European Actuarial Journal
https://doi.org/10.1007/s13385-024-00387-3

SURVEY PAPER

Fairness: plurality, causality, and insurability

Matthias Fahrenwaldt[1,2] · Christian Furrer[1] · Munir Eberhardt Hiabu[1] · Fei Huang[3] · Frederik Hytting Jørgensen[1] · Mathias Lindholm[4] · Joshua Loftus[5] · Mogens Steffensen[1] · Andreas Tsanakas[6]

Received: 23 April 2024 / Revised: 21 May 2024 / Accepted: 24 May 2024

Abstract
This article summarizes the main topics, findings, and avenues for future work from the workshop *Fairness with a view towards insurance* held August 2023 in Copenhagen, Denmark.

Keywords Artificial intelligence · Discrimination · Insurance · Machine learning

1 Introduction

Fairness in insurance has always been a critical concern as insurance addresses inequalities, and, therefore, the nature and origin of such inequalities are key ingredients in designing insurance systems and contracts. Historically, discriminatory practices in insurance led to social and economic disparities. This includes not just practices such as gender-based pricing but also redlining [35]. Ensuring fairness in insurance is essential for social justice and equal access to crucial financial services. Biases must be addressed to maintain trust in the insurance industry and to contribute to a more equitable future.

Insurance is not just about justice, trust, and disparities, though. It is also about compensation for unpredictable financial losses under strict conditions. Alongside the legal contracts and stipulations lie their formalization and methodical analyses written in the language of mathematics and statistics, offering the needed precision.

About the workshop On August 17 and 18, 2023, the Department of Mathematical Sciences at the University of Copenhagen hosted a workshop titled *Fairness with a view towards insurance*. The workshop brought together 29 researchers from multiple areas to discuss actuarial, statistical, economic, sociological, and regulatory fairness in light of recent technical and algorithmic developments. It was organized by Christian Furrer, Munir Eberhardt Hiabu, and Mogens Steffensen. Eight talks by invited speakers illuminated the multifaceted nature of fairness, while a roundtable discussion pulled the threads together and pointed the way forward. The invited speakers were Matthias Fahrenwaldt, Thomas Hildebrandt, Fei Huang, Frederik Hytting Jørgensen, Mathias Lindholm, Joshua Loftus, Liz McFall, and Andreas Tsanakas, and the roundtable was moderated by Christian Furrer, who was joined by the panelists Matthias Fahrenwaldt, Fei Huang, Joshua Loftus, Liz McFall, and Andreas Tsanakas. Fynske Købstæders Fond and Danmarks Frie Forskningsråd supported the workshop.

Extended author information available on the last page of the article

Published online: 19 June 2024

Statistics and computer science have given birth to machine learning and artificial intelligence methods, updating the toolbox of the insurance industry's quants. These technologies support the analysis of vast amounts of data to identify patterns and predict outcomes. Insurers may use machine learning algorithms to customize policies and premiums based on individual risk characteristics or to enhance customer experience and satisfaction. These methods have, however, their own notions of fairness related to but not rooted in the historically crucial insurance context, for example, group fairness criteria such as demographic parity and individual fairness criteria such as counterfactual fairness; see also Table 3.5 in [1] for an overview of group fairness criteria. It is essential to understand the relations, consistencies, and contradictions, not only mutually among statistical notions of fairness but also between the legal and statistical notions of fairness. This is a basic condition for the insurance industry: To deliver the 'best solutions', given the demand from policyholders and society as well as the conceptual supply from statistics and computational supply from technological advances. That equilibrium is moving with the societal and computational evolution as well as the innovation power of the industry. Actuarial scientists and statisticians worldwide will be uniquely positioned to deliver the gold standard—if they seize the emerging opportunities.

Many authors have attacked that challenge already, including several of the authors of the present paper. Prominent examples include [30], in which it is discussed that the conception of fairness is dynamic and, in particular, changes over time. One recurring theme is whether characteristics that an individual has no control over and hence cannot change should be considered in insurance pricing [36]. Another theme is whether insurance should be understood as an expression of solidarity between homogeneous groups or as a fair contract between an insurer and an individual insured [2, 11]. The actuarial literature also offers specialized fairness concepts and technical solutions, see for instance [5, 23]. For an overview of different fairness concepts with a view towards insurance, consult [4, 10].

