Bias diagnostics for Black-Box AI Systems

François HU
Milliman & ISFA
Lead AI engineer & Associate Professor

Bertille TIERNY
Milliman
AI engineer & Actuary

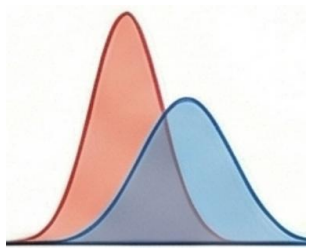
6th European
Congress of Actuaries
www.eca2026.org

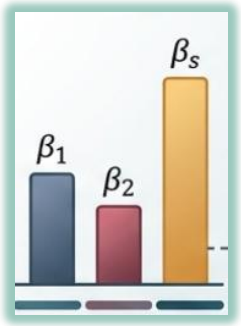
When Fairness Turns Linear Models into Black Boxes

Why standard coefficients no longer explain feature influence under fairness constraints

Standard (linear model)

$$Y = \beta X + \beta_0$$



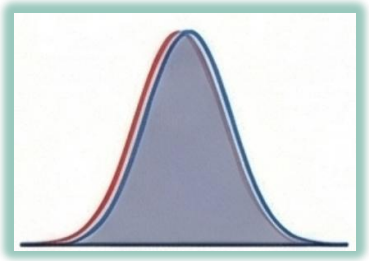


✓ **Interpretable coefficients:** we directly see how each feature affects the outcome.
Easy audit: simple to justify decisions to regulators or reviewers.

✗ **Unfair outcomes:** the model may treat groups differently without control.
No protection: nothing prevents the use of proxies or hidden bias.

SOTA (with fairness constraint)

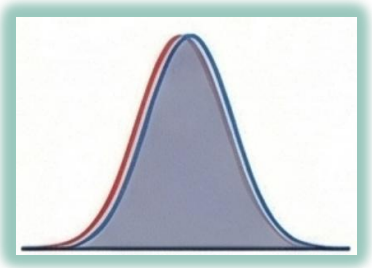
$$\cancel{Y = \beta X + \beta_0}$$

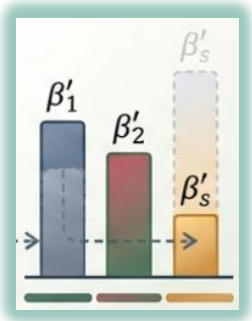


✓ **Fairness improved:** measurable gaps between groups are reduced.
Regulatory alignment: easier to show efforts toward fairness compliance.

✗ **Lost interpretability:** coefficients no longer reflect the true decision logic.
Hidden bias: fairness constraints can hide which features still drive unfairness.

Our method (fairness constraint)





Fairness improved: measurable gaps between groups are reduced.
Regulatory alignment: easier to show efforts toward fairness compliance.

✓ **Interpretability restored:** we recover a meaningful view of how each feature is used after fairness adjustments.
Bias visibility: we can see remaining bias and how fairness constraints changed the model.

The General Case

Direct Bias ☐

Indirect (Mean) Bias ☐

Interaction ☐

Indirect Structural Bias ☐

(Quadratic) Risk: $\mathcal{R}(f) := \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$

Gaussian Linear framework:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta^*} \rangle + \gamma^* S + \beta_0^* + \zeta,$$
$$\boldsymbol{X} \mid S = s \sim \mathcal{N}(\boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})$$

