Fairness and Confidence in Insurance markets, a Practitioners perspective

François HU & Antoine LY

SCOR
FOUNDATION FOR SCIENCE

1) Overview – focus on group fairness

Definitions

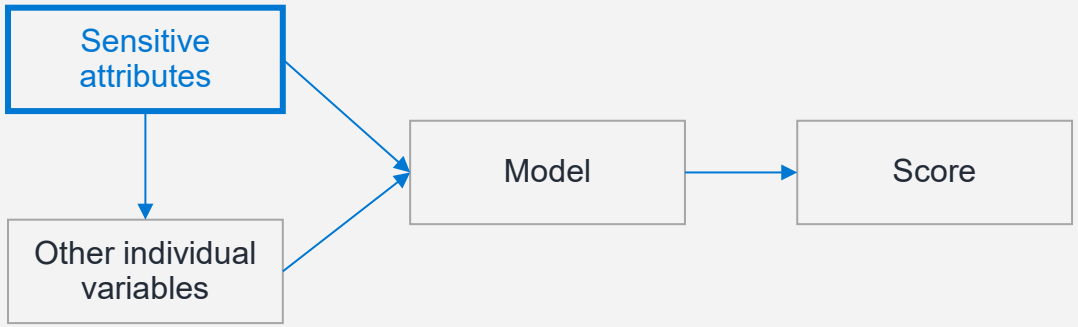
Dimensions of fairness

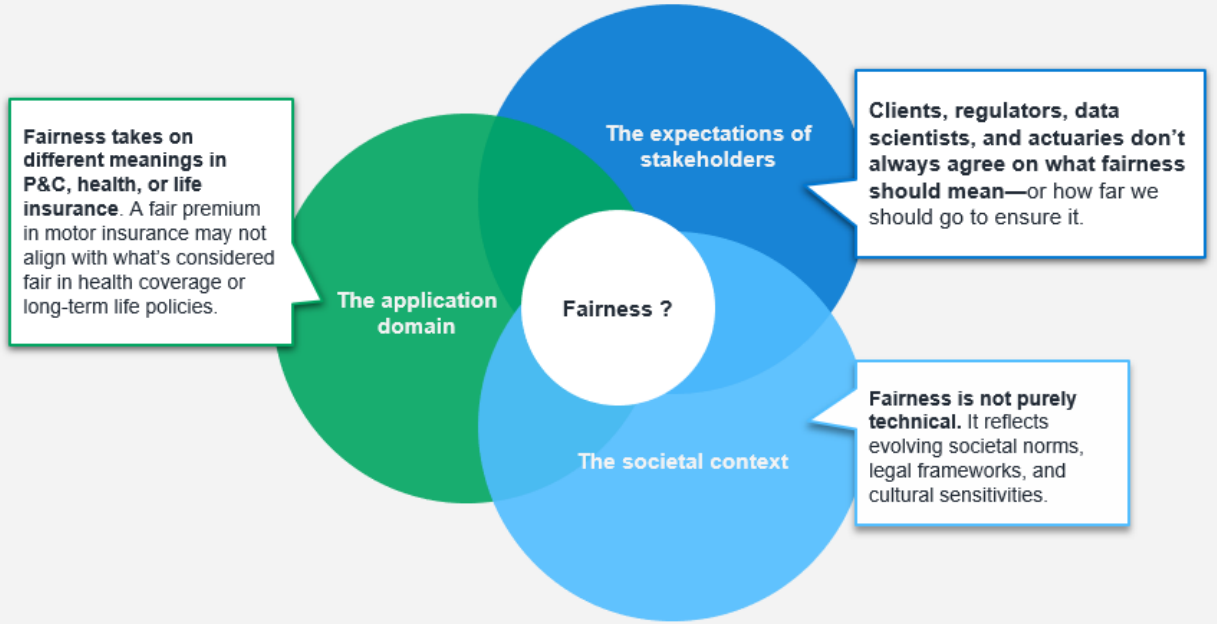
Group: Are individuals in the privileged group and those in the unprivileged group treated differently?