The remainder of the article is structured as follows. Section 2 contains the main insights drawn from the workshop and some avenues for future research; this includes findings related to plurality and causality as well as implications on privacy and regulation. Section 3 provides an outlook and concludes. Finally, Appendix A contains summaries of the eight workshop talks by the invited speakers.

2 Findings, implications, and opportunities

This section collects the main insights across the workshop's talks and discussions. This includes areas where consensus was established and areas with potential for academic discord. In both of these cases, important avenues for future research are highlighted.

2.1 Plurality of fairness

Models and algorithms for prediction and risk analysis must have discriminatory effects, but these may be undesired or even unlawful. Discrimination and fairness depend on context, and even within a narrow context, fairness may be contested by the different parties involved, including the insured and the insurers. One party may believe they have been harmed or otherwise negatively impacted by discrimination acting through an algorithm or model, and another party that designed or used the algorithm may believe their decisions can be legitimately justified. Thus, even without disagreement about the observed facts, there can be a disagreement about their legal, ethical, or social interpretation. It should not be surprising that different definitions of fairness can be mutually incompatible [12]. It is still worth noting the consequence: a regulation attempting to prevent one type of discrimination may also enforce discrimination of another kind. For example, a rule requiring formal non-dependence—where a model or algorithm cannot use a sensitive or protected variable as an input—can still allow so-called proxy or indirect discrimination; confer, for instance, with the 'red car scenario' of [21].

In recent years, emerging research in the actuarial community has focused on mitigating potential indirect insurance discrimination in insurance risk pricing. Meanwhile, various fairness criteria have been proposed and flourished in the statistics and machine learning literature [29]. In evaluating these methods and criteria, one should examine what they imply in concrete cases and if their application leads to counterintuitive consequences. If our intuitions about concrete cases conflict with the abstract definitions, we may try to modify the definitions. In other cases, our intuitions about concrete cases may be uncertain, and the abstract definitions can inform us about what should be done. Coming up with definitions that are robust to different scenarios is crucial.

It is important to note that fairness is a broad term that needs to be made more precise. The aforementioned actuarial, statistics, and machine learning literature is concerned with *algorithmic fairness*: a rule-based approach that aims at ensuring uniform treatment of groups of people. Usually, this entails statistical metrics. In addition, there is *discrimination* in the stricter legal sense—enshrined in consumer protection legislation and concerning individual rights—and finally *bias*, for instance, in the data used for training an algorithm, but also introduced by the user when applying the algorithm.

In [23], a discrimination-free insurance pricing technique is proposed. It is consistent with the model introduced in [32] and aims to mitigate proxy discrimination. It has been shown that the resulting fairness concept conflicts with group fairness: satisfying the one does not mean the other is satisfied; adjusting for one may undermine the other [25]. The degree to which the technique of [23] specifically, and fair pricing techniques in general, rely on the notion of causality is contested—and no consensus was reached during the workshop. However, the technique may be embedded in a causal framework, whence its appropriateness in different scenarios may be studied through that lens [15].

2.2 Causal notions

From the previous subsection, it is evident that some mathematically founded guidance on comparing and choosing among different fairness definitions could be helpful, given the mutual incompatibility of a plurality of definitions. Causal fairness is one high-level framework or research program that attempts unification by asserting that fairness or discrimination must be understood in terms of causal relationships using explicitly causal methodology [6, 18, 20–22, 31]. For a recent survey, see [27]. This can also be viewed as one application area of a larger program on causal machine learning [17].

Many statistical fairness criteria are defined in terms of conditional probability distributions. One source of the multiplicity of definitions is the choice of which variables to include in the condition of such conditional distributions. By distinguishing between observation and intervention, causal models can help decide which variables to condition on for observational purposes. Causal fairness criteria recast the choice of which variables to condition on as a choice about pathways in a graphical causal model [6, 18, 21, 22, 31]. Visual representation of causal pathways in a graph could help guide such choices and facilitate understanding for a wider group of stakeholders. More generally, using interpretable models or visualization methods [26] could help all parties better understand the limitations and consequences of using a particular model. Intersectional fairness and discrimination can become an important issue when there are multiple categories or sensitive variables involved [37], and in such cases, the expressiveness of causal models may be particularly helpful [3].

