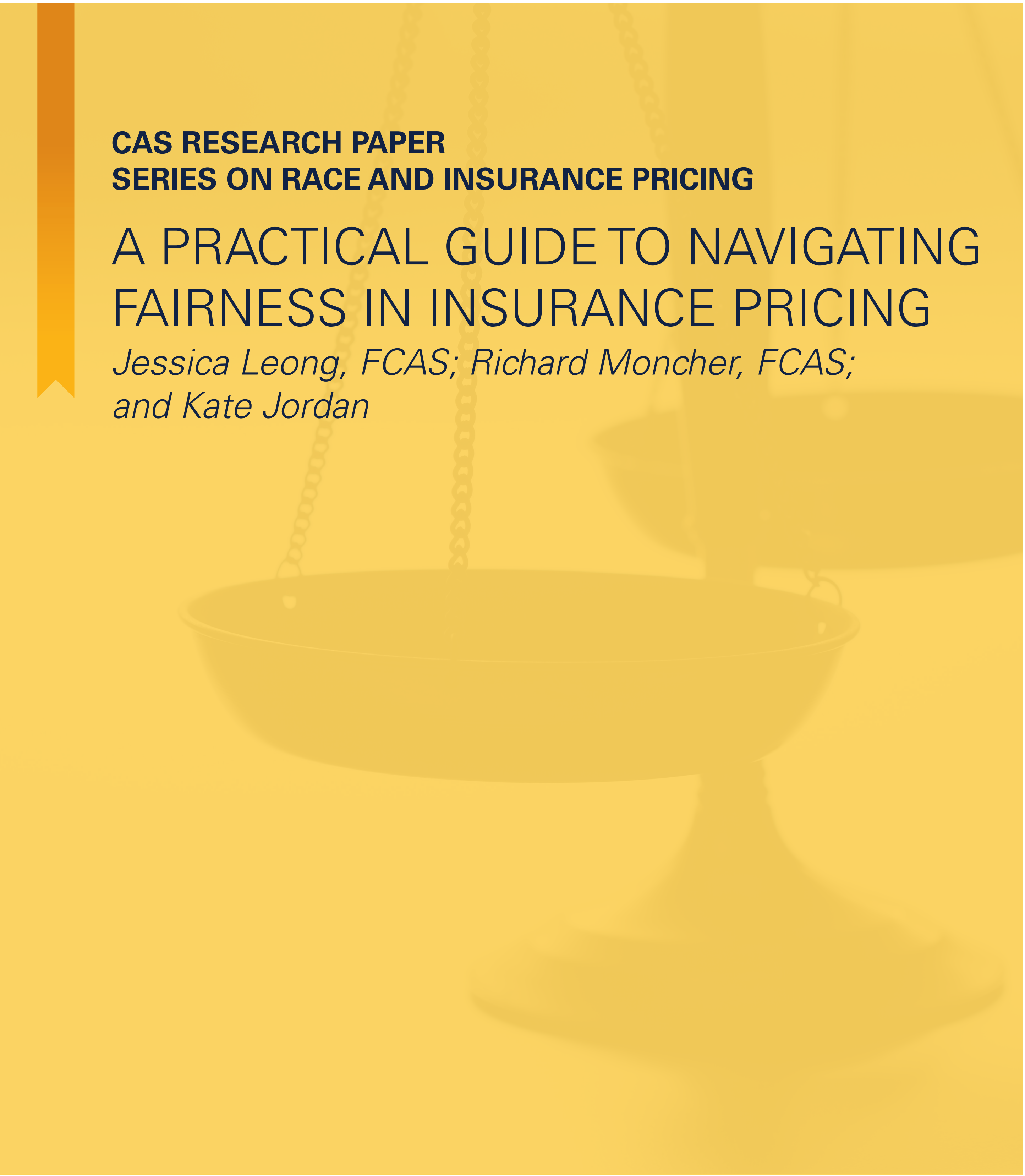
CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

A PRACTICAL GUIDE TO NAVIGATING FAIRNESS IN INSURANCE PRICING

Jessica Leong, FCAS; Richard Moncher, FCAS; and Kate Jordan



CASUALTY ACTUARIAL SOCIETY

The CAS Research Paper Series on Race & Insurance Pricing was created to guide the insurance industry toward proactive, quantitative solutions that address potential racial bias in insurance pricing. These reports explore different aspects of unintentional potential bias in insurance pricing, address historical foundations and offer forward-looking solutions to quantify and handle possible bias. Through these reports, the CAS aims to encourage actuaries to discuss this topic with their stakeholders across all areas of insurance pricing and operations. For more information on the series, visit casact.org/raceandinsuranceresearch.

