Feedback on FSRA's "Proposed Automobile Insurance Rating and Underwriting Supervision Guidance"

To: FSRA Auto Insurance Regulatory Team
From: Arthur Charpentier (UQAM), Marie-Pier Côté (U. Laval), Olivier Côté (U. Laval) and Agathe Fernandes Machado (UQAM)
Date: November 15, 2024

Context

We are submitting this constructive feedback on the "Proposed Guidance: Automobile Insurance Rating and Underwriting Supervision Guidance" as part of FSRA's consultation closing on November 15, 2024. In this document, we provide a few major comments, followed by additional general comments. Section 1 presents an executive summary, and in Section 2 and 3, we discuss in more detail some of the points to clarify our critique and provide suggestions.

About the authors

The authors of this letters are Arthur Charpentier, Agathe Fernandes Machado from Université du Québec à Montréal, Marie-Pier Côté and Olivier Côté from Université Laval (Québec). They have multiple ongoing projects regarding fairness in insurance. Arthur and Marie-Pier jointly organized the first two editions of the workshop in fairness and discrimination in insurance, for which the last edition was in May 2024.

Arthur Charpentier PhD, Fellow of the French Institute of Actuaries, is full professor at Université du Québec à Montréal, Montreal, Canada, and Université de Rennes, in France. He is member of the editorial board of the *Journal of Risk and Insurance*, the *ASTIN Bulletin: The Journal of the IAA*, *Risks* and previsouly the *European Actuarial Journal*. He edited, a few years ago, Computational Actuarial Science with R (CRC, with the associate R package CASdatasets), and more recently wrote the reference book entitled Insurance, Biases, Discrimination and Fairness published by Springer. He is also on the board of the Chapman & Hall/CRC Series in Actuarial Science. His recent interests are climate change and predictive modeling insurance, more specifically in the context of fairness and discrimination. He is currently the principal investigator of a research project funded by the SCOR Fundation for Science, on "Fairness of predictive models: an application to insurance markets".

Marie-Pier Côté PhD, FSA, ACIA, is an associate professor at the School of Actuarial Science of Université Laval (Québec, Canada) since 2018 and holds the Chair of Educational Leadership in Big Data Analytics for Actuarial Science — Intact. Her research on statistical learning for actuarial science has won her the Best Paper Award in the *North American Actuarial Journal* (2021) and the Patrick Brockett & Arnold Shapiro Award from the American Risk and Insurance Association. Her interests include algorithmic fairness in actuarial science and statistical modeling of dependence and actuarial risks.

Olivier Côté is a PhD student under the supervision of Marie-Pier Côté and Arthur Charpentier. His ongoing thesis is titled "A Causal Perspective on Direct and Indirect Discrimination from Sensitive Characteristics in Insurance Predictive Models" and expected for mid 2027, in collaboration with one of the biggest insurer in Canada. This year, he joined the select group of Hickman Scholars, a prestigious program of the Society of Actuaries aimed at increasing the number of professors with professional actuarial designations.

Agathe Fernandes Machado is a PhD student under the supervision of Arthur Charpentier and Ewen Gallic (Aix Marseille Université, France). Her ongoing thesis, titled "Algorithmic Fairness and Discrimination in Predictive Models", is expected to be completed in 2027. This year, Agathe was awarded a scholarship by the International Observatory on the Societal Impacts of Artificial Intelligence and Digital Technology in Montreal.

Contents

1 Executive summary

We salute FSRA for addressing fairness in insurance regulation with a nuanced perspective, thereby taking a position as a positive leader among international regulatory entities. This is an important effort, and we congratulate the people who are making this possible at FSRA.

Here are our main suggestions of improvements on the proposed guidance, detailed in Section 2:

- **Refine the fairness principles of Fair Consumer Outcomes (FCO)**: The principles of "Absence of unfair discrimination", "Absence of unfair bias", and "Absence of proxies" are central to the fairness objectives within FCO. However, their definitions appear overlapping and lack clarity on how they differ in purpose and application. Fairness is complex, with persistent ambiguity and practical challenges. Practitioners often hesitate to act due to limited understanding of what truly matters within these concepts. We recommend refining these three principles to reduce overlap, clarify key aspects, and define each principle's distinct role.