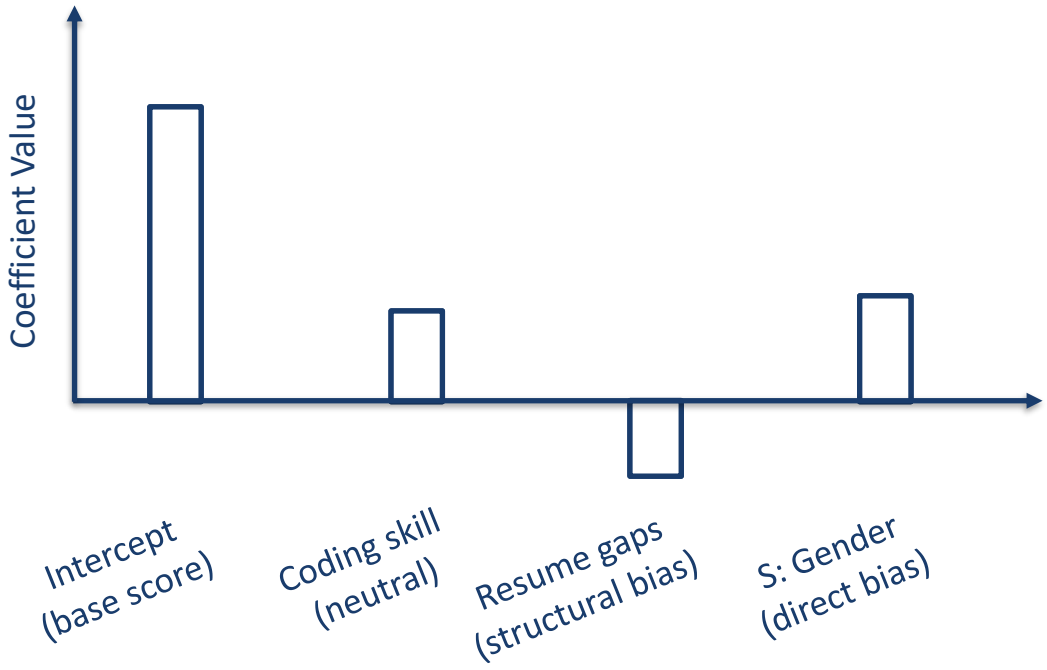
Bayes Optimal:

$$f^* \in \arg\min_f \mathcal{R}(f)$$
$$f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta^*} \rangle + \gamma^* s + \beta_0^*$$

Fairness constraint:

$\mathbb{P}(f(X,S) = 1 | S = $ 🚹 $) = \mathbb{P}(f(X,S) = 1 | S = $ 🚺 $)$

Hiring score



The General Case

Direct Bias ☐

Indirect (Mean) Bias ☐

Interaction ☐

Indirect Structural Bias ☐

(Quadratic) Risk: $\mathcal{R}(f) := \mathbb{E}(Y - f(\boldsymbol{X}, S))^2$

Gaussian Linear framework:

$$Y = \langle \boldsymbol{X}, \boldsymbol{\beta^*} \rangle + \gamma^* S + \beta_0^* + \zeta,$$

Left: Women $(S = 0)$ Right: Men $(S = 1)$

Bayes Optimal:

$$f^* \in \arg\min_f \mathcal{R}(f)$$

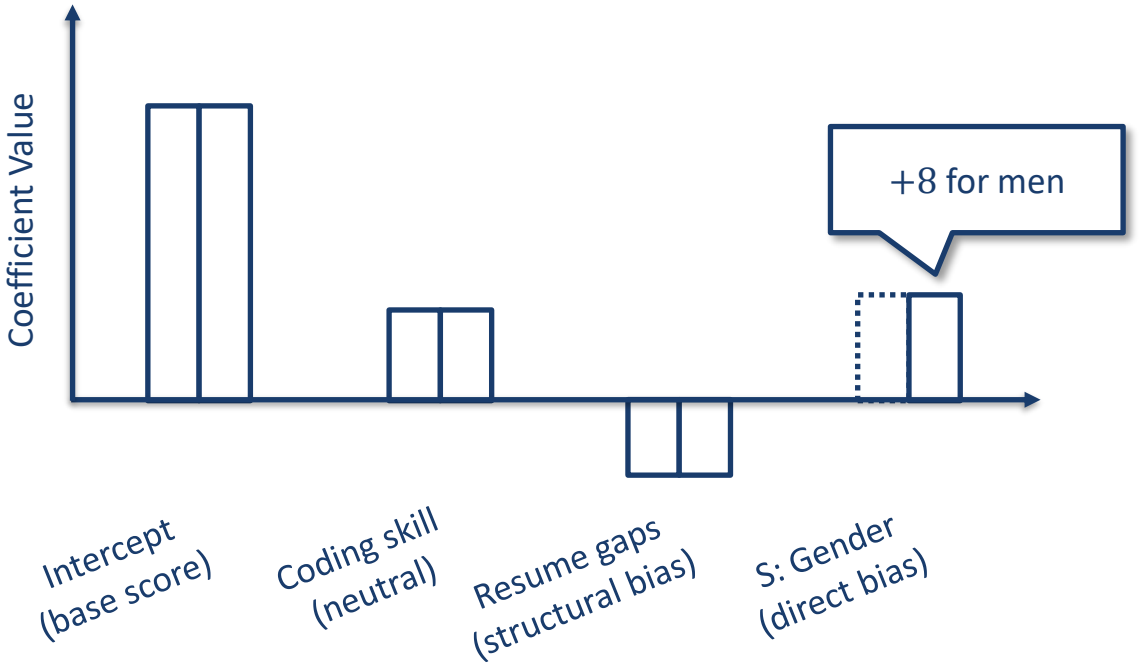
$$f^*(\boldsymbol{x}, s) = \langle \boldsymbol{x}, \boldsymbol{\beta^*} \rangle + \gamma^* s + \beta_0^*$$