Individual: How would the model have made a prediction if the same individual have belonged to the privileged group?

Counterfactual: Would the decision have been the same for an individual if the value of their sensitive attribute had been different?

Discrimination within the modelling process





2) Growing attention to fairness across different sectors

Deficiencies that already pose reputational risks through concrete cases across the value chain

Scoring

A 2019 *Science* article (1), later publicized by Microsoft (2), revealed that a commercial health risk scoring algorithm used in hospitals was biased with respect to ethnicity: *"At a given risk score,* ***Black patients are significantly sicker than white patients****."* The article highlights the impact of this bias, noting that *"correcting this disparity would raise the percentage of Black patients receiving additional care from 17.7% to 46.5%."*

Fraud Detection

In France, a recent investigation (4) focused on the Family Allowance Funds (Caisses d'Allocations Familiales), which use an internal algorithm to assign fraud risk scores and allocate the work of ~700 inspectors across ~14 million beneficiary households. According to the study, **the scores rely heavily on protected attributes, including disability and family status**, and the system has not been recently audited for potential discriminatory bias.

Claims Settlement 1

In 2022, a study (4) assessed **racial discrimination** in the resolution of insurance claims related to earthquakes. It found that claims originating from **areas with higher Black populations were less likely to result in a payout**—and when a payout did occur, it was typically lower in those regions.

Claims Settlement 2

A study (5), central to an ongoing lawsuit against the insurer State Farm in the U.S., revealed disparities in how Black and white policyholders were treated. **Differences were observed in claim processing times, the volume of additional documentation requested, and the number of interactions** required with State Farm representatives before payouts were made. The case also examines automated rules generated by third-party software, which reportedly **relied on data related to individuals' place of residence.**

(1) Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
(2) https://www.microsoft.com/en-us/research/video/dissecting-racial-bias-in-an-algorithm-that-guides-health-decisions-for-millions/
(3) https://www.lemonde.fr/les-decodeurs/article/2023/12/06/algorithme-des-caf-le-departement-de-seine-saint-denis-saisit-la-defenseure-des-droits_6204180_4355770.html
(4) Lin, X., Browne, M. J., and Hofmann, A. (2022). Race discrimination in the adjudication of claims: Evidence from earthquake insurance. Journal of Risk and Insurance.
(5) https://www.insurancejournal.com/news/national/2022/12/16/699788.html

2) Mandatory Documentation (Fairness) in the U.S. and Europe

Applicable Laws	Fairness Requirements	Mandatory Documentation	Entities in charge
Insurance Laws, with specific focus on AI	**Focus on prohibited discriminations** (e.g., race, color, origin, religion, sex, orientation, disability, gender identity/expression under CO SB21-169). **Governance of ECDIS** (e.g., credit scores, social media habits, locations, purchasing habits, home ownership, educational attainment, occupation, licensures, civil judgments, and court records).	• Comprehensive documentation and regular testing for discrimination. • Testing for discrimination of models containing ECDIS (Colorado SB21-169). • Narrative report submission on compliance (Colorado SB21-169). • Risk Management Framework (Colorado SB21-169). • Detailed reporting requirements (Circular Letter of the NYDFS). • Policies, Procedures, Internal Controls (Circular Letter of the NYDFS). • Submission of AI-based algorithms and training datasets to the Pennsylvania Department of Insurance for transparency and certification to minimize bias risks (Pennsylvania, HB1663). • Compliance documentation and record keeping, metrics to monitor and review products and ensuring fair customer treatment (European Insurance & Occupational Pensions Authority).	• **Life Insurers** (Colorado SB21-169) • **All insurers in NY** (Circular Letter of the NYDFS) • Bills: automobile, motor vehicle, health insurers
AI Laws	High-Risk AIS are those involving consequential decisions in specific sectors, including the: • **financial sector** (Colorado AI Act) • **life insurance sector** (EU AI Act) • **insurance sector** (Colorado AI Act) • **health insurance sector** (Colorado/EU AI Act) • **evaluation of creditworthiness or establishment of credit scores** (EU AI Act)	**USA** • Documentation re. absence of algorithmic discrimination (Developer) • Risk Management Policy and Program on algorithmic discrimination (Deployer) • Impact Assessment on algorithmic discrimination (Deployer) **Europe** • Fundamental Right Impact Assessment outlining the specific risks of harm (EU AI Act – Deployer). • Conformity Assessment documenting the data governance and measures to detect, prevent and mitigate possible biases (EU AI Act – Developer).	• Developer of a High-Risk AI System (EU AI Act/Colorado AI Act) • Deployer of a High-Risk AI System (EU AI Act/Colorado AI Act).
Data Privacy Laws	Personal Information/data must me processed lawfully, fairly and in a transparent manner. Automated decision making, in particular in the insurance and financial sector, and the profiling requires a Data Protection Impact Assessment.	• Data Privacy Impact Assessments. • Reinforced information necessary towards the individuals, with meaningful information about the logic involved.	• Data Controller (Insurer) • Assistance by Data Processor

