Award Winner

Equity Underwritten: Mitigating Bias in Risk Assessment and Pricing Processes

Marco Pirra, AFFI, CAS

October 2025

The views and ideas expressed in this essay are the author's alone and do not represent the views or ideas of the Society of Actuaries, the Society of Actuaries Research Institute, Society of Actuaries members, or the author's employer.

INTRODUCTION

Fairness in insurance has always been a critical concern. Insurance inherently engages with inequalities, aiming to compensate for unpredictable financial losses while distributing risk across populations. However, some argue that insurance systems have at times contributed to social disparities through practices such as redlining or gender-based pricing.[1] . These practices highlight the need to re-examine fairness not only as a regulatory or ethical obligation but as a foundational principle for trust and access to financial services.

The evolution of insurance has introduced increasingly complex methods for evaluating and pricing risk, particularly through data-driven systems and artificial intelligence (AI). While these tools offer precision and efficiency, they also pose new challenges: statistical and algorithmic approaches introduce their own fairness frameworks, which may be disconnected from historical and legal understandings of equity. As insurers shift toward machine learning-based underwriting and pricing, it becomes essential to interrogate how *bias*—defined here as systematic unfairness in outcomes across social groups—may arise, how it can be measured, and how it might be mitigated. Similarly, this essay approaches *fairness* as a pluralistic concept, encompassing group parity, individual equity, and procedural transparency, depending on the context.

This essay focuses specifically on underwriting and pricing in life insurance, exploring how bias manifests in AI-driven systems and what solutions are emerging at the intersection of actuarial science, statistics, and computer science. It draws on conceptual distinctions between intended and unintended bias, discusses practical fairness techniques, and places these in the context of broader legal, societal, and regulatory developments. These broader regulatory and societal forces help explain why fairness in underwriting has become a central concern and provide essential context for the technical and organizational strategies discussed in the sections that follow.

---

[1] Mosley, Roosevelt, and Radost Wenman, "Methods for quantifying discriminatory effects on protected classes in insurance," *CAS research paper series on race and insurance pricing* 26 (2021), https://www.casact.org/sites/default/files/2022-03/Research-Paper_Methods-for-Quantifying-Discriminatory-Effects.pdf; and Squires, Gregory D., "Racial profiling, insurance style: Insurance redlining and the uneven development of metropolitan areas," *Journal of Urban Affairs* 25.4 (2003): 391-410, Print.

MOTIVATIONS FOR FAIRNESS: TRANSPARENCY, REGULATION, AND ACCOUNTABILITY

The push to mitigate bias in underwriting is not only an ethical initiative—it is also being driven by growing demands for transparency, legal accountability, and regulatory compliance. As insurers adopt algorithmic tools to evaluate risk and set prices, regulators have begun to intervene. Under the EU's proposed AI Act, insurance underwriting and pricing fall under the "high-risk" category, requiring explainable systems, human oversight, and formal documentation of fairness practices. U.S. state regulators and consumer protection advocates are also intensifying scrutiny of algorithmic decision-making in insurance.

These developments are transforming fairness from an ethical aspiration into a compliance requirement. Insurers are expected to demonstrate proactive assessment of discrimination risks, not just statistical accuracy. Internal governance mechanisms such as fairness review boards and bias audit systems are emerging as essential tools for accountability.

Importantly, transparency also matters to policyholders. Consumers increasingly expect to understand the factors behind their premiums and to have access to clear dispute processes. Addressing these expectations reinforces trust in insurance institutions and helps align the industry with evolving public values.

Fairness, therefore, must be seen not just as a mathematical property but as a legal, social, and reputational mandate. Its definition varies across historical, cultural, and legal contexts. As a result, efforts to ensure fairness must be interdisciplinary, involving actuarial science, data ethics, law, and public policy.

BIAS IN UNDERWRITING: INTENDED AND UNINTENDED

Bias can enter underwriting systems in multiple ways. Intended bias occurs when known disparities in data representation or feature selection are consciously accepted, typically for predictive performance. For instance, if a model is trained predominantly on male applicants and uses gender in rating, the resulting premium recommendations may disadvantage women—even if their risk is equal or lower. While technically rational, such design choices can be ethically and legally problematic, potentially constituting discrimination.

Unintended bias, by contrast, can arise from proxy variables or unrepresentative data. A model trained on urban policyholder data may fail to generalize to rural populations, leading to poor predictions and irrelevant recommendations. Similarly, socioeconomic variables like credit score or ZIP code, while predictive, often correlate with race and income, which may embed systemic inequities into pricing.

Understanding these forms of bias is key. Fairness cannot be reduced to excluding sensitive attributes from models. Indeed, the literature distinguishes between "fairness through unawareness," where protected variables are omitted, and "fairness through awareness," where these variables are included explicitly to monitor and mitigate disparities. The latter approach enables more transparent and equitable model behavior.

Risk classification is foundational to underwriting. Insurers categorize applicants based on health status, lifestyle, and other factors to assign premiums. However, classifications such as BMI thresholds or geographic location can disproportionately affect certain groups.

For example, BMI cutoffs may not account for ethnic variations in body composition. Similarly, while ZIP code is more commonly used in property insurance, it has been incorporated indirectly in some life insurance contexts—such as through credit-based scoring or third-party data enrichment—and may serve as a proxy for racial segregation or environmental inequality. These issues call for regular audits of classification systems, testing for disparate impact across demographics and allowing flexibility for individual improvements, such as wellness participation or lifestyle changes.

A pluralistic view of fairness is needed here: different conceptions of fairness (group parity, individual justice, causal attribution) may conflict, and no single metric captures equity in all contexts. This complexity should not deter action, but it requires transparency in choosing and justifying fairness criteria.