Fairness constraint:

$$\mathbb{P}(f(X,S) = 1 | S = \text{🚹}) = \mathbb{P}(f(X,S) = 1 | S = \text{🚺})$$

Hiring score

Left: Women $(S = 0)$ Right: Men $(S = 1)$



Initial Bias Diagnosis. This graph reveals two major flaws in the historical data: an explicit **Direct Bonus** for men (+15) and a general **Penalty** for resume gaps (-8). This model replicates the historically biased outcomes.

The ε-optimal fair linear model

An interpretable post-processing model with feature-level bias correction

Proposition (informal)

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}} \right) + \mu_\varepsilon^{(s)}$$

$\mu^{(s)}$ and $\sigma^{(s)}$ corresponds respectively to the group-conditional mean and the group-conditional variance

$\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$) is the population-level average of the group-conditional mean (resp. variance).

$\mu_\varepsilon^{(s)}$ (resp. $\sigma_\varepsilon^{(s)}$) is the interpolation between the group-conditional mean $\mu^{(s)}$ (resp. variance $\sigma^{(s)}$) and its population-level average $\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$).

$$\beta_{0,\varepsilon}^{(s)} = \mu_\varepsilon^s - \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \langle \mu^{(s)}, \beta^* \rangle \quad \text{and} \quad \beta_\varepsilon^{(s)} = \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \beta^*$$

Direct Bias ☐

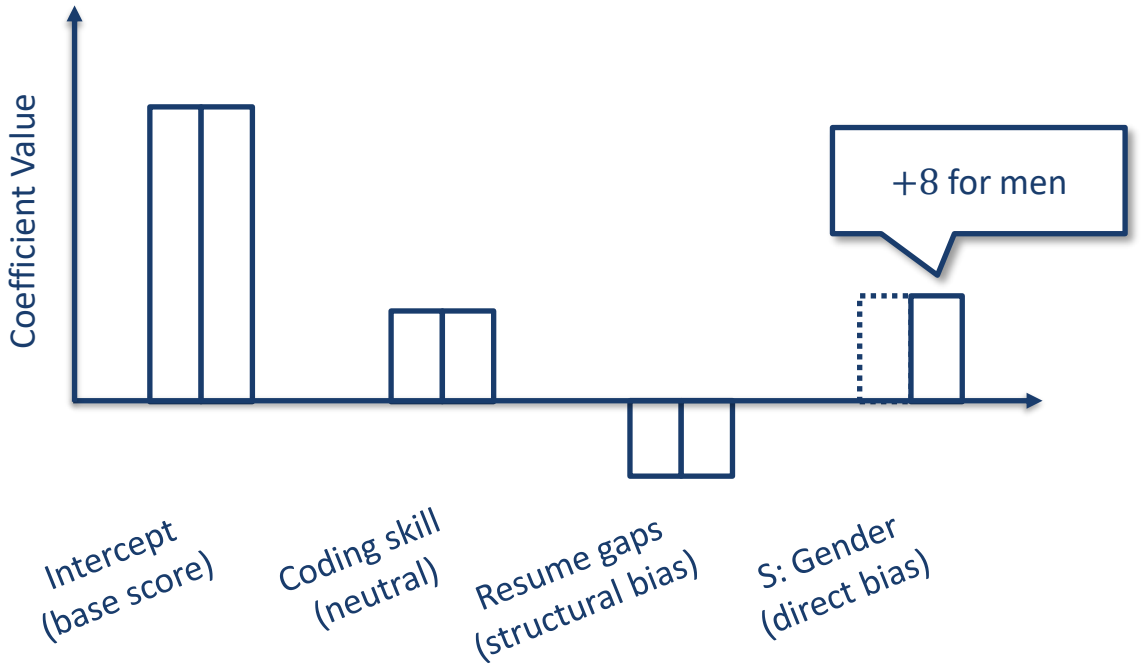
Indirect (Mean) Bias ☐

Interaction ☐

Indirect Structural Bias ☐

Hiring score

Left: Women ($S = 0$) Right: Men ($S = 1$)



Initial Bias Diagnosis. This graph reveals two major flaws in the historical data: an explicit **Direct Bonus** for men (+15) and a general **Penalty** for resume gaps (-8). This model replicates the historically biased outcomes.