3) Integrating bias remediation into the modeling process

Features → Model → Score

1

Pre-processing

Consists in producing a set of corrected data (modifications to existing data / addition of synthetic data).

- **If biased data**
- **Model agnostic**
- No « direct » control of model output bias

2

In-processing

Integrating fairness constraints into the objective function used to calibrate the model

- **Remains in model class**
- Dependent on this model class
- Time-consuming (hyper-parameters...)

3

Post-processing

Adjustment of predicted scores to optimize fairness metrics under maximum performance constraints

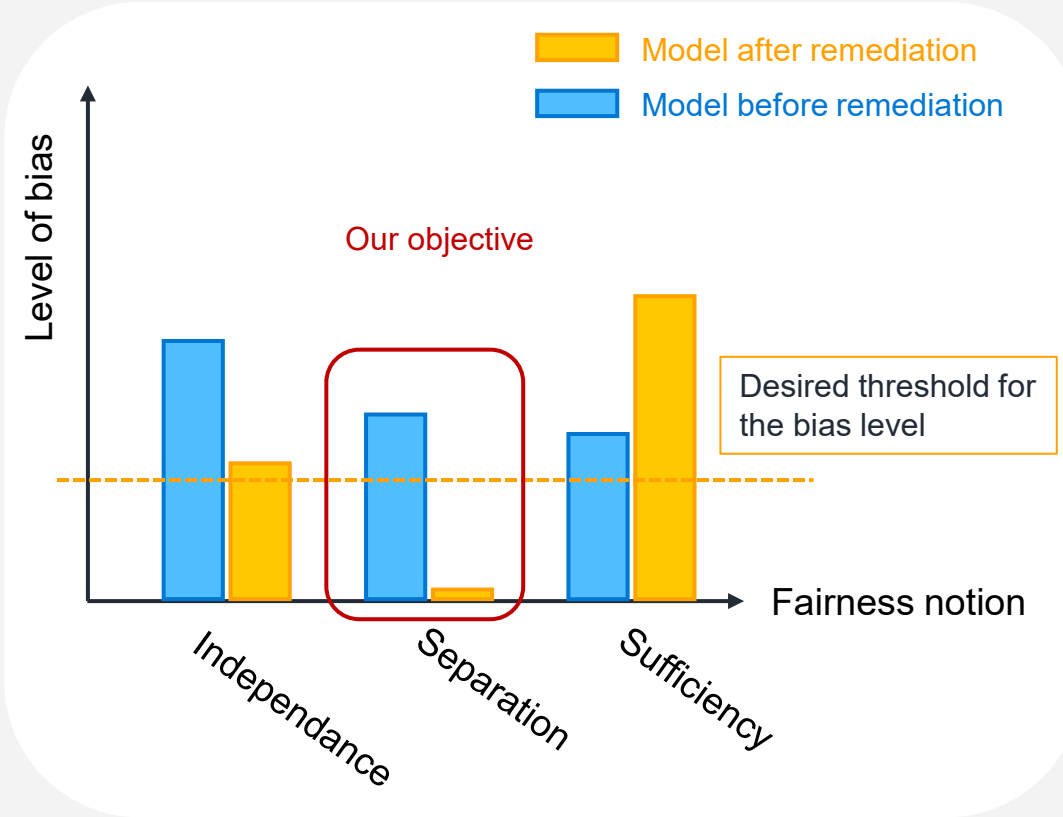
- **« Direct » bias control**
- **Model agnostic**
- Final predictions do not remain in the model class (e.g. not GLM)