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing, professional education and research. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property-casualty risks and serves over 10,000 members worldwide. CAS members are sought after globally for their insights and ability to apply analytics to solve insurance and risk management problems. As the world's premier P&C actuarial research organization, the CAS reaches practicing actuaries across the globe with thought-leading concepts and solutions. The CAS has been conducting research since its inception. Today, the CAS provides thousands of open-source research papers, including its prestigious publication, *Variance* — all of which advance actuarial science and enhance the P&C insurance industry. Learn more at casact.org.

Caveat and Disclaimer

This research paper is published by the Casualty Actuarial Society (CAS) and contains information from various sources. The study is for informational purposes only and should not be construed as professional or financial advice. The CAS does not recommend or endorse any particular use of the information provided in this study. The CAS makes no warranty, express or implied, or representation whatsoever and assumes no liability in connection with the use or misuse of this study. The views expressed here are the views of the authors and not necessarily the views of their current or former employers.

CAS RESEARCH PAPER
SERIES ON RACE AND INSURANCE PRICING

A PRACTICAL GUIDE TO NAVIGATING FAIRNESS IN INSURANCE PRICING

Jessica Leong, FCAS; Richard Moncher, FCAS; and Kate Jordan

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

Contents

A Practical Guide to Navigating Fairness in Insurance Pricing

By Jessica Leong, FCAS; Richard Moncher, FCAS; and Kate Jordan

Executive Summary

As the insurance industry increases its use of models, machine learning, and artificial intelligence, regulatory scrutiny has intensified, particularly with regard to concerns about unfair discrimination. In the past, regulation centered around model inputs, specifying certain variables that could not be used. Now, new regulations are emerging around testing model outcomes for unfair discrimination, with requirements to report the findings.

This paper provides actuaries with information and tools to proactively consider fairness in their modeling process and navigate this new regulatory landscape. The paper's key points are summarized as follows:

- **Unfair discrimination can arise from several sources.** The data may be insufficient or have underlying bias. The model may have bias – for example, producing different error rates for different groups of interest. Or the result of the model may be deemed unfair even if the underlying data shows that there are differences between groups of interest. For example, in several U.S. states and the European Union, auto insurance rates cannot be set by gender. These jurisdictions have deemed this unfair, even if the data shows true differences in outcomes.
- **Regulators have started to mandate the testing of model outcomes for unfair discrimination.** Some examples include Colorado's insurance statute broadening the definition of "unfair" and the District of Columbia conducting a test for unfair discrimination for private passenger auto insurance in 2023. Outside the insurance sector, New York City now requires an independent bias audit of the selection rate by race/ethnicity and gender when automated employment decision tools are used.
- **Fairness can be considered throughout the modeling life cycle.** The model development process includes many iterative steps. Many choices are made along the way, and many potential alternative models are created. When it comes to measuring unfair discrimination, these alternative models will fall along a spectrum of results, and a reasonable solution may be among them. Fairness considerations arise during the following phases of model development:
  1. **Model governance.** Setting the stage for a carrier's consideration of unfair discrimination, model governance establishes the organization's

broad philosophy and approach. How the organization will follow that philosophy is an essential starting point. In particular, how will the organization approach the potential need to infer data on race and ethnicity?