The ε-optimal fair linear model

An interpretable post-processing model with feature-level bias correction

Direct Bias ☑
Indirect (Mean) Bias ☑
Interaction ☐
Indirect Structural Bias ☐

Proposition (informal)

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}} \right) + \mu_\varepsilon^{(s)}$$

$\mu^{(s)}$ and $\sigma^{(s)}$ corresponds respectively to the group-conditional mean and the group-conditional variance

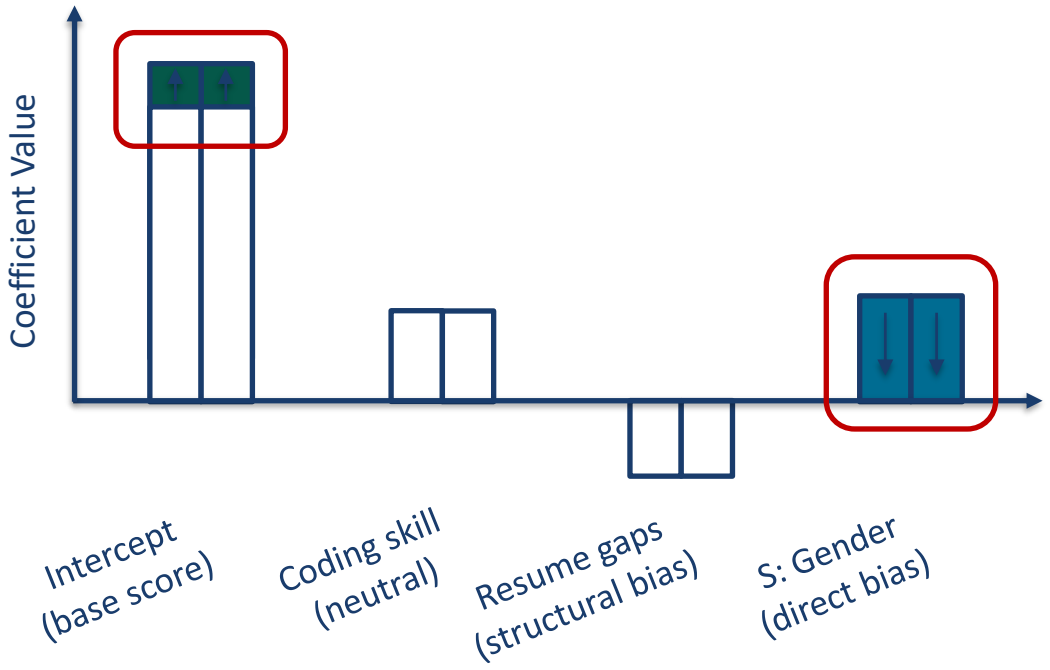
$\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$) is the population-level average of the group-conditional mean (resp. variance).

$\mu_\varepsilon^{(s)}$ (resp. $\sigma_\varepsilon^{(s)}$) is the interpolation between the group-conditional mean $\mu^{(s)}$ (resp. variance $\sigma^{(s)}$) and its population-level average $\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$).

$$\beta_{0,\varepsilon}^{(s)} = \mu_\varepsilon^s - \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \langle \mu^{(s)}, \beta^* \rangle \quad \text{and} \quad \beta_\varepsilon^{(s)} = \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \beta^*$$

Hiring score

Left: Women ($S = 0$) Right: Men ($S = 1$)



Direct Bias. We eliminate the explicit gender term. The Intercept shifts globally to the new fair population mean. At this stage, the **average score is fair**, but the distributions (variance) are still unequal due to the structural bias (unreliable "Resume gaps" penalty).

The ε-optimal fair linear model

An interpretable post-processing model with feature-level bias correction

Direct Bias ☑
Indirect (Mean) Bias ☑
Interaction ☑
Indirect Structural Bias ☑

Proposition (informal)

$$f_\varepsilon^*(\boldsymbol{x}, s) = \sigma_\varepsilon^{(s)} \left(\frac{\langle \boldsymbol{x} - \boldsymbol{\mu}^{(s)}, \boldsymbol{\beta}^* \rangle}{\sigma_{f^*}^{(s)}} \right) + \mu_\varepsilon^{(s)}$$

$\mu^{(s)}$ and $\sigma^{(s)}$ corresponds respectively to the group-conditional mean and the group-conditional variance