- **Target the tests for minorities, aligned with FCO principles** : The mentioned FCO tests appear focused on population-level metrics, which may overlook the minority-centered nature of fairness. Even when the sensitive attributes are clearly defined and observed in the data, fairness requires protecting the most vulnerable subpopulations within these groups — often minorities among minorities. Practitioners need tests that target what truly matters, are easy to interpret, directly applicable, and clearly aligned with the broader fairness principles of FCO.

- **Address the sensitive variable challenges**: Accessibility to sensitive variables is unaddressed, creating challenges for effective fairness assessments. Insurers need guidance on whether to collect or estimate the protected characteristics of their insureds, or to rely on third parties. Even with access to the sensitive information, an insurance company's data often misaligns with broader population distributions, complicating the application of uniform fairness standards. Clear guidance on handling sensitive data is essential to ensure meaningful, consistent fairness across the industry.

And here are other general comments, detailed in Section 3:

- **Emphasis on credit scores**: Why prioritize credit scores over other factors? (Sec. 1.3.1)

- **Transparency:** Transparency in Sec. 1.4 does not mean model transparency. Model transparency refers to the ease with which a model's inner workings can be understood by humans, as in Lipton (2018). Black-box models and interpretation tools fall short. For example, the partial dependence plots, as discussed in Xin et al. (2024), can lead to misguiding insights into the model. The same applies to popular explainable tools such as SHAP and LIME (Slack et al., 2020). We recommend using models that are directly interpretable.

- **Fair consumer outcomes conflict:** "The Fair Consumer Outcomes are not listed in order of priority and may come into conflict in practice." (Sec. 1.4) Unfortunately, "may" is not the correct word in that sentence, it should be "will" unless there is no link whatsoever between the protected grounds and the risk; we discuss compromise between actuarial fairness and solidarity in Côté et al. (2024b).

- **Proxies and bias:** Causal inference should be part of the discussion in order to allow rating factors associated with protected grounds while avoiding undesired proxy effects. Potential for proxy effect is everywhere, so causal tools come handy in controlling for those biases. A variable can suffer from a "proxy effect" even when it is a risk factor. Causal inference should enter the discussion as it can help to isolate the "true risk component" of rating variables.

- **Reporting requirements:** The distinction between self-reporting and audit requirements in fairness assessments remains unclear. Self-reporting offers flexibility but lacks objectivity and consistency, while audits provide uniformity but are heavier and restrict insurers' methodological choices. Greater clarity on which approach is required would strengthen the assessment process.

- **Adverse Selection Risk:** If an insurer prioritizes solidarity, adverse selection may occur if competitors do not adopt similar measures. Without regulatory mandates, expecting insurers to pursue solidarity is unrealistic, as it would weaken their competitiveness. The balance between actuarial fairness and solidarity must be explicitly defined.

- **Unintended Consequences Risk:** Clearer guidance is needed to identify and address potential unintended risks from fairness measures. For example, if a vulnerable group faces higher premiums and loss ratios, reducing premiums could increase the loss ratio gap, which may lead insurers to incur losses and could affect coverage availability for this group. These risks would benefit from careful consideration and proposed solutions.

- **Defining the target population for fairness:** Each insurer's portfolio reflects a skewed sample of the entiere population of policyholders shaped by specific demographics and business factors, making data sparsity and the gap with the insured population unique to each insurer's view. As shown in Côté et al. (2024a), achieving demographic parity in a specific portfolio often conflicts with market-wide parity, creating a trade-off that demands prioritization. What should be the scale of the target population for which fairness is intended?

2 Details on major comments

This section gives details on all the major suggestions summarized in Section 1.

