \\ \widehat{y}_{s=0,\star}(x) = \alpha_0^\star + \beta_0^\star x \end{cases}$$

See e.g., Charpentier (2024) or Côté et al. (2025).

# Route A — change the information used by the model

But if $X$ carries information about $S$, the unaware approach does not remove proxy discrimination, see Causal Mediation Analysis, Fernandes Machado et al. (2026c).




- Lindholm et al. (2022): construct prices free of direct and proxy discrimination.
- Lindholm et al. (2024): proxy discrimination and demographic disparities are distinct objects.
- Input-side adversarial / representation approaches provide another way to reduce recoverability of $S$ from $X$, Grari et al. (2022).

# Route B — change the optimization problem

- Minimize quadratic risk

$$(\alpha_\star, \beta_\star) = \underset{(a_0,a_1),(b_0,b_1)}{\arg\min} \sum_i \big(y_i - (a_{s_i} + b_{s_i} x_i)\big)^2$$

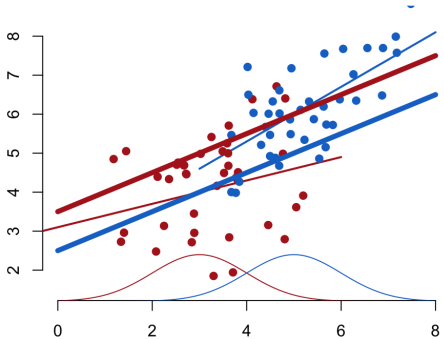
under the constraint that

$$\underbrace{(\widehat{Y} \mid S=0)}_{\sim\mathcal{N}(a_0+b_0\mu_0, b_0\sigma_0)} \overset{\mathcal{L}}{=} \underbrace{(\widehat{Y} \mid S=1)}_{\sim\mathcal{N}(a_1+b_1\mu_1, b_1\sigma_1)}$$

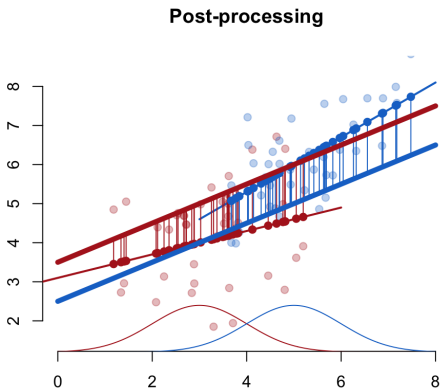
i.e., $\gamma_s + \delta_s\mu_s = m$ and $\delta_s\sigma_s = \tau$.

$$\min_{m,\tau} \sum_s \pi_s \sum_{i:s_i=s} \left(y_i - m - \frac{\tau}{\sigma_s}(x_i - \mu_s)\right)^2$$

$$\widehat{y}_{s,\star}(x) = m^\star + \frac{\tau^\star}{\sigma_s}(x - \mu_s)$$



# Route C — change the distribution of prices




- Use best estimate (on full information)

$$\begin{cases} \widehat{y}_{s=1}(x) = \alpha_1 + \beta_1 x \\ \widehat{y}_{s=0}(x) = \alpha_0 + \beta_0 x \end{cases}$$

- Shift the predictions $\widehat{y} \xrightarrow{T_s} \widehat{y}'$

$$\underbrace{(\widehat{Y}' \mid S=0)}_{\sim\mathcal{N}(a_0+b_0\mu_0, b_0\sigma_0)} \overset{\mathcal{L}}{=} \underbrace{(\widehat{Y}' \mid S=1)}_{\sim\mathcal{N}(a_1+b_1\mu_1, b_1\sigma_1)}$$

and again, since $\widehat{y}_{s,\star}(x) = T_s(\widehat{y}_s(x))$,

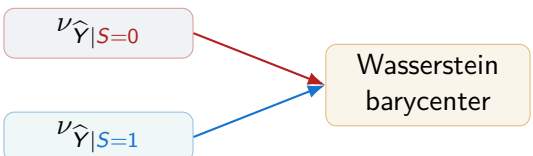
$$\widehat{y}_{s,\star}(x) = m^\star + \frac{\tau^\star}{\sigma_s}(x - \mu_s)$$