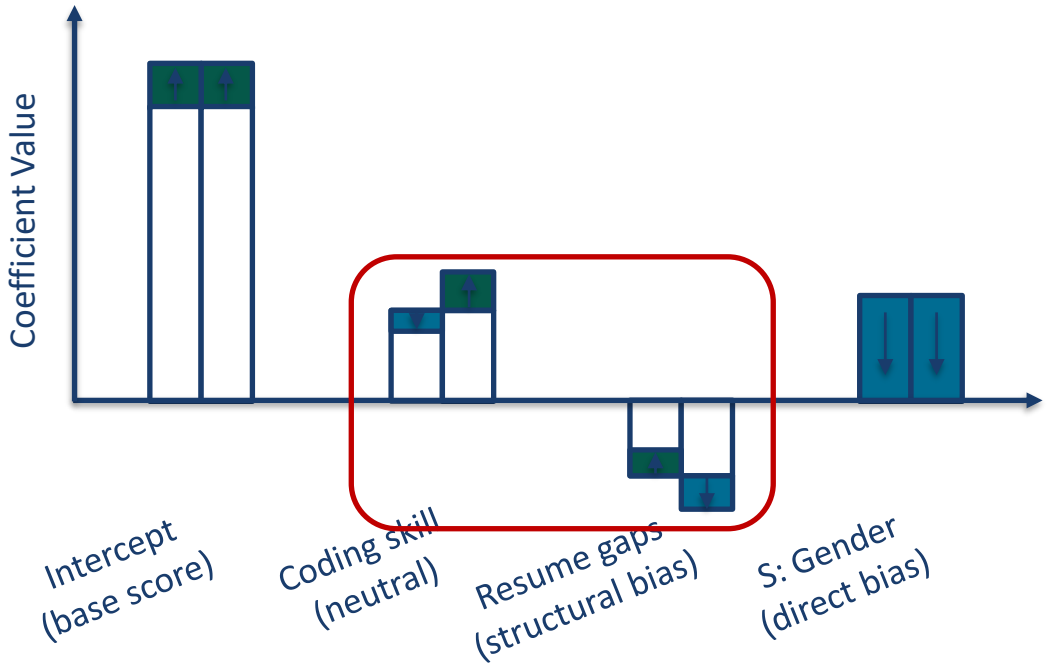
$\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$) is the population-level average of the group-conditional mean (resp. variance).

$\mu_\varepsilon^{(s)}$ (resp. $\sigma_\varepsilon^{(s)}$) is the interpolation between the group-conditional mean $\mu^{(s)}$ (resp. variance $\sigma^{(s)}$) and its population-level average $\bar{\mu}_{f^*}$ (resp. $\bar{\sigma}_{f^*}$).

$$\beta_{0,\varepsilon}^{(s)} = \mu_\varepsilon^s - \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \langle \mu^{(s)}, \beta^* \rangle \quad \text{and} \quad \boxed{\beta_\varepsilon^{(s)} = \left(\frac{\sigma_\varepsilon^{(s)}}{\sigma_{f^*}^{(s)}} \right) \beta^*}$$

Hiring score

Left: Women ($S = 0$) Right: Men ($S = 1$)



Coefficient Scaling. This is the core of our method. To achieve **Strong Demographic Parity**, the model globally rescales the coefficients. The penalty for "Resume gaps" is **dampened** for women (due to data unreliability) but is **retained** for men (where the data is reliable).

Numerical evaluation with real data

Model	CRIME			LAW			GOSSIS		
	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness	GWR^2	RMSE	Unfairness
Base Model Unaware	$.45 \pm .05$	0.15 ± 0.01	0.55 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.13 \pm .01$	$.69 \pm .01$	10.3 ± 0.1	$.14 \pm .01$
Base Model	$.46 \pm .05$	0.15 ± 0.01	0.61 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.43 \pm .02$	$.69 \pm .01$	10.3 ± 0.1	$.15 \pm .01$
CS22	$.46 \pm .05$	0.15 ± 0.01	0.54 ± 0.04	$.15 \pm .01$	$0.37 \pm .00$	$.08 \pm .01$	$.69 \pm .01$	10.3 ± 0.1	$.14 \pm .01$
FS23	$.35 \pm .09$	0.19 ± 0.01	0.20 ± 0.05	$.08 \pm .05$	$0.39 \pm .01$	$.15 \pm .05$	$.51 \pm .40$	12.5 ± 3.3	$.13 \pm .07$
Our model	$.38 \pm .07$	0.19 ± 0.01	$\mathbf{0.12 \pm 0.04}$	$.15 \pm .01$	$0.37 \pm .00$	$\mathbf{.07 \pm .02}$	$.69 \pm .01$	10.4 ± 0.1	$\mathbf{.03 \pm .01}$