2.1 Refine the fairness principles of FCO

We found that the terms used across the proposed guidance to discuss fairness often lack precise definitions, making it challenging to navigate the document effectively. Terms such as “Unfair discrimination” (Sec. 1.4, 3), “directly or indirectly proxies” (Sec. 1.4), “Unfair Bias” (Sec. 3), and “disparate impact” (Sec. 3), “non-risk factors” (Sec. 3) are central but remain unclear without mathematical or legal definitions. For example, the term “unfair treatment” is explained as “a non-exhaustive list of specific conduct that constitutes **unfair treatment**, which includes: [...] engaging in **unfair discrimination**”, which seems a circular definition. Narrowing down the language to a few well-defined, central concepts could significantly improve clarity, reducing confusion and ensuring consistency in interpretation.

The Fair Consumer Outcomes (FCO) guidance introduces three core principles to support fairness: “Absence of unfair discrimination”, “Absence of unfair bias”, and “Absence of proxies”. While these principles are foundational to fairness within FCO, their titles and definitions seem to overlap, creating blurred boundaries that could lead to inconsistent application. For example, “absence of unfair discrimination” requires “demonstrating how the use of these data elements [...] relates to the risks [...]”, whereas “absence of unfair bias” involves the evaluation of “deviations from actuarially indicated rates”. Furthermore, the principle of “accurate pricing and underwriting” ask that the “Pricing actuaries explain any deviations [...] from actuarial indications”. The distinction between these principles becomes difficult to make, as they all mention staying close to the risk, which means that they are aligned with actuarial fairness (Barry, 2020). Since fairness is known to be conflicting (See, e.g., Sec. 8.5 in Charpentier, 2024), the apparent alignment among the proposed principles suggests a framework that may not fully capture its complexities.

Despite this, the FCO in Chapter 3 is structured around three core principles. We interpret FSRA’s intention as follows:

- **Absence of Unfair Discrimination**: Insurers must ensure that models comply with regulations and are actuarially justified.
- **Absence of Unfair Bias**: Insurers must mitigate bias and ensure fair premiums for all customers.
- **Absence of Proxies**: Insurers must ensure they do not misuse any variables in a way that indirectly targets protected groups, whether intentionally or not.

While the description of the “absence of unfair bias” principle seems aligned with disparate impact mitigation, its description is too broad. It should focus on the concept of solidarity, a crucial element in fairness assessments. Currently, the FCO framework emphasizes minimizing loss ratio differences between protected groups to improve model accuracy, and make sure that the insurer does not make more profit on some protected groups. Yet, this approach focuses solely on “disparate impact” from a loss-ratio perspective, bringing it in line with actuarial fairness and accuracy. This perspective **encourages the use of proxy factors** that could reduce demographic disparities, placing it as justifiable on the basis of “absence of unfair bias”. Solidarity as a fairness goal requires to balance accuracy with demographic parity in terms of premiums. Demographic parity seeks equal premium distribution across protected groups, ensuring they receive similar rates. It reflects solidarity by promoting levelled premiums across these groups. A focus on solidarity for one FCO principle would ensure that the framework accounts for both actuarial fairness and the (distinct) principle of solidarity.

To improve clarity and distinctiveness, we recommend reorganizing FCO’s fairness principles as follows:

- **Absence of Unjustified Discrimination**: This principle ensures that models avoid discrimination that cannot be justified by actuarial standards or regulations.
- **Absence of Demographic Disparities**: This principle emphasizes a solidarity-based focus, aiming to balance premiums across protected groups to prevent undue demographic disparities.

- **Absence of Proxies**: This principle prevents the use of variables that indirectly target protected groups, in order to reduce bias from proxy indicators.

This reorganization enhances each principle's focus on a distinct and critical dimension of fairness: actuarial fairness, solidarity, and proxy avoidance. For instance, renaming "Absence of Unfair Discrimination" as "Absence of Unjustified Discrimination" provides clearer alignment with actuarial standards. Similarly, "Absence of Demographic Disparities" better reflects the goal of promoting solidarity by reducing demographic imbalances. Moving the element "deviations from actuarially indicated rates" under "Absence of Unjustified Discrimination" also aligns it more closely with actuarial fairness, while solidarity would remain a separate goal that sometimes warrants deviations from strictly actuarial rates.