# Route C — change the distribution of prices

Distributional demographic parity asks for

$$(\widehat{Y} \mid S = 0) \overset{\mathcal{L}}{=} (\widehat{Y} \mid S = 1) .$$

In one dimension, optimal transport makes the correction particularly transparent: group-conditional score laws can be mapped toward a common Wasserstein barycenter.

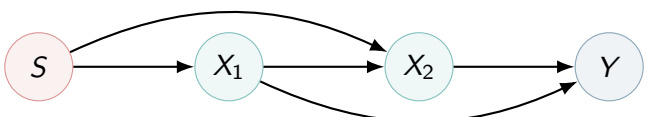



This gives a model-agnostic post-processing mechanism and extends naturally to approximate / sequential fairness.
Sequential multi-sensitive post-processing: Hu et al. (2024); calibration and post-processing cautions: Fernandes Machado et al. (2024a).

# Route A' — counterfactual transport with a mechanism

Pure input transport can destroy meaningful dependencies. Pure output transport may hide the mechanism. A causal / graphical route asks:

*If $S$ were different, which covariates should change with it — and in what order?*




- Causal graphs distinguish direct and indirect pathways in insurance pricing.
- Sequential transport combines a DAG with conditional optimal transport to build interpretable counterfactuals.

See Fernandes Machado et al. (2025a,b) and Fernandes Machado et al. (2026c).

# Fairness is also a diagnostic problem: where does the disparity live?

**Linear models**
Coefficient-level decomposition of direct and indirect effects under a fairness correction.

**GLMs**
The inverse link can compress or amplify group disparities through curvature.

**Trees**
Localize where a post-processing correction changes information gain across splits / features.

A complementary perspective measures the *distance* to the set of non-proxy-discriminatory prices and attributes it to covariates using sensitivity analysis / Shapley values.

**declare unfair?** $\longrightarrow$ **measure how much** $\longrightarrow$ **locate where / why**

Tierny et al. (2026a,b); Hu et al. (2026); and Lindholm et al. (2026).

# Fairness constraints do not replace actuarial constraints

A premium still has to behave like a premium.

### Balance / calibration

Premiums should match claims at relevant aggregation levels.

### Multicalibration

Combine actuarial autocalibration with fairness-type conditional mean constraints.

### Several criteria

Individual fairness, group fairness and actuarial adequacy can be studied jointly.

- Denuit et al. (2026): balance + fairness through multicalibration.
- Jamotton and Hainaut (2026): multiple fairness paradigms in one probabilistic pricing model.
- Miao and Pesenti (2026): pricing principles made insensitive to protected-covariate perturbations.

Post-calibration can improve calibration while worsening distributional alignment: Fernandes Machado et al. (2024a).

# What assumptions did we make?

$S$ is **single**

$S$ is **observed**

$S$ is **well-defined**

$Y$ is **scalar**

portfolio $=$ **population**

prediction $=$ **decision**

**Most open problems begin**
**by dropping one of these assumptions.**

# Problem 0 — what is $\widehat{Y}$ ? what does an individual probability mean?

For a binary claim indicator, insurance needs a probability, not only a classification:

$$\widehat{p}_i = \widehat{\mathbb{P}}(Y_i = 1 \mid \mathcal{I}_i) \in [0, 1].$$

Here $\mathcal{I}_i$ specifies the information used, for a stated event and time horizon.

## An individual prediction

The score is assigned to one person, but learned from comparisons with others. Different information sets and reference classes can produce different probabilities for that person.

## A collective validation

Calibration requires $\mathbb{E}[Y \mid \widehat{p}] = \widehat{p}$.
It checks average outcomes among people receiving similar scores. One observed outcome cannot establish the correctness of one person's probability.