2. **Project planning.** How a business problem is translated into an analytical problem can significantly affect the potential for unfair discrimination and disparate impact. For example, Elzayn et al. (2023) researched U.S. Internal Revenue Service models created to select targets for audits. Disparate impact was reduced by changing from a model that predicted a binary outcome (compliance or noncompliance) to one that predicted the amount of money people failed to report.
3. **Data preparation.** Data quality, credibility, and variability are critical factors in assessing potential unfair discrimination. Actuaries should analyze data sets for early indicators of unfair discrimination and consider data improvement strategies. They can also keep in mind fairness considerations as they transform data, such as how they infer missing values, as well as develop and cap losses.
4. **Modeling.** During modeling, fairness can be considered when grouping data for variables, as well as when interpolating or extrapolating variables. Other levers that can be pulled include changing training/test data set partitions, removing or replacing variables, and penalizing models by incorporating a fairness metric.
5. **Model implementation.** A model itself is neither fair nor unfair, but how it is used and how it impacts people may result in unfair discrimination. Therefore, model implementation must be included in considerations on how to increase fairness. For example, a “human in the loop” may be considered to mitigate unfair discrimination.

Compliance with evolving regulations requires collaboration across the organization. By following the outlined steps, implementing proactive measures, and adapting to new developments, actuaries can play a vital role in achieving fair and equitable outcomes in pricing, underwriting, claims, and other models.

1. Introduction, Goal, and Scope

In the current artificial intelligence (AI) and machine learning environment, regulations regarding unfair discrimination are evolving. In particular, a new practice is emerging around testing model outcomes for unfair discrimination, with requirements to report on the findings. This paper aims to arm actuaries with the tools to consider fairness at each step of the modeling process so they can comply with emerging U.S. regulations.

2. What Do We Mean When We Say, “Unfair Discrimination”?

The traditional actuarial understanding of “fair” comes from the Casualty Actuarial Society’s Statement of Principles Regarding Property and Casualty Insurance Ratemaking, which states:

> A rate is reasonable and not excessive, inadequate, or unfairly discriminatory if it is an actuarially sound estimate of the expected value of all future costs associated with an individual risk transfer.

In many states, regulations require that rates not be excessive, inadequate, or unfairly discriminatory.

Society and the state of analytics are changing, as are regulations. Over the last decade, there has been a significant increase in the use of predictive analytics and external data and the rise of insurtech. Outside the insurance sector, many industries have seen the dawn of a new age of AI.

Some states, such as Colorado, are adopting a broader definition of unfair discrimination, which is explained as follows (Colo. Rev. Stat. § 10-3-1104.9(8)(e)):

> “Unfairly discriminate” and “unfair discrimination” include the use of one or more external consumer data and information sources, as well as algorithms or predictive models using external consumer data and information sources, that have a correlation to race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression, and that use results in a disproportionately negative outcome for such classification or classifications, which negative outcome exceeds the reasonable correlation to the underlying insurance practice, including losses and costs for underwriting.

A key difference in Colorado’s approach is the focus on model outcomes. Indeed, as models grow increasingly sophisticated, some regulators inside and outside the insurance sector are moving toward an outcomes-based approach. This reduces the need to police the ever-growing diversity of input data and understand increasingly complex models, and it cuts to the heart of what matters – the outcomes.

The concepts of bias and unfair discrimination in algorithms are used broadly, and it’s helpful to understand all the potential sources of unfair discrimination. Following is a list, but not an exhaustive one, categorizing sources of unfair discrimination and examples:

1. **Statistical bias in the model.** The mathematical approach has different error rates for different groups of interest.
2. **Bias due to data completeness or diversity.** The analytics team has data on protected classes only in urban areas and extrapolates that data to rural areas where the same patterns may not hold.
3. **Unfair discrimination in the underlying data.** Data on traffic violations may be biased if law enforcement is applied inconsistently.
4. **Something is deemed unfair even if the data shows that a particular result differs for certain groups of interest.** In several U.S. states and the European Union (EU), for example, auto insurance rates cannot be set by gender. These jurisdictions have deemed this practice unfair, even if the data shows true differences in outcomes. For an actuary, this can seem unfair, as this may result in some risks being charged more than their loss costs imply and some being charged less. However, fairness is not purely a mathematical idea; it is for society to decide. What is considered fair can change over time and among groups.

3. Overview of Evolving Regulations on Unfair Discrimination

It's useful for a practitioner to have an overview of the evolving regulations relating to AI and unfair discrimination. Note that the regulations continue to change, and the information in this paper is accurate as of the time of writing.