With this proposal, we highlight more concretely the balance between actuarial fairness and solidarity. However, insurers favoring actuarial fairness may gain a competitive edge due to adverse selection. Without standardized industry practices, this balance is likely to shift toward actuarial fairness. Achieving solidarity requires clear regulatory guidelines or collective risk-sharing mechanisms. As coined in Côté et al. (2024b), there is a "price to pay for fairness" (Chzhen and Schreuder (2022) use the term "risk-fairness trade-off", stressing that asking for fairness will reduce accuracy).

Applying this concept allows to specify the emphasis on solidarity within this balance, particularly the relative reduction in unfairness achieved by using a model targeting FCO, compared to the model optimized for predictive performance. Requiring a minimum relative improvement as part of regulatory guidelines could foster more standardized practices.

Finally, drawing a clear link between inputs and outcomes can enhance understanding of how each fairness principle interconnects. For instance, demographic disparities in outcomes (unequal premium distributions across groups) may result from input-based proxies (variables indirectly targeting protected characteristics). Identifying these relationships allows practitioners to see how input choices affect fairness outcomes, guiding them in selecting variables that uphold fairness. This approach provides clearer guidelines, reduces interpretive ambiguity, and enables a more consistent application of fairness principles across the industry.

2.2 Targeted Tests for Minorities, Aligned with FCO Principles

Assuming that FCO principles capture the core facets of fairness, it is now critical to ensure that the set of tests fully covers each principle. These tests must also reflect the minority-centered nature of fairness, ensuring that fairness efforts yield results for the vulnerable groups they aim to protect.

Each test should explicitly target a specific FCO principle, with clarity about which one it assesses. Together, the tests must cover all principles, offering a complete view of fairness. For instance, some tests may check for deviations from actuarial standards, while others focus on demographic disparities to support solidarity. This will naturally make the tests inherently conflicting, but such conflict demonstrates that we are addressing fairness comprehensively. This alignment ensures fairness assessments are relevant and actionable.

The purpose and design of mentioned FCO tests, such as control variable, balancing, and loss ratio tests, remain vague. Discussions with our industrial collaborators reveal that further detail is needed, particularly as these tests seem to rely primarily on aggregate, population-wide metrics.

While aggregate metrics provide a broad view, they are incomplete in fully addressing fairness for vulnerable subpopulations within protected groups. In Ontario, credit score often features in fairness debates around insurance underwriting and pricing. However, in focusing solely on credit scores risks, we might overlook the socio-economic realities that fairness policies aim to address; credit-score-based fairness policies are often designed to protect groups facing systemic barriers, not individuals with poor credit scores due to consistently high-risk behaviours. For example, systemic factors may lead certain ethnic groups to have lower average credit scores, resulting in higher premiums. To ensure these fairness policies have their intended results, it is essential to assess their effects specifically on the vulnerable subpopulations they are designed to support.

Population-wide metrics can obscure fairness by overlooking outcomes experienced by individuals behind the data points. Aggregates may hide disparities within subpopulations, where unfairness can vary widely. To grasp the full picture, we need more than population-wide metrics; we must measure how unfairness is distributed across various segments. Distributional methods can then identify subpopulations for closer scrutiny, ensuring fairness reaches those who need it most. By incorporating distributional or subgroup-based metrics, FCO tests would ensure that fairness is monitored across all layers of the population. This approach goes beyond averages to capture a fuller picture of fairness, covering the most vulnerable, not just the majority in a protected class. Such metrics capture the nuances needed to reflect the minority-centred nature of fairness. As a tool to ensure the consideration of minorities within minority groups, intersectional fairness, defined as crossing multiple sensitive groups, allows for the assessment of unfairness by accounting for the relationships between multiple sensitive variables. If correction methods are applied, it ensures that sensitive variables are not corrected independently, preventing the creation or amplification of unfair treatment for one group at the expense of another.