Outperforms recent fairness baselines on real datasets

- **74%** of unfairness reduction on **medical** data (GOSSIS)
- **43%** of unfairness reduction on **economical** data (CRIME)
- **13%** of unfairness reduction on **admission** data (LAW)

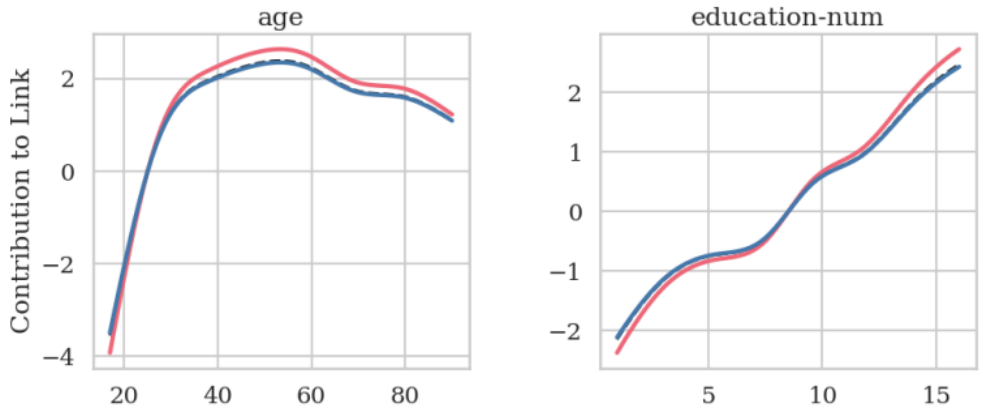
With negligible performance degradation across datasets

- ✓ **Mathematically grounded**
- ✓ Applicable to **any linear model** (low-cost post-processing)
- ✓ **Feature-level decomposition** of model bias

Extending the bias diagnostics to GLMs

First-order decomposition on the mean scale

Extension to GLMs and GAMs



Fair generalized additive model (GAM, logistic) on the *Adult* dataset to predict high income (> $50k). Visualisation of the smoothing functions.

From the linear benchmark to the GLM mean scale

Reuse the linear split, then pass it through the link with a Taylor approximation

First-order (Taylor) expansion: same four channels + two curvature

Four linear channels (preserved)

- **Direct:** the explicit effect of the sensitive variable
- **Indirect:** groups differ in their average profiles (proxies)
- **Interaction:** direct and indirect effects add up or cancel out
- **Structural:** groups have different spread in their scores

Two new curvature terms (from inverse link)

- **Curvature coupling:** the link's bend mixes the mean gap with the spread
- **Curvature amplification:** the bend enlarges the gap when group means are far apart

Latent-to-mean-scale transport in GLMs

From score to prediction: a Taylor approximation through the link

Score scale

score η = sensitive + covariates

- The group gap splits cleanly into **direct** + **indirect**
- Here the decomposition is exact

Taylor

1st-order: tangent line
+ curvature correction

Prediction scale

prediction = link(score)

- Straight part of the link keeps the same direct/indirect split
- Curve part adds the two curvature terms

Actuarial specializations of the inverse link:

Logistic
logit link · binary risk
Claim occurrence

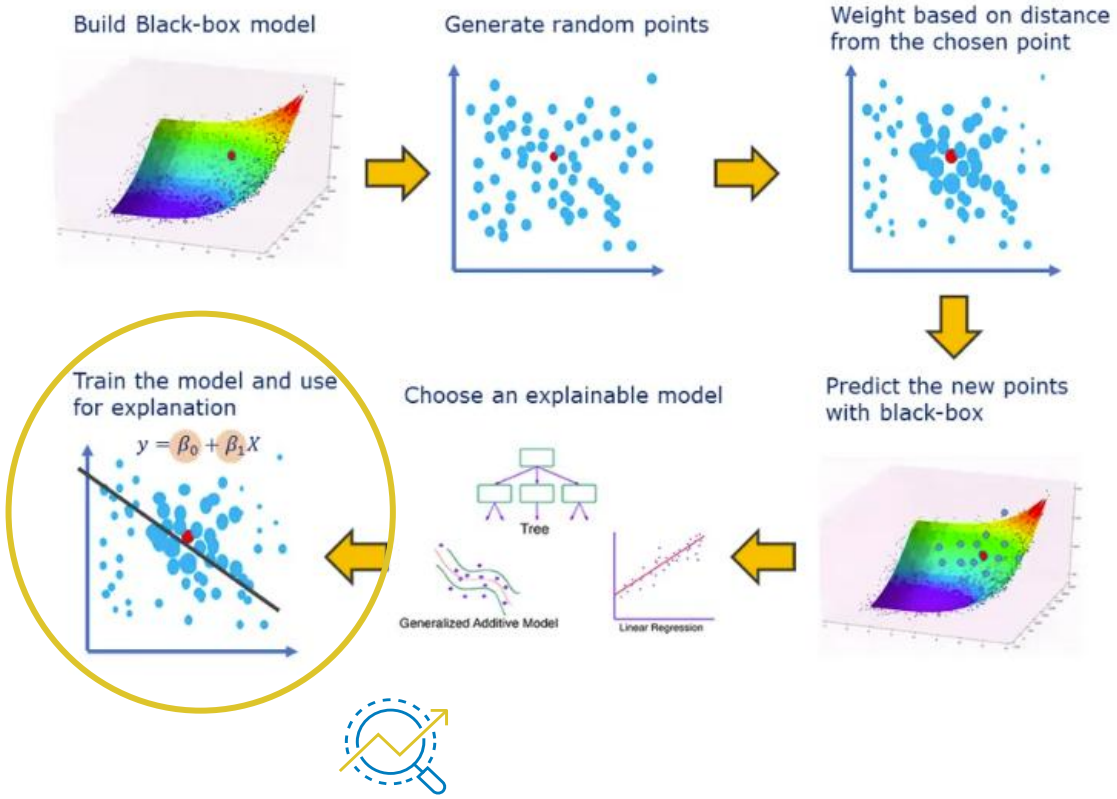
Poisson
log link · counts
Claim frequency

Tweedie
power link · aggregate loss
Pure premium

Extending the bias diagnostics to Black-box models (1/2)

Experimentation on the Medical Expenditure Panel Survey Household Component (MEPS-HC)

Extension to any « black box » model using Lime



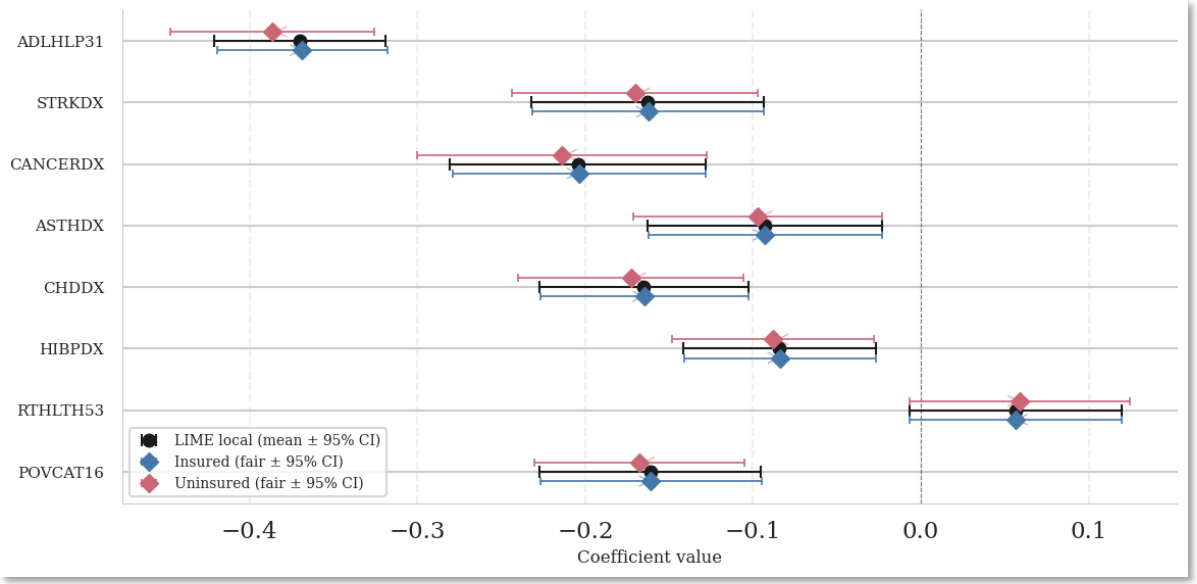
From local surrogate to bias decomposition

- **Model-agnostic:** any black-box predictor (GBM, neural net, ensemble) is handled without touching its internals.
- **Local linear surrogate:** LIME fits an interpretable score $y = \beta_0 + \beta x$ around each point —> exactly the linear predictor our framework assumes.
- **Same decomposition applies:** the surrogate coefficients feed the four-channel split —> direct, indirect, interaction, structural.
- **Local diagnostic:** reveals where, in covariate space, sensitive effects and proxies drive prediction disparities.