Table 1 shows the potential expected impacts of new regulations or guidance for property and casualty (P&C) insurers from selected jurisdictions. The text that follows provides more details for each jurisdiction.

California

In response to concerns about potential bias in insurance, Commissioner Ricardo Lara issued California Department of Insurance (CDI) Bulletin 2022-05 in 2022. The bulletin emphasized that insurers must adhere to principles of fairness, treating all similarly situated individuals equally in their marketing, policy issuance, pricing, fraud investigation, and claims handling. Commissioner Lara warned insurers against total reliance on algorithms.

The bulletin explicitly mandates that California insurers actively avoid conscious and unconscious bias or discrimination when using AI and other forms of big data. It further requires insurers to provide specific reasons for using algorithms to restrict or decline insurance coverage, to increase premiums, or to take other adverse actions.

Table 1. Jurisdiction Summary Comparison

Jurisdiction	Date	Legislation/ Regulation/Circular	Potential Impact for P&C Insurers
California	June 30, 2022	Bulletin 2022-05 issued by Insurance Commissioner Ricardo Lara	• Insurers must adhere to fairness principles and avoid bias in marketing, policy issuance, pricing, fraud investigation, and claims handling. • Insurers face increased scrutiny of external consumer data and information sources (ECDIS) and algorithms and potential disciplinary action by the California Department of Insurance.
Colorado	July 2021	Legislation S.B. 21-169	• For personal lines rating, risk acceptance, marketing, and claims management, carriers will need to test their model outcomes for unfair discrimination. • A robust governance and reporting framework is also required.
Connecticut	September 1, 2022 deadline	Certification Notice: Big Data and Avoidance of Discriminatory Practices	• Insurers are mandated to comply with anti-discrimination laws and complete data certification annually.
District of Columbia	April 28, 2023 Data call deadline	DISB Request for Data – Private Passenger Auto	• The Department of Insurance, Securities and Banking (DISB) conducted its own evaluation of unfair discrimination in the pricing of private passenger auto insurance using carrier data and released its results in May 2024.

(continued on next page)

Table 1. Jurisdiction Summary Comparison *(Continued)*

Jurisdiction	Date	Legislation/ Regulation/Circular	Potential Impact for P&C Insurers
			• The conclusion was: "Our review showed that Black and to a lesser degree Hispanic drivers paid higher premiums than white and Asian and Pacific Islander drivers. However, our analysis of losses showed even larger differentials than premiums by race. From this we concluded that a difference in premiums by race is not sufficient to establish bias."
Illinois	March 10, 2023	HB 2203 re-referred to Rules Committee	• Insurers cannot decline to issue or renew a personal auto liability policy based on certain prohibited rating/ underwriting factors. • Personal auto insurers must show that their rating, underwriting, claims, fraud, marketing, and other predictive models don't disparately impact specified protected groups.
New York	January 17, 2024	New York Department of Financial Services circular on use of ECDIS in underwriting	• The circular proposes additional scrutiny of algorithms and potential testing, risk management, and governance requirements.
USA (Federal)	October 2022 Publication	Blueprint for an AI Bill of Rights	• No real impact is expected for P&C insurers, but actuaries should understand these broader trends.

Table 1. Jurisdiction Summary Comparison *(Continued)*

Jurisdiction	Date	Legislation/ Regulation/Circular	Potential Impact for P&C Insurers
	October 30, 2023 Publication	Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (Order 14110)	• The federal blueprint acts as guidance but does not mandate any laws or regulations. State insurance regulations generally already follow this guidance.
Ontario, Canada	June 1962 Effective	Human Rights Code	• Use of various protected classes is prohibited in models with a customer-facing impact; this prohibition extends to data that may act as a proxy for protected classes.
	September 20, 2022 Effective	Operational Risk Management (ORM) Framework in Rating and Underwriting of Auto	• Regulation mandates that personal auto insurers develop an ORM framework with procedures to evaluate model/output fairness and detect/mitigate unfair discrimination.
Québec, Canada	June 1976 Effective	Charter of Human Rights and Freedoms	• Québec's charter provides a more extensive range of protections compared to those offered by other provinces.
European Union	December 2023 Proposal	Transparent Requirements for General Purpose AI Models with Global Implications and Exemptions for Open-Source Models	• The EU regulation is more traditional, with a focus on inputs and how models are created. • No real impact is expected for P&C insurers, as "high-risk" applications exclude P&C pricing and underwriting.