There is a long tradition of using predictive models, with no explicitly causal assumptions, to make decisions. This is based on hopes that decision-makers understand the differences between observation and action, prediction and intervention. This status quo has been broadly criticized [38]. And now, the increasing prevalence of causal modeling provides an alternative, shifting the focus from passive prediction to actions, interventions, and consequences.

Justification processes for indirect discrimination are potentially too permissive. There may be many associated or correlated variables, and an algorithm or model user seeking to avoid responsibility can search among these to find ones that excuse the appearance of discrimination. A trustworthy causal model for how the world works could limit such a search. When a model or algorithm is used for many impactful decisions—consider credit risk scores—the output of such a model becomes what we might call a ‘universal collider’, and its use for decisions will induce associations among the input variables [9]. Measuring the harms or costs of this problem is empirically challenging and may be a Sisyphean task if these same dynamics invalidate the attempt to measure them. But this may be unavoidable: these issues can invalidate justifications for any given fairness criteria if, for example, the reason for a particular association is due to something like collider bias rather than individual circumstances and decisions.

2.3 Data availability and privacy

To achieve many of the notions of fairness, we need to collect the protected attributes of individuals. However, this information is usually wholly missing or only partially available. How can the lack of data be overcome while satisfying privacy requirements? Discrimination-free insurance prices may still be calculated if partial data on sensitive attributes are available [24]. Still, depending on the jurisdiction and the attribute, even partial collection of such data may be problematic—this also highlights the role of regulators in establishing when, how, and for which purposes sensitive data can be collected, used, and stored. Furthermore, regulators and insurers must be able to communicate the privacy implications to policyholders.

In some areas of insurance, the many discriminatory effects of pricing would perhaps dissipate if sufficiently detailed policyholder information is collected. Indeed, demographic characteristics are often only proxies for risk drivers such as policyholder behavior. Policyholder behavior may, increasingly, be measured via individualized data collection (for example, wearables and telematics). But even beyond concerns around surveillance and privacy, words of caution are due. Constructivist theories about social categories such as race and gender imply that almost all other variables will be associated with these because historical inequalities make these attributes influence nearly all life experiences [14, 19]. After all, categorizing attributes as sensitive may itself be the result of historical circumstances.

While the increasing use of data does not alter the fundamental issues of insurance discrimination, it is changing how insurers do business [28]. A concern is that the resulting highly individualized or personalized rates can make insurance unaffordable or unavailable for some high-risk consumers. For example, as the insurance market has moved towards more individualized risk pricing in the past decades, with the availability of new technologies to measure better and understand risks, more homes in the most disaster-prone regions are facing difficulties in purchasing home insurance.

Considering the broader picture, decision systems based on predictions will often be favored in well-known, data-rich environments. If we rely on these too much and refuse to take action outside such environments, we could miss a lot of potential benefits. This lost potential would be distributed unequally, concentrated among the most underserved people as they often reside in the most data-poor environments. Further, decision systems targeting small units, such as individual people, can miss both risks and opportunities of investments in groups of dependent units or the transformation of environments. More generally, the closed nature of fully specified formal models artificially constrains action spaces. This leads to lost opportunities when those models and constrained action spaces differ from the real world and its action spaces. The growth of systems of data collection and automated decisions should provide us with more free resources and options for innovation and not simply predict—and enforce—more of the same.

2.4 Societal attention

Insurance can be treated as a social good, an economic commodity, or something in between. When insurance is mandatory, or nearly so, it becomes less of a financial commodity and more of a social good, resulting in different attitudes toward fairness. Modern-day insurance is generally sponsored at significant levels, either by governments or private corporations, of which policyholders can own the latter, as is the case for mutual and takaful companies, or by investors. Pool members may feel a form of insurance solidarity, but the responsibilities of the pool depend on its nature. For example, in the case of the pool of contracts issued to individuals by a for-profit stock company, the pool can be thought of as a sum of bilateral agreements that leaves out the collective dimension of insurance. Actuarial fairness applied in this context might mean that each customer should pay for their own risk and only their own risk. However, subsidies from one group to another are typical for social insurance, where a government entity owns the pool.

The pool's responsibility, protected attributes, and economic considerations such as adverse selection, moral hazard, and financial efficiency jointly determine the appropriateness of insurance discrimination (differentiation), which depends on the context and varies by lines of business and jurisdiction. No intent on the insurers' side is necessary for discriminatory effects to occur.