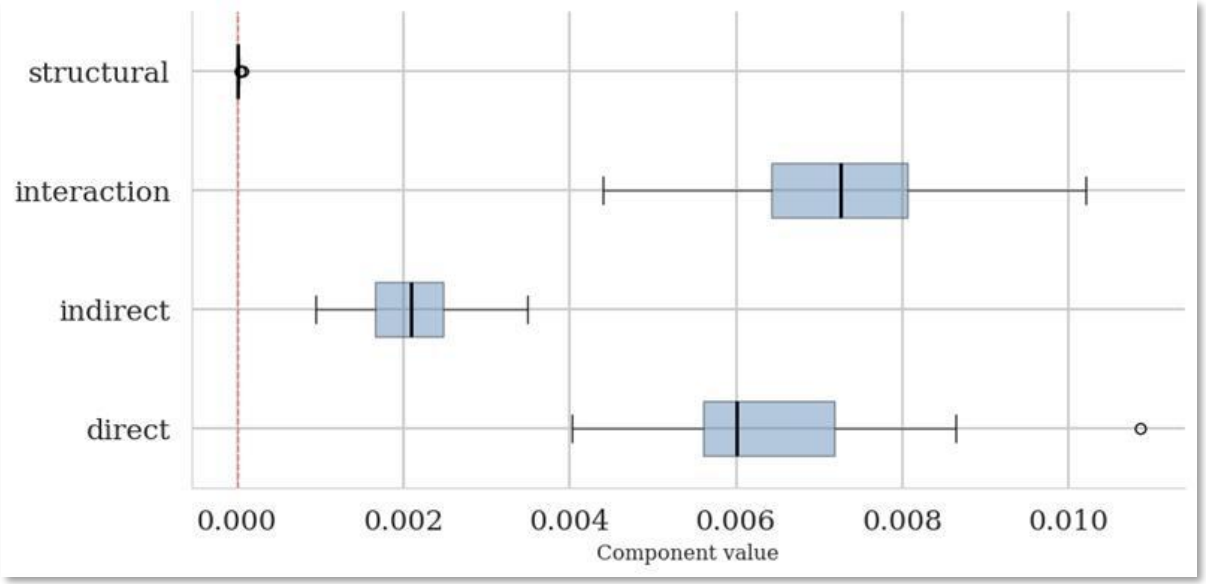
Extending the bias diagnostics to Black-box models (2/2)

Experimentation on the Medical Expenditure Panel Survey Household Component (MEPS-HC)

Use case: Prediction of the total healthcare costs (TOTEXP16) using a Lightgbm model



Coefficients Shifts



Direct and Indirect Biais Decomposition

Conclusion

A unified, interpretable diagnostic for direct and indirect discrimination

What the framework delivers:

Exact linear benchmark

The Wasserstein criterion reduces to two moments and splits disparity into four interpretable channels: direct, indirect, interaction, structural.

GLM extension

A first-order mean-scale decomposition keeps the four channels and adds two curvature terms from the inverse link : logistic, Poisson, Tweedie.

Model-agnostic reach

Through a LIME local surrogate, the same diagnostic applies to any black-box predictor, localising where sensitive and proxy effects arise.

In one sentence

We measure how unfair a model is, and break that number down to show **why** the gap is there.

It works on any model

The same method covers simple linear models, GLMs, and complex black boxes.

Good to know

It is a fast, practical check (a useful guide), not a perfect proof of fairness.

Thank you for your attention !

François HU
Milliman AI Solutions & ISFA
Lead AI engineer & Associate Professor

Bertille TIERNY
Milliman AI Solutions
AI engineer & Actuary

- Milliman is among the world's largest independent actuarial and consulting firms.
- Milliman AI Solutions is a global team of AI engineers, actuaries, and AI architects focused on delivering trustworthy and efficient AI solutions for regulated environments, with an emphasis on interpretability, auditability, robustness, and responsible use.