4) In practice

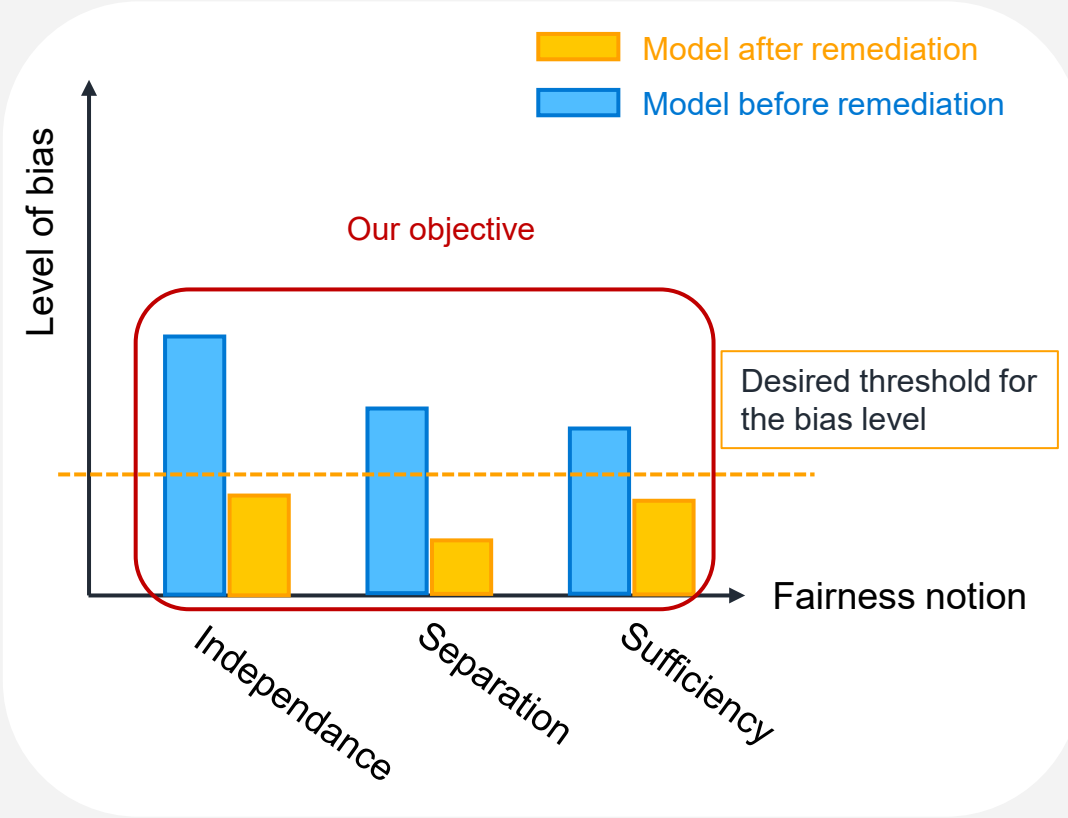
Confronted with the impossibility theorem

Illustrative example

Scenario 1 We achieved **perfect fairness** in terms of separation, but at the **expense of sufficiency**



Scenario 2 We **achieved approximate fairness** in terms of separation, while also **reducing other notions of unfairness**



4) Overview of algorithmic fairness packages

Focus on group fairness

Note : Existing Python packages focus on demographic parity and alternative metrics in binary classification. However, they do not provide a suitable framework for multi-class classification or regression for key metrics such as Equal Opportunity and Predictive Parity.