Following Chapter 8 of Charpentier (2024), the **solidarity** aspect of fairness can be represented by ensuring equal average premiums across protected groups (Def. 8.5, *weak demographic parity*). For a more rigorous approach, we could require the same premium distribution across protected groups (Def. 8.6, *strong demographic parity*), achieved by minimizing the *Wasserstein distance* (Prop. 8.1) between distributions of protected populations. Monitoring the extremes of distributions is also useful; for example, we might track the mean premium for the riskiest 5% within each protected group. If some allowed variables (e.g., vehicle type and usage) are identified for permissible full risk variation, even if they introduce demographic disparities, we could require similar premium distributions across protected subpopulations within each segment of this "allowed subset".

In contrast, the **actuarial fairness** component can be assessed by calculating the difference between commercial rates and actuarially indicated rates (considering all allowed and protected attributes). This difference can then be evaluated across groups, subgroups, and key metrics (e.g., the mean for the top 5% deviations within protected groups). Similarly, loss ratios can provide further insight: by calculating loss ratios for 250 random subpopulations and comparing distributions across protected groups, we can perform comparable fairness tests (e.g., checking for similar average loss ratios across groups, or consistent loss ratio quantiles). Following Baumann and Loi (2023), actuarial fairness aligns with the *sufficiency* criterion (Def. 6.22) which ensures similar expected claims across protected subpopulations within each predicted premium group, accounting for risk-relevant factors.

About the **absence-of-proxy** principle, no variable is a proxy in itself, but its use in a model may create proxy effects. Comparing commercial rates to the price based purely on the causal effect of ratemaking variables can reveal such effects. If certain segments show skewed prices seemingly targeting protected groups, this flags a need for monitoring to prevent extra premiums tied to protected status.

Practitioners grasp the basics of fairness and the impact of over- or under-charging individuals, but applying fairness principles to populations remains challenging. We need simple and principle-driven assessment tools. This invites practitioners into fairness discussions, helping them see its importance rather than just complying.

2.3 Address sensitive variable accessibility

Fairness assessments hinge on the use of sensitive variables, yet there is a significant challenge in accessing and managing this data accurately. Insurers must decide whether to collect or predict sensitive data themselves or rely on third-party sources. Without this data, insurers lack a foundational tool to identify and confirm the absence of unfair discrimination, risking unverified and unmonitored fairness outcomes. When predicting sensitive data, the reliability of predictions for minorities must be carefully checked due to the data scarcity for minority groups, to avoid unfairness in the methodology itself (Imai et al., 2022) or miscalibrated predictions (Fernandes Machado et al., 2024).

There are also multiple potential "protected grounds", which mandates clarification. The Ontario Human Rights Code lists protected grounds such as citizenship, race, place of origin, ethnic origin, color, ancestry, disability, age, creed, sex/pregnancy, family status, marital status, sexual orientation, gender identity, gender expression, public assistance (for housing), and record of offenses (in employment). However, most insurance datasets only cover minority status via aggregated census data, leaving other protected grounds unaccounted for and thus limiting possible fairness assessments. This gap emphasizes the need for a clearly defined set of protected groups within industry standards.

Even when sensitive attributes are observed, the scale at which fairness is assessed presents another obstacle. Fairness practices typically apply across the market-wide insured population, yet many fairness techniques depend on access to the full distribution of sensitive attributes — something that individual insurers cannot achieve. An insurer's own distribution will differ from a market-wide distribution due to multiple factors described in the general comment "Defining the target population for fairness".

Finally, to implement effective fairness practices, insurers need not only accurate sensitive data but also population-level benchmarks that align with the scope of fairness desired. Without a clear, collective standard for sensitive attributes, fairness interventions may inadvertently overlook vulnerable subpopulations, especially minorities within protected groups. We recommend that guidance explicitly address these issues, setting standards for sensitive variable access and ensuring that insurers can anchor their fairness assessments in relevant, industry-wide data.

3 Details on general comments

This section gives details on some of the general comments summarized in Section 1.

3.1 Transparency and Model Interpretability

Sec. 1.4 mentions transparency but does not address model transparency: the ability to understand how a model makes predictions. In fairness assessments, true transparency requires more than outcome insights; it demands clarity on how models work, especially in fields like insurance.