- Denuit et al. (2021); Fernandes Machado et al. (2024a,b); Denuit et al. (2026)
- Reichenbach (1971); Roth and Tolbert (2025); Barry and Charpentier (2027)
  Ongoing work on individual probabilities and actuarial pricing.

# Open problem 1 — more than one sensitive attribute

Marginal fairness is not joint fairness:

$$\widehat{Y} \perp S_1 \text{ and } \widehat{Y} \perp S_2 \quad \not\Rightarrow \quad \widehat{Y} \perp \boldsymbol{S} = (S_1, S_2).$$

Three related questions emerge:

1. How do we detect purely intersectional disparities?
2. How do we correct several sensitive dimensions without order artifacts?
3. How do we attribute a joint disparity back to individual dimensions?

**sequential transport** **joint Wasserstein** **Shapley allocation** **KL barycenter**

Hu et al. (2024), Fernandes Machado et al. (2026a,b), or Miao and Pesenti (2026).

# Open problem 2 — what if the sensitive attribute is only partly observed?

In practice, protected attributes may be available only in a designed audit, while a demographic proxy $q_i$ exists for everyone.

A useful principle is:

$$\widetilde{S}_i = q_i + \frac{R_i}{\pi_i}(S_i - q_i).$$

### Identification

The probability audit identifies the complete-demographics fairness target.

### Efficiency

The proxy is auxiliary information: better $q_i$ reduces variance, but does not define demographic "truth".

Use demographic inference to gain precision, not to silently replace identity.
On going work, Huynh et al. (2027).

# Open problem 3 — what if $S$ itself is not well-defined?

A fairness audit must operationalize identity. But several non-equivalent categories may coexist:

self-identified | administrative | socially perceived | algorithmically inferred

The category used for an audit can differ from the category through which a system effectively treats people:

$$S^{\text{audit}} \neq S^{\text{treatment}}.$$

Open question: fairness with respect to *which* identity — and who gets to define it?
Algorithmic ascription and the distinction between identity-for-audit and identity-for-treatment: Charpentier (2026a).

# Open problem 4 — what if the output is multivariate?

Insurance often prices several coverages simultaneously:

$$\widehat{\boldsymbol{Y}} = (\widehat{Y}_{\text{liability}}, \widehat{Y}_{\text{collision}}, \widehat{Y}_{\text{theft}}, \ldots) \in \mathbb{R}^d.$$

Fairness of each coverage separately does *not* imply fairness of the joint premium vector:

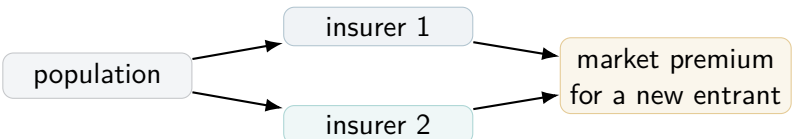
$$\widehat{Y}_j \perp S, \ \forall j \quad \not\Rightarrow \quad \widehat{\boldsymbol{Y}} \perp S.$$

This is the output-side analogue of intersectionality:

Joint multivariate post-processing via entropic / sliced Wasserstein barycenters: Huynh and Charpentier (2026).

# Open problem 5 — local or global fairness ?

A fairness correction is always relative to a reference distribution. But insurer portfolios are selected subpopulations.




A rule can satisfy a fairness criterion *within each insurer* and still produce direct or proxy discrimination at market level when portfolio inclusion depends on $X$ and/or $S$.

Côté et al. (2026), what is the correct unit of fairness — policy, portfolio, insurer, or market?

# Open problem 6 — what about privacy?

Insurers hold separate datasets. The relevant fairness question may concern their combined population.

## Why local audits can mislead

Group composition and score distributions vary across institutions.
Local demographic parity need not imply global demographic parity.

## A federated audit

Each institution sends group counts and a small quantile sketch of each group-score distribution.
One communication round allows the server to approximate global distributional disparity without pooling individual records.