Main Python Packages \ Fairness Metrics	**Demographic Parity** $\hat{Y} \perp S$		**Equalized Odds** $(\hat{Y} \perp S) \mid Y$		**Predictive Parity** $(Y \perp S) \mid \hat{Y}$	
	Classification	**Regression**	**Classification**	**Regression**	**Classification**	**Regression**
AIF360 (IBM)	**Pre-processing approach**: Reweighing		Binary classification	✕	Binary classification: Optimal bias-perf thresholding	✕
Fairlearn (Microsoft)	Binary classification	**In-processing**: Exponentiated Gradient	Binary classification	✕ (Other metrics)	✕	
Fairness Indicators (Google)	Diagnosis only					
Equilibration/EquiPy	**Post-processing** approach: manages multiple sensitive attributes and directly proposes well-calibrated scores		✕		✕	

Maturity level : Developed | Limited | Not yet developed

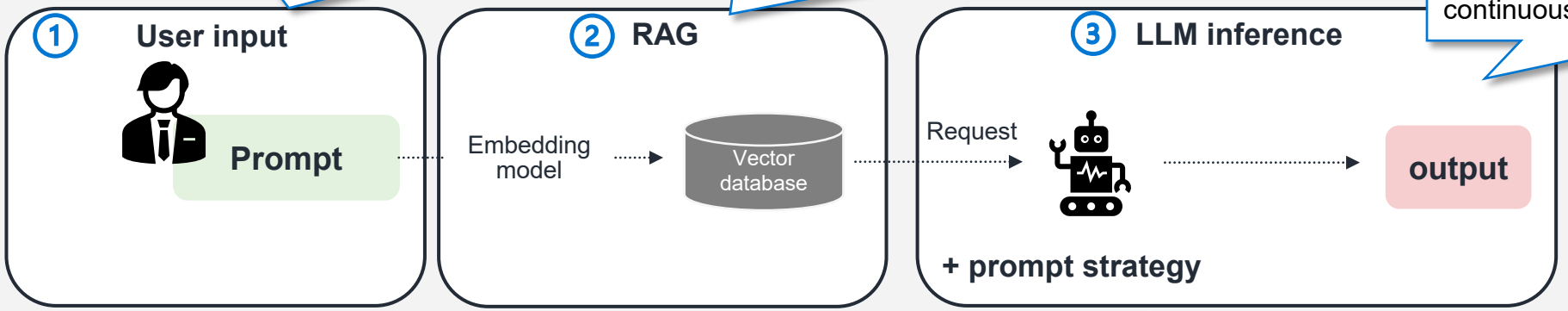
5) Generative AI ?

Content moderation, toxicity, bias and fairness

Best practices:
Design clear and specific prompts to guide the model and obtain the desired output. The *system prompt* can be used to provide ethical behavior instructions.

Best practices:
Select and maintain source documents carefully: build a reliable and domain-specific data source and ensure ongoing quality control of the documents it contains.

Best practices:
Monitoring and User Feedback: Set up monitoring to track system performance and risks. Leveraging user feedback helps identify potential weaknesses and continuously improve the tool.



Potential Bias: If the model was trained on biased data, certain prompts may trigger or amplify those biases in its responses.

Example: *"What high-paying jobs are best suited for women?"* may lead to a sexist answer.

Potential Bias: Documents used for RAG may contain historical or systemic biases. If the AI relies on such sources, it risks reproducing those biases.

Example: If the RAG corpus includes outdated bank reports from the redlining era, it may label predominantly Black neighborhoods as "high-risk."

Potential Bias: A domain-specific LLM is more likely to hallucinate, show bias, or reveal vulnerabilities when prompted outside its area of expertise.

Example: A financial advice chatbot exposed via API could be prompted to give medical advice (out-of-scope content) or respond with inappropriate or offensive suggestions—posing reputational and operational risks.


THANK YOU
SCOR
FOUNDATION FOR SCIENCE