"Black-box" models, such as complex machine learning algorithms, often lack this clarity. Interpretation tools like SHAP and LIME attempt to explain these models but can be misleading, as shown by Slack et al. (2020). Similarly, partial dependence plots (PDPs), discussed by Xin et al. (2024), may obscure feature interactions, creating a false sense of understanding.

To achieve real transparency, we recommend defining transparency to include model interpretability, ensuring models are understandable by design. This focus reduces reliance on post-hoc tools.

When assessing FCO principles, transparency must be considered to "effectively monitor and respond to inquiries and complaints" from consumers. In this context, communicating population-wide metrics to consumers may be replaced by considerations of individual fairness, with causal inference supporting the calculation of individual-level metrics, as highlighted in Kusner et al. (2017).

3.2 Fair consumer outcomes conflict

The Fair Consumer Outcomes are described as "not listed in order of priority and may come into conflict in practice" (Sec. 1.4). However, "may" is misleading; "will" is more accurate, as well-defined fairness principles inevitably conflict. Fairness in actuarial science is multi-dimensional, with inherent tensions between objectives like "actuarial fairness" and "solidarity" (Lindholm et al., 2024b). When fairness is framed solely by metrics such as "deviation from actuarial indicates", "deviation from risk estimates", and "loss ratio parity", it lacks true trade-offs, as these pillars all align with actuarial fairness. Without real tension among principles, the framework risks overlooking essential dimensions of fairness, like solidarity.

A comprehensive fairness framework should reveal and manage competing priorities, as each principle contributes uniquely to broader goals of equity and fairness. The absence of conflict suggests a critical element may be missing, reducing the framework's ability to address fairness comprehensively. In Côté et al. (2024b), we discuss how recognizing and balancing these trade-offs enables insurers to approach fairness as a complex balance between inherently conflicting values rather than a singular, conflict-free goal.

3.3 Proxies and bias

Potential for proxy effect is everywhere. In the era of big data, many combinations of covariates will inevitably correlate with protected attributes. However, correlation alone does not make a variable a "proxy". As Tschantz (2022) puts it, "it may be difficult to avoid using features that are not at least weakly proxies", highlighting that a **variable's role as a proxy depends not on the variable itself but on how it is used** in relation to others and protected attributes.

Causal inference is essential for managing proxy effects, particularly when rating factors are tied to protected characteristics. Imagine building a predictive model where two correlated variables both appear to impact the outcome, yet something feels off. The model assigns too much importance to one variable and downplays the other. Intuition suggests that part of the downplayed variable's effect is mistakenly absorbed by the other. This points to a deeper question; how the influence of each explanatory variable on the response variable Y should be properly allocated.

As a further complexity, the direction of the perceived proxy relationship can shift based on the perspective taken. As Tschantz (2022) explains, depending on whether the auditor examines X_1 as a proxy for X_2 or vice versa, the assessment might suggest that either variable is the proxy. Domain knowledge is key in deciding which is the case and actuaries working in industry have a valuable expertise: they could take better advantage of it through causal assumptions. Causal inference provides essential tools for navigating these ambiguities, helping to isolate the true risk effects from proxy effects.

Many rating variables, such as territory or income, are valid predictors of risk, but can unintentionally act as proxies for characteristics like ethnicity or socioeconomic status, depending on their use. A variable doesn't "become" a proxy; rather, it is the use of the variable in the model that qualifies as a proxy, not the variable itself. For instance, territory may accurately capture geographical risk, but if misused, it can effectively target specific ethnic communities concentrated in certain areas, leading to unfair premiums as the variable inadvertently targets ethnicity.

Without clarity on proxy, we risk misallocating effects between variables, leading to incorrect assessment of the relationships between variables: bias. To address biases such as the proxy effects, we need more than correlation, we need to think causally. Causal inference methods help clarify how variables genuinely relate to each other. For example, if territory is strongly tied to claims but also aligns closely with ethnic or economic groupings, causal tools can help separate the "true risk" component from social factors, so the variable's effect in the model reflects only the territory's risk.