According to Lara (CDI 2022), the increasing use of AI, algorithms, and big data had resulted in a surge of consumer complaints in California and beyond. To address this, the CDI investigated potential bias and alleged unfair discrimination. Then, Commissioner Lara declared that the CDI has the authority to scrutinize insurers through market conduct exams by evaluating criteria, algorithms, and models and that the DCI could take disciplinary action, if necessary, especially when a model lacks a clear connection to actual losses and has the potential to discriminate against protected classes unfairly.

Colorado

The Colorado Department of Insurance (DOI) took the lead in shaping bias regulation when it enacted S.B. 21-169 in July 2021.[1] The bill requires that insurers assess their "external consumer data and information sources" (ECDIS), algorithms, and predictive models to prevent unfair discrimination against consumers based on protected classes. This prohibition applies to personal and small commercial lines and any insurance practice, including pricing, underwriting, marketing, and claims management.

The Colorado DOI gathered feedback from stakeholder meetings and decided to direct insurers to focus on race and ethnicity as a first step.

The Colorado DOI began by issuing a draft regulation for life insurance. The draft regulation on the use of algorithms and predictive models was released by the DOI in 2023.[2] It mandates that insurers infer the race and ethnicity of life insurance applicants and test algorithms and models that use ECDIS for potential unfair discrimination. This draft regulation requires annual testing of model outcomes for unfair discrimination and mandates corporate governance via a risk management framework. This testing of model outcomes is new for insurers, as it deviates from the historical regulatory focus on inputs and justification of rates based on loss costs. The draft requirements for the risk management framework include detailed inventories and descriptions of ECDIS, algorithms, and models, as well as the establishment of cross-functional governance committees.

Please see section 7. Appendix for a more detailed look at Colorado's draft regulation on Quantitative Testing for Unfairly Discriminatory Outcomes for Algorithms and Predictive Models Used for Life Insurance Underwriting, current as of this paper's August 2024 publication date.

Connecticut

In 2022, the Connecticut Insurance Department (CID) issued a notice on the use of big data and the avoidance of discriminatory practices, instructing insurers to comply with state and federal antidiscrimination laws and requiring them to complete a data certification by September 2022 and annually after that (CID 2022). This certification confirms that the carrier has a process for addressing the use of third-party data that is in line with CID's guidance, and that upon the department's request, the carrier will make available the data used to build the models included in all rates, forms, and underwriting filings.

Taking an additional step toward comprehensive oversight, in October 2022, CID consolidated all actuarial and data science functions into a single unit. This strategic move aimed to enhance regulatory oversight over AI, big data, and machine learning, focusing on consumer protection.

District of Columbia

The D.C. Department of Insurance, Securities and Banking (DISB) pioneered an initiative among all U.S. states and jurisdictions by initiating a data call to all insurers writing personal auto insurance in D.C. (D.C. DISB 2023). As in Colorado, model outcomes will be tested for unfair discrimination, and, DISB said, "The central purpose of these tests is to measure differences in underwriting decisions or pricing between applicants of different races or ethnicities."

The market conduct draft report (D.C. DISB 2024), released in May 2024, is a good illustration of what a test of outcomes looks like. Key findings include the following:

- "The average annual premium is $705 for white drivers, $1,031 for Black drivers. . . . This shows a 'Black/white premium gap' of $326."
- "Black policyholders pay more in premium compared to white policyholders – a factor of 1.39 – but generate 2.4 times the losses, on average. That means Black drivers as a group have a higher loss ratio."
- Even after accounting for several explanatory variables, such as age of driver, driving record, payment type, age of car, gender, marital status, coverage limits, policy year, new car, and prior lapse in coverage, the Black/white gap in premiums is $271.
- The authors reflect:

  > The analysis shows there is a race gap in premiums that is not explained by the explanatory factors DISB collected in this data call and analyzed here. The race gap is mirrored – in fact, magnified – in actual losses; so, while Black drivers pay higher premiums, they represent (even) higher costs to insurers.