**The population being audited matters as much as the metric.**

Charpentier et al. (2026) for mathematical guarantees.

# Open problem 7 — prediction is not the final decision

Fairness of uncertainty quantification creates its own trade-offs.

### Conformal prediction

Pooled calibration can hide group-wise under-coverage. Group calibration can equalize coverage, but heterogeneity can reappear as unequal set size.

### Downstream use

Procedural guarantees do not automatically imply equitable downstream outcomes when people or systems act on prediction sets.

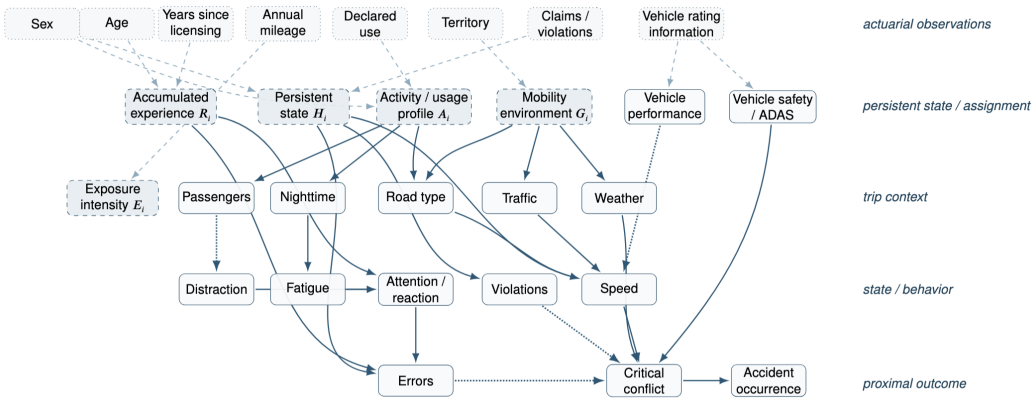
**equalized coverage** $\leftrightarrow$ **equalized set size** $\longrightarrow$ **substantive fairness**

**Calibration does not remove heterogeneity**

Gao et al. (2026); downstream substantive fairness: Liu et al. (2026).

# Open problem 8 — from prediction to causal mechanisms

A rating factor can predict claims without identifying the mechanism that generates the risk. Annual rating variables, trip-level driving states, crashes and recorded claims are different objects (important for prevention)




Multi-resolution causal framework in Charpentier (2026b)

# From risk modeling to pricing, from fairness to welfare

The distinction“ risk factors vs. rating factors ” in Lindholm’s talk points to a broader problem (moving from actuarial science to public policy)

### Risk sharing

Which differences become individual charges? Which remain pooled, and who bears the transfers?

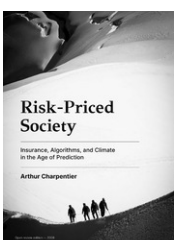
### Effective protection

Who can afford coverage? How much loss remains with households through deductibles, exclusions or withdrawal?

### Risk reduction

Does better information finance prevention and give people the means to act on their exposure?

- A fairness metric evaluates a property of predictions.
- A welfare assessment also examines who receives protection, who bears losses, and how the system changes exposure.
- When should predictive information change a price,
- And when should it guide prevention or collective protection?
- See Charpentier (2026c) for more (currently under open review)



# References


Barry, L. and Charpentier, A. (2027). What does an actuarial risk score mean? calibration, reference classes, and individual risk. *On going*.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, 8(4):436–455.

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

Charpentier, A. (2026a). Algorithmic ascription: Inferred identity and the governance of algorithmic discrimination. Working paper.

Charpentier, A. (2026b). From risk prediction to risk mechanisms: A multi-resolution causal representation for road safety and motor insurance. *arXiv*, 2608.09441.

Charpentier, A. (2026c). *Risk-Priced Society*. https://freakonometrics.github.io/risk-priced-society/.