Causal inference goes beyond correlations to show how variables relate in the real world — the world where fairness efforts must ultimately deliver results. As almost any variable can unintentionally serve as a proxy, causality is essential.

3.4 Reporting Requirements

It remains unclear whether fairness assessments require self-reporting, audits, or both. Each approach has strengths but also limitations. Self-reporting allows insurers flexibility to assess and address internal biases in a way that fits their unique processes. However, this flexibility can result in less objective, inconsistent evaluations across the industry. Audits, by contrast, provide an unbiased, uniform assessment that better aligns with regulatory standards. Yet audits are heavier, reduce insurers' flexibility in choosing assessment methods, and may come with higher costs for insurers—costs likely to be passed on to consumers, potentially undermining consumer welfare.

If audits are required, it is essential to clarify who would perform them—whether FSRA directly or a designated third party. FSRA-led audits would ensure alignment with regulatory expectations, but third-party audits could provide added independence and potentially lessen FSRA's administrative burden. Clarity on these points would help insurers better prepare for compliance, while ensuring that fairness assessments remain both rigorous and feasible.

3.5 Defining the target population for fairness

Vulnerable individuals are at the centre of fairness concerns. How they are treated in the entire insurance market matters. However, if each insurance company make specific adjustments on its own portfolio, it does not guarantee a fair treatment of individuals overall. As detailed in Côté et al. (2024a), "[m]odels are trained on each insurer's portfolio, potentially biased subsets of the full insured population. Each insurer's portfolio reflects a skewed sample shaped by specific demographics and business factors," meaning that the fairness criteria do not necessarily behave the same way for an insurer's portfolio than for the complete population of policyholders. As explained in Côté et al. (2024a), "achieving demographic parity in a specific portfolio often conflicts with market-wide parity, creating a trade-off that demands prioritization. What should be the scale of the target population : a given insurer's clients, a provincial insurance pool for a specific coverage, a product line in a country? Fairness comes at a cost for consumer welfare (Shimao and Huang, 2022) and is already complex due to inherent conflicts (Lindholm et al., 2024a; Kleinberg et al., 2016). To avoid unnecessary costs and added complexity, regulators should guide insurers by defining a target population on which fairness is intended, aligning efforts toward a single, people-centred fairness standard."

References

Barry, L. (2020). Insurance, big data and changing conceptions of fairness. *European Journal of Sociology*, 61(2):159–184.

Baumann, J. and Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, 36(3):45.

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

Chzhen, E. and Schreuder, N. (2022). A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50.

Côté, M.-P., Côté, O., and Charpentier, A. (2024a). Selection bias in insurance: why portfolio-specific fairness fails to extend market-wide. *Available at SSRN 5018749*.

Côté, O., Côté, M.-P., and Charpentier, A. (2024b). A Fair price to pay: exploiting causal graphs for fairness in insurance. *Available at SSRN 4709243*. https://dx.doi.org/10.2139/ssrn.4709243.

Fernandes Machado, A., Charpentier, A., and Gallic, E. (2024). Post-calibration techniques: Balancing calibration and score distribution alignment. *Advances in Neural Information Processing Systems*.

Imai, K., Olivella, S., and Rosenman, E. T. (2022). Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *Science Advances*, 8(49):eadc9824.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint*. https://arxiv.org/abs/1609.05807.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30.

Lindholm, M., Richman, R., Tsanakas, A., and Wüthrich, M. V. (2024a). Sensitivity-based measures of discrimination in insurance pricing. *Available at SSRN*. https://ssrn.com/abstract=4897265.

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

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Shimao, H. and Huang, F. (2022). Welfare implications of fairness and accountability for insurance pricing. *Available at SSRN 4225159*. https://ssrn.com/abstract=4225159.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA.

Tschantz, M. C. (2022). What is proxy discrimination? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1993-2003.

Xin, X., Hooker, G., and Huang, F. (2024). Why you should not trust interpretations in machine learning: Adversarial attacks on partial dependence plots. arXiv preprint. https://arxiv.org/abs/2404.18702.