Consumer protection is a cornerstone of financial services regulation. Financial services such as banking or insurance services must not unfairly discriminate against customers. The European Union's legal framework distinguishes between direct and indirect discrimination. While unfair discrimination has always been part of a supervisor's agenda, machine learning, and the corresponding ever-growing data requirements can scale the problem, for example, in automated decision-making with little or no human oversight.

Accordingly, the European Union is introducing new legislation in the form of the AI Act. This aims to classify the use of machine learning in three risk tiers. The AI Act guides the use of such models in the 'high risk' category, such as disclosure towards customers, governance, and human oversight. This risk tier currently explicitly includes risk assessment and pricing in life/health insurance, but the scope may be adjusted at a later stage.

The challenge for financial institutions, regulators, and supervisors alike is to find a practical way to ensure practices are free of unfair discrimination against customers. Supervisors can expect banks and insurance companies to proactively identify sources of such unfairness and take measures to avoid unfair discrimination. Such measures will likely exceed simple metrics and require extensive business knowledge and additional human oversight.

3 Outlook

The above discussions have, at least at certain stages, made the implicit assumption that the technical actuarial prices are the prices charged in an insurance market. However,

insurance pricing is a complicated process that may involve cost modeling (risk pricing), demand modeling, and price optimization, depending on the line of business and jurisdiction. The existing actuarial research predominantly focuses on the risk pricing stage, which is a narrow focus as discrimination could appear at all stages of the pricing process; see however [33]. More research—and broader research—is needed, covering a more comprehensive range of insurance practices, including underwriting, pricing, marketing, claims processing, fraud detection, etc.

Major open questions can be split into three related dimensions. First, from a technical perspective, mathematical methods are needed to directly avoid or at least identify potential cases of unfair discrimination with a high accuracy rate. Typically used explainability tools, which supposedly improve machine learning methods' transparency, are lacking in many regards. For instance, it is easy to construct examples where an algorithm suggests a highly unfair decision, while these tools do not identify this issue [34]. This calls for further research on interpretability, building on sophisticated statistical frameworks such as causal graphs.

Furthermore, regulators and supervisors must—in conjunction with the financial services industry and based on current technical developments—define supervisory expectations relating to fairness and unfair discrimination, not least in the context of artificial intelligence and machine learning. These expectations should include minimum skills, processes, and human oversight requirements.

Finally, the broader actuarial community should develop guidelines that support informed decision-making regarding fairness criteria across insurance contexts. The fairness and accuracy trade-off usually discussed in the machine learning literature conflicts with the goals of business decision-making. On the contrary, machine learning scholars might take insight from the interdisciplinarity of the actuarial literature [13]. Nevertheless, a more business- and society-focused framework based on stakeholder analyses is needed. This is closely related to another critical yet underexplored question: How does fairness impact stakeholders, and who pays the cost of fairness? While a first attempt to empirically answer this question is provided in [33], more research is needed to understand the impacts of various fairness policies to inform optimal decision-making by businesses and regulators. We, therefore, call for high-quality datasets to perform empirical research in this area.

Across all these dimensions, technical actuarial considerations are essential to clarify the contours. Let us, however, conclude by echoing the message of [7]: Resolutions require the interdisciplinary engagement of expertise from fields such as actuarial science, statistics, and computer science, but also the social sciences, including jurisprudence and economics.

A Summary of talks

This appendix briefly summarizes the eight talks by the workshop's invited speakers. They are presented in an appropriate sequence according to topic and focus.

Thomas Hildebrandt: Transparent, Adaptable and Explainable Decision and Process Models as Basis for Fairness in Automated Decision Support

Hildebrandt presented the current status of more than 15 years of research and development of methods for transparent, adaptable, and explainable modeling of decisions and processes using the Dynamics Condition Response graph technology. Focusing on the EcoKnow Innovation Foundation Grand Solutions Project (2017–2021), which concerned effective co-created and compliant adaptive case management for knowledge workers, he summarized a range of pitfalls and challenges. These included too late involvement of users affected by the system, poor model documentation, vague relation between data and practice, bias in data, and lack of explainability. Hildebrandt concluded that although AI models, like all models, are wrong, using them can still make sense. They give rise to new challenges and pitfalls during all phases of development, from the idea to the system's disposal. This calls for the use of standards and guidelines for software development.