- The report highlights two outstanding questions: "How are insurers finding and charging these higher-cost drivers, since the obvious rating factors investigated didn't seem to explain much?" and "Why are the race disparities in claims and losses so much larger than can be explained by the factors analyzed in this data call?"

Illinois

H.B. 2203 (Motor Vehicle Insurance Fairness Act) amended the Illinois Insurance Code to state that insurers cannot decline to issue or renew a personal auto liability policy based on certain prohibited rating/underwriting factors. The proposed bill requires that personal auto insurers show that their rating, underwriting, claims, fraud, marketing, and any other predictive models don't disparately impact specified protected groups.

New York

The New York Department of Financial Services (NYDFS) investigated the use of external data in underwriting, which had the potential to conceal discriminatory practices prohibited by state laws for life insurance. An NYDFS circular addressed the use of sources such as geographical data, homeownership details, education level, credit information, licensures, civil judgments, and court records, which pose a risk of hidden race-based underwriting.

The NYDFS cautioned against models that use data from sources such as retail purchase history, social media activities, internet usage, mobile activity, location tracking, and the type of electronic device owned by an applicant. As defined by New York and federal laws, these practices were flagged as potentially impacting protected classes disproportionately.

On January 17, 2024, the NYDFS released a proposed circular letter, Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing (NYDFS 2024). The circular asks insurers to evaluate ECDIS and AI utilization by testing, risk management, and governance, and states that "11 NYCRR § 90.2 requires an insurer to have a corporate governance framework that is appropriate for the nature, scale, and complexity of the insurer." The circular further states that the testing, risk management, and governance should at least include documenting processes, auditing, and senior management and board oversight. Further, this NYDFS circular letter suggests, "To ensure appropriate oversight of third-party vendors, insurers should develop written standards, policies, procedures, and protocols for the acquisition, use of, or reliance on ECDIS and AIS developed or deployed by a third-party vendor." Finally, the NYDFS requested feedback on the proposed guidance by March 17, 2024.

U.S. Federal Actions

The U.S. government has issued a Blueprint for an AI Bill of Rights (White House 2022) and an Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (White House 2023). Neither impacts P&C insurers directly, instead focusing on other industries such as defense, education, and healthcare.

The federal blueprint acts as guidance but does not mandate any laws or regulations. State insurance regulations generally already follow this guidance, except for calling for a "human in the loop" to address issues on request. For example, having a person on call to articulate why the algorithm denied a customer insurance coverage is not required by state regulations.

This blueprint and executive order show the increasing regulatory scrutiny in this area in general and illustrate the broader trends relating to the regulation of algorithms.

Ontario, Canada

Ontario has a Human Rights Code that prohibits the use of various protected classes in models that have a customer-facing impact; this prohibition extends to data that may act as a proxy for protected classes, such as census data about race.

The Human Rights Code allows for four insurance-related exceptions. The law explicitly allows the use of age, sex, marital and family status, or handicap if they are bona fide and reasonable. Except for handicap, all exceptions are applied to auto insurance in Ontario.

Ontario also has an "Operational Risk Management (ORM) Framework in Rating and Underwriting of Automobile Insurance," effective September 2022 (FSRAO 2022). ORM guidance reflects the Financial Services Regulatory Authority's overall strategy to reform Ontario's auto insurance rate regulation.

This regulation mandates that personal auto insurers "should have processes and tools to ensure there is no unfair discrimination in models used for rating and underwriting, throughout the modelling process." An insurer's ORM framework should outline how it will monitor risks being managed, report risk levels to relevant stakeholders, and address risks that fall outside acceptable levels. Insurers must build tools and procedures to ensure that their rating and underwriting models are not unfairly discriminatory throughout the modeling life cycle. These procedures and tools must balance the model's predictive performance with fairness constraints, so that output is not solely maximized for accuracy.

European Union

The EU's proposed AI Act[3] takes a "risk-based approach," in that the riskier an AI application is deemed to be, the more stringent the rules. For practical purposes, there is no real impact for P&C insurers, as "high-risk" applications exclude P&C pricing and underwriting.