Charpentier, A., Machado, A. F., Côté, O., and Hu, F. (2026). Federated measurement of demographic disparities from quantile sketches. *ArXiv*.

Côté, M.-P., Côté, O., and Charpentier, A. (2026). Recoverability of market-wide fair insurance premiums under selection bias. *Insurance: Mathematics and Economics*, 129:103268.

Côté, O., Côté, M.-P., and Charpentier, A. (2025). A fair price to pay: Exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*, 92:33–75.

# References


Denuit, M., Charpentier, A., and Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497.

Denuit, M., Hainaut, D., and Trufin, J. (2019). *Effective statistical learning methods for actuaries I: GLMs and extensions*, volume 1. Springer.

Denuit, M., Michaelides, M., and Trufin, J. (2026). Balance and fairness through multicalibration in nonlife insurance pricing.

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

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

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025a). Optimal transport on categorical data for counterfactuals using compositional data and dirichlet transport. *arXiv*, 2501.15549.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2025b). Sequential conditional transport on probabilistic graphs for interpretable counterfactual fairness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(18):19358–19366.

# References


Fernandes Machado, A., Jamotton, C., and Charpentier, A. (2026a). Multi-sensitive distributional fairness: Joint wasserstein audits and shapley allocation. Preprint; under review.

Fernandes Machado, A., Jamotton, C., and Charpentier, A. (2026b). Robbing peter to pay paul: Cross-attribute externalities of fairness interventions. On going work.

Fernandes Machado, A., Voitsitska, I., Charpentier, A., and Gallic, E. (2026c). Graph-aligned sequential transport for distributional disparity decomposition. *ArXiv*, 2603.15182.

Gao, Z., Liu, P., Yang, A. Y., Belbahri, M., Cresswell, J. C., and Asgharian, M. (2026). On the burden of achieving fairness in conformal prediction. *arXiv*, 2605.14260.

Grari, V., Charpentier, A., and Detyniecki, M. (2022). A fair pricing model via adversarial learning.

Hu, F., Fernandes Machado, A., and Charpentier, A. (2026). TRACE: Statistical inference for localizing the structural effects of fairness corrections in tree-based models. Preprint; under review.

Hu, F., Ratz, P., and Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Huynh, S. and Charpentier, A. (2026). A transport framework for fair insurance premiums across multiple coverages. Working paper.

Huynh, S., Jamotton, C., Fernandes Machado, A., and Charpentier, A. (2027). Audit-certified fair GLMs with partially observed demographics. On going work.

# References


Jamotton, C. and Hainaut, D. (2026). A multi-criteria fair gaussian regressor for insurance premium. Discussion Paper ISBA 2026/03, LIDAM, UCLouvain.

Kitagawa, E. M. (1955). Components of a difference between two rates. *Journal of the american statistical association*, 50(272):1168–1194.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin*, 52(1):55–89.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2024). What is fair? proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*, (9):935–970.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2026). Sensitivity-based measures of discrimination in insurance pricing. *European Journal of Operational Research*, 333:601–615.

Liu, P., Yu, Z., Belbahri, M., Charpentier, A., Asgharian, M., and Cresswell, J. C. (2026). Beyond procedure: Substantive fairness in conformal prediction. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*.

Miao, K. E. and Pesenti, S. M. (2026). Discrimination-insensitive pricing. *arXiv*, 2603.16720.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 14(3):693–709.

# References


Reichenbach, H. (1971). *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability*. University of California Press, Berkeley, CA, 2 edition.

Roth, A. and Tolbert, A. W. (2025). Resolving the reference class problem at scale. *Philosophy of Science*, 92(4):868–882.

Tierny, B., Charpentier, A., and Hu, F. (2026a). Decomposing direct and indirect biases in linear models under demographic parity constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tierny, B., Charpentier, A., and Hu, F. (2026b). Direct and indirect discrimination in generalized linear models. *ArXiv*, 2606.23116.