Andreas Tsanakas: A multi-task network approach for calculating discrimination-free insurance prices

Following a brief introduction to proxy discrimination and demographic unfairness, Tsanakas introduced discrimination-free insurance pricing [23]. This approach suggests building a best estimate model using all policyholder characteristics (including protected ones) and then averaging out the protected characteristics. However, this requires full knowledge of the policyholder's protected characteristics, which may be problematic, not least due to regulatory restrictions. Following [24], Tsanakas showed how, given only partial information on protected characteristics, a multi-task network structure can be used to estimate discrimination-free prices. The method was showcased as performing well on both synthetic and real-world data.

Mathias Lindholm: Discrimination and fairness in insurance pricing

Lindholm discussed differences between indirect discrimination and popular group fairness axioms based on [25]. Simply disregarding protected policyholder characteristics to form the unawareness price still allows for the possibility of using the protected attributes from non-protected characteristics; this is what leads to so-called indirect or proxy discrimination. Lindholm showed that while discrimination-free insurance pricing prevents proxy discrimination, this is not generally the case for various criteria for algorithmic fairness, such as demographic parity. Demographic parity and discrimination-free insurance pricing are thus different concepts with different objectives. However, the independence between protected and non-protected policyholder characteristics is a sufficient condition for the unawareness price to satisfy demographic parity and be a discrimination-free insurance price. This suggests searching for transformed non-protected policyholder characteristics with such features, using, for example, input or output optimal transport.

Joshua Loftus: Using-causality for model explanations and fairness

Loftus began his talk by giving an overview of causality and its application as a research program to fairness and discrimination as well as explainability. The concept of counterfactual fairness [21] and various extensions and alternatives based on pathway analyses and decompositions were introduced and discussed. Further, Loftus showed how structural causal models such as causal dependence plots [26] can

be used to probe the limitations of popular interpretable machine learning tools like partial dependence plots.

Frederik Hytting Jørgensen: Unfair Utilities and First Steps Towards Improving Them

Many fairness criteria constrain which or how policyholder characteristics can be used. Jørgensen presented the paper [16], which proposes a different framework for thinking about fairness, namely to consider instead which utility a policy is optimizing for. Defining the criterion *value of information fairness*, it is suggested not to use utilities that do not satisfy this criterion. Protected characteristics have *value of information* relative to a set of decision inputs if one can obtain a strictly larger expected utility in case the policy may also depend on the protected characteristics, and a utility is *value of information fair* if this is not the case. Through various examples, Jørgensen showed how the theory might be employed and how it could impact prediction.

Fei Huang: Fairness-aware Insurance Pricing: Principles, Fairness Criteria, and Welfare Implications

Based on her recent work [10, 33, 39], Huang discussed the following three research questions: What are social and economic principles to assess the appropriateness of insurance discrimination? How can we match the existing and potential anti-discrimination insurance regulations with fairness criteria and develop pricing models to mitigate indirect discrimination? What are the welfare implications of existing and potential fair insurance pricing regulations on consumers and firms? Among the insights presented were that fairness depends on the context and whether insurance is considered a social good or an economic commodity, that there is a need to explore the welfare consequence of regulation on cost modeling and pricing, and that even a small change in accuracy can leave behind a significant and heterogeneous welfare impact.

Liz McFall: The trouble with fairness in insurance: a situated account

McFall considered challenges in insurance presented by fairness as a goal and a claim from a sociological and ethnographic perspective. Based on the provocation that fairness in insurance is a situated judgment that may not be resolvable statistically, she presented one historical and two contemporary examples of the challenges presented by fairness: Industrial Life (poor) Assurance, the Patient Protection and Affordable Care Act (Obamacare), and EIOPA's 2021 report on artificial intelligence governance principles [8], in particular the section on fairness.

Matthias Fahrenwaldt: Fairness and explainability of machine learning methods in risk models

Existing and emerging regulation for financial institutions entails fairness, ethics, and accountability. While this is more prominent in customer-facing processes, Fahrenwaldt highlighted how it may also be relevant in risk models insofar as the fairness of an internal model, for instance, might be assessed by explaining the workings of the model. From this point of departure, Fahrenwaldt provided examples of regulatory approaches to fairness and a supervisory perspective on the explainability of complex models. This included the FEAT framework of Singapore and the upcoming EU AI

Act. Concluding, caution was expressed about the ability of explainability techniques to 'guarantee' fairness and to be accepted by courts of law.