The regulation concerns inputs and how models are created rather than requiring the testing of outputs. The AI Act will take effect two years after final approval from European lawmakers, who are expected to vote in early 2024. As a global standard, the act could influence other regulations.

Similar to U.S. federal activity, this EU proposal highlights the increased regulatory scrutiny in general and illustrates the broader trends in the regulation of algorithms.

4. How Other Industries Address Discrimination

Unfair discrimination in algorithms is also a concern for industries outside insurance. Understanding how these industries have approached this concern can help educate our own approach.

In employment, housing, and credit, long-standing civil rights laws address discrimination through three federal acts:

- **The Equal Credit Opportunity Act (ECOA)**, originally signed into law in 1974, prohibits discrimination in any aspect of a credit transaction (FDIC 2021).
- **The Fair Housing Act (FHA)**[4] was signed into law in 1968 as part of the Civil Rights Act. It prohibits discrimination in all aspects of "residential real-estate related

transactions," including but not limited to making loans to buy a home or renting a property. While the FHA regulates more than just lending, it has significant impact on fair lending through its oversight of the mortgage industry.

- **Title VII of the Civil Rights Act of 1964** prohibits employment discrimination based on race, color, religion, sex, or national origin. It covers all employment decisions, including recruitment, selection, and terminations.

These laws prohibit both disparate treatment and disparate impact. Disparate impact is when a seemingly neutral practice unduly impacts a protected group. Disparate treatment is when someone is treated differently because they are a member of a protected group; this requires discriminatory intent.

Fair Lending

Overall, equal opportunity to credit has been a highly regulated and considered area for decades, leading to a body of analysis and litigation around the concept of "fair lending." There has been significant attention to the aspects of "disparate impact," which focuses on unacceptable outcomes and is, therefore, a relevant parallel to some of the new regulation in the P&C insurance space. For example, our discussions with practitioners in the fair lending space indicate that many lenders review their model outcomes by race and ethnicity. In Section 5, "How to achieve fair outcomes in modeling," we reference several approaches from the fair lending space.

There are a few notable differences between fair lending and P&C insurance. Both the ECOA and FHA are largely reactive, relying on lawsuits to identify and enforce their regulations. This is a meaningful difference from the P&C insurance industry, which is largely focused on proactive oversight. Whereas P&C insurance has rules and tests, fair lending practices have been established and refined over many decades through the accumulation of case law.

NYC Hiring Algorithm

In 2021, the New York City Department of Consumer and Worker Protection implemented Local Law 144,[5] which requires an independent bias audit for automated employment decision tools. The bias audit calculates the selection rate for each race/ethnicity and sex category and compares the selection rates to the most selected category to determine an impact ratio. There is no threshold requirement; instead, the only requirement is to conduct the audit and share the results.

While the law became effective in July 2023, a study led by Cornell University found that only 18 out of 391 employers published hiring algorithm audit reports. The study concluded that this is because the law gives employers considerable discretion over compliance (Wright et al. 2024).

The regulation illustrates the trend of regulating model outcomes, and, in this instance, the regulation does not allow for the adjustment of any explanatory variables.

5. How to Achieve Fair Outcomes in Modeling

As insurance practices and regulations evolve, it is becoming increasingly important to identify and mitigate unfair discrimination across the entire algorithmic life cycle.

In this section, we will review the five steps of model development and discuss how fairness can be considered in each step. The options presented do not imply that these are recommendations or best practices. They are possible options; actuaries should seek guidance to determine whether they suit their organization's situation.

Broadly, the steps are as outlined in Table 2. Each of the steps is discussed in greater detail below.

Some actuaries may fear that unfair discrimination regulations will mean that rates no longer reflect the underlying loss costs. However, as the proposed options indicate, the model development and deployment process has many iterative steps. Many choices are made along the way, and many potential alternative models are created. These alternative models will fall along a spectrum of results when it comes to measuring unfair discrimination, and a reasonable solution may be among them.

Table 2. Five Steps of Model Development

Step	Substeps
1. **Model governance** (applies to all models, while steps 2-5, below, apply to a