FEATURE SELECTION AND PROXY DISCRIMINATION

One major source of bias lies in feature selection. Variables like education level, employment type, or housing status may serve as stand-ins for sensitive attributes, leading to indirect discrimination. Such proxy discrimination may not violate formal model constraints but can still result in unfair outcomes.

To identify such effects, insurers can use *sensitivity analyses* and *counterfactual testing*, in which protected attributes—such as ZIP code or gender—are altered in synthetic test cases to observe if the model's outputs change significantly. While this process involves hypothetical modifications, it does not necessarily introduce personal bias if implemented in a controlled and systematic way. Rather than evaluating real individuals, these methods use matched or simulated records to assess how much sensitive attributes alone influence predictions.

Admittedly, counterfactual fairness testing relies on assumptions about which variables can be changed independently and what constitutes a "fair world." These assumptions are not free of normative judgment. However, they provide a practical lens to uncover structural dependencies in models. Additional tools like adversarial debiasing—where an auxiliary model tries to predict protected attributes from the main model's output—can further expose hidden correlations, offering a more data-driven and less subjective means of identifying bias.

The analytical tool Shapley Additive Explanations (SHAP) values and causal graphs enable users to visualize how the model functions during prediction tasks. These tools do not eliminate bias but enable stakeholders to understand and challenge model decisions. The most promising approach involves using causal inference frameworks because these methods determine whether observed relationships between variables and outcomes represent real effects or random associations.

MODELING TECHNIQUES AND TECHNICAL SOLUTIONS

Machine learning introduces powerful tools for bias mitigation. The counterfactual fairness framework tests hypothetical situations by evaluating how an applicant would be treated if their sensitive attribute were modified. The model shows unfairness when it produces different results for applicants based on their sensitive characteristics. The method of adversarial debiasing requires repeated model modifications to identify and eliminate bias in the predictions.

Explainable AI (XAI) introduces transparency as a fundamental enhancement to artificial intelligence systems. Through feature decision explanations XAI enables developers along with auditors to detect biased patterns and make necessary adjustments for bias remediation. SHAP values provide specific measurements of sensitive proxy variable impact on model predictions thus aiding model improvement as well as compliance verification.

Data preprocessing techniques provide insurers with a proactive way to decrease algorithmic bias that may surface during modeling. The process includes de-biasing, cleaning, class distribution balancing, and dataset reweighting to prevent systematic underrepresentation of any demographic group. A high-quality, diverse training dataset is essential because models developed with narrow or unbalanced data tend to replicate any embedded social inequalities. However, how can one know if a dataset meets this standard? Several diagnostics can help: distributional analysis across protected attributes (such as age, race, gender, income), missing data rates by subgroup, and coverage comparisons with population-level statistics (e.g., census or public health data). These assessments can identify whether some groups are over- or underrepresented, or if key variables are biased in how they are recorded or collected. Additionally, fairness-aware data audits can flag latent disparities in data that may

not be immediately visible. Though no dataset is perfect, transparent evaluation of representativeness is a crucial first step toward fair modeling.

When fairness constraints are directly incorporated into model training processes, they help algorithms produce balanced outcomes. The constraints function as restrictions that discourage discriminatory patterns while promoting equitable prediction distributions between subgroups. Loss functions should integrate these constraints to enable simultaneous optimization of performance and fairness.

Periodic bias audits are essential. Model performance evaluation examines different demographic groups to measure output variations, including approval rates, false positives, and pricing differentials. When model parameters reveal discrepancies through bias identification, the model parameters should be modified and feature weights adjusted to minimize inequities.

Multiple methods stacked together starting from data collection through preprocessing and modeling constraints and ending with auditing and explainability create a resilient framework for AI-driven underwriting which eliminates bias as much as possible.

CONCLUSION AND FUTURE DIRECTIONS

The potential for underwriting bias has existed since the inception of underwriting practices, but machine learning technologies have reshaped both its nature and extent. Technical solutions present possibilities, but they are not sufficient on their own. The concept of fairness needs implementation across data sources, models, organizational structures, and how society views insurance products.

Underwriting's future development requires multiple disciplines to work together. The definition of fairness criteria needs collaboration between actuarial science, statistics, computer science, law, and social science to create both technically valid and socially acceptable standards. Insurers need to work with regulators and the public through transparent engagement to embed fairness as a concrete objective within their organizational mission.

Such measures will not only ensure compliance and reduce reputational risk but also reaffirm the role of insurance as a tool for solidarity and protection, one designed to unite rather than divide.

* * * * *

REFERENCES

Charpentier, A. (2024). "Insurance, biases, discrimination and fairness." Berlin: *Springer*. https://www.springerprofessional.de/en/insurance-biases-discrimination-and-fairness/27085462

De Giovanni, Domenico, Marco Pirra, and Fabio Viviano. (2025). "Joint mortality models based on subordinated linear hypercubes." *ASTIN Bulletin: The Journal of the IAA*: 1-20. https://doi.org/10.1017/asb.2025.8

Fahrenwaldt, Matthias, et al. (2024). "Fairness: plurality, causality, and insurability." *European Actuarial Journal* 14.2: 317-328. https://link.springer.com/article/10.1007/s13385-024-00387-3

Leong, Jessica, Richard Moncher, and Kate Jordan. (2024). “A practical guide to navigating fairness in insurance pricing.” CAS RESEARCH PAPER. https://www.casact.org/sites/default/files/2024-08/A_Practical_Guide_to_Navigating_Fairness_in_Insurance_Pricing.pdf

Giudici, P., Pirra M., Zieni R. (2025). “Cyber Risk Management with time varying artificial intelligence models.” Accepted XAI-2025: The 3rd World Conference on eXplainable Artificial Intelligence https://xaiworldconference.com/2025/