Acknowledgements We want to thank all participants for stimulating discussions. Special thanks to Liz McFall from the School of Social and Political Science at the University of Edinburgh and Thomas Hildebrandt from the Department of Computer Science at the University of Copenhagen for their insightful and thought-provoking presentations. Thanks also to Thomas Mikosch from the Department of Mathematical Sciences at the University of Copenhagen for helping to support the initiative.

Funding Open access funding provided by Copenhagen University

Declarations

Conflict of interest The author (s) reported no potential conflicts of interest. The views expressed are those of the author(s) and do not necessarily reflect those of the organizations with which the author(s) are affiliated. This paper should not be reported as representing the views of any of those organizations.

References

1. Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: limitations and opportunities. https://fairmlbook.org/pdf/fairmlbook.pdf. Accessed 21 May 2024
2. Barry L, Charpentier A (2020) Personalization as a promise: Can Big Data change the practice of insurance? Big Data Soc 7:1–12. https://doi.org/10.1177/2053951720935143
3. Bright L, Malinsky D, Thompson M (2016) Causally interpreting intersectionality theory. Philos Sci 83:60–81. https://doi.org/10.1086/684173
4. Charpentier A (2024) Insurance, biases, discrimination and fairness. Springer Actuarial. Springer, Cham (**to appear**)
5. Chen A, Vigna E (2017) A unisex stochastic mortality model to comply with EU Gender Directive. Insur Math Econ 73:124–136. https://doi.org/10.1016/j.insmatheco.2017.01.007
6. Chiappa S (2019) Path-specific counterfactual fairness. In: Proceedings of the AAAI Conference on artificial intelligence, pp 7801–7808, https://doi.org/10.1609/aaai.v33i01.33017801
7. Dolman C, Frees E, Huang F (2021) Multidisciplinary collaboration on discrimination—not just "Nice to Have". Ann Actuarial Sci 15(3):485–487. https://doi.org/10.1017/S174849952100021X
8. EIOPA's Consultative Expert Group on Digital Ethics in insurance (2021) Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the European insurance sector. Tech. rep., European Insurance and Occupational Pensions Authority (EIOPA), https://www.eiopa.europa.eu/system/files/2021-06/eiopa-ai-governance-principles-june-2021.pdf. Accessed 21 May 2024
9. Elwert F, Winship C (2014) Endogenous selection bias: the problem of conditioning on a collider variable. Ann Rev Sociol 40:31–53. https://doi.org/10.1146/annurev-soc-071913-043455
10. Frees E, Huang F (2023) The discriminating (Pricing) actuary. N Am Actuarial J 27:2–24. https://doi.org/10.1080/10920277.2021.1951296
11. Frezal S, Barry L (2020) Fairness in uncertainty: some limits and misinterpretations of actuarial fairness. J Bus Ethics 167:127–136. https://doi.org/10.1007/s10551-019-04171-2

12. Friedler S, Scheidegger C, Venkatasubramanian S (2021) The (Im) possibility of fairness: different value systems require different mechanisms for fair decision making. Commun ACM 64(4):136–143. https://doi.org/10.1145/3433949
13. Fröhlich C, Williamson R (2024) Insights from insurance for fair machine learning, preprint, arxiv:2306.14624
14. Hu L (2023) What is Race in algorithmic discrimination on the basis of race? J Moral Philos 1:1–26. https://doi.org/10.1163/17455243-20234369
15. Ituria C, Hardy M, Marriott P (2022) A discrimination-free premium under a causal framework, preprint, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4079068
16. Jørgensen F, Weichwald S, Peters J (2023) Unfair utilities and first steps towards improving them, preprint, arxiv:2306.00636
17. Kaddour J, Lynch A, Liu Q, et al (2022) Causal machine learning: a survey and open problems, preprint, arxiv:2206.15475
18. Kilbertus N, Rojas Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B (2017) Avoiding discrimination through causal reasoning. Adv Neural Inf Proc Syst 30
19. Kohler-Hausmann I (2019) Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. Northwest Univ Law Rev 113(5):1163–1228
20. Kusner M, Loftus J (2020) The long road to fairer algorithms. Nature 578(7793):34–36. https://doi.org/10.1038/d41586-020-00274-3
21. Kusner M, Loftus J, Russell C, et al (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, et al (eds) Advances in neural information processing systems
22. Kusner M, Russell C, Loftus J, Silva R (2019) Making decisions that reduce discriminatory impacts. In: International Conference on Machine Learning, pp. 3591–3600. PMLR
23. Lindholm M, Richman R, Tsanakas A et al (2022) Discrimination-free insurance pricing. ASTIN Bull 52:55–89. https://doi.org/10.1017/asb.2021.23
24. Lindholm M, Richman R, Tsanakas A et al (2023) A multi-task network approach for calculating discrimination-free insurance prices. Eur Actuarial J 11:1–41. https://doi.org/10.1007/s13385-023-00367-z
25. Lindholm M, Richman R, Tsanakas A, et al (2023) What is fair? Proxy discrimination vs. demographic disparities in insurance pricing, preprint, https://openaccess.city.ac.uk/id/eprint/30549/1/. Accessed 21 May 2024
26. Loftus J, Bynum L, Hansen S (2023) Causal dependence plots for interpretable machine learning, preprint, arxiv:2303.04209
27. Makhlouf K, Zhioua S, Palamidessi C (2022) Survey on causal-based machine learning fairness notions, preprint, arxiv:2010.09553
28. McFall L, Meyers G, Hoyweghen IV (2020) Editorial: the personalisation of insurance: data, behaviour and innovation. Big Data Soc 7(2):10. https://doi.org/10.1177/2053951720973707
29. Mehrabi N, Morstatter F, Saxena N et al (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607
30. Meyers G, Hoyweghen IV (2018) Enacting actuarial fairness in insurance: from fair discrimination to behaviour-based fairness. Sci Cult 27(4):413–438. https://doi.org/10.1080/09505431.2017.1398223
31. Nabi R, Shpitser I (2018) Fair inference on outcomes. In: Proceedings of the AAAI Conference on artificial intelligence, https://doi.org/10.1609/aaai.v32i1.11553
32. Pope D, Sydnor J (2011) Implementing anti-discrimination policies in statistical profiling models. Am Econ J Econ Pol 3(3):206–231. https://doi.org/10.1257/pol.3.3.206
33. Shimao H, Huang F (2022) Welfare implications of fairness and accountability for insurance pricing, preprint, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4225159
34. Slack D, Hilgard S, Jia E, et al (2020) Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 180–186. https://doi.org/10.1145/3375627.3375830
35. Squires G (2003) Racial profiling, insurance style: insurance redlining and the uneven development of metropolitan areas. J Urban Aff 25(4):391–410. https://doi.org/10.1111/1467-9906.t01-1-00168
36. Thiery Y, Schoubroeck CV (2006) Fairness and equality in insurance classification. Geneva Papers on Risk Insur-Issues Pract 31(2):190–211. https://doi.org/10.1057/palgrave.gpp.2510078
37. Wang A, Ramaswamy V, Russakovsky O (2022) Towards intersectionality in machine learning: including more identities, handling underrepresentation, and performing evaluation. In: Proceedings of the

2022 ACM Conference on Fairness, Accountability, and Transparency, pp 336–349, https://doi.org/10.1145/3531146.3533101
38. Wang A, Kapoor S, Barocas S et al (2023) Against predictive optimization: on the legitimacy of decision-making algorithms that optimize predictive accuracy. ACM J Responsib Comput. https://doi.org/10.1145/3636509
39. Xin X, Huang F (2023) Antidiscrimination insurance pricing: regulations, fairness criteria, and models. N Am Actuarial J. https://doi.org/10.1080/10920277.2023.2190528

Authors and Affiliations

Matthias Fahrenwaldt[1,2] · Christian Furrer[1] · Munir Eberhardt Hiabu[1] · Fei Huang[3] · Frederik Hytting Jørgensen[1] · Mathias Lindholm[4] · Joshua Loftus[5] · Mogens Steffensen[1] · Andreas Tsanakas[6]

✉ Christian Furrer
furrer@math.ku.dk

✉ Munir Eberhardt Hiabu
mh@math.ku.dk

✉ Mogens Steffensen
mogens@math.ku.dk

1 Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

2 Federal Financial Supervisory Authority (BaFin), Bonn, Germany

3 School of Risk and Actuarial Studies, University of New South Wales, Sydney, Australia

4 Department of Mathematics, Stockholm University, Stockholm, Sweden

5 Department of Statistics, London School of Economics and Political Science, London, UK

6 Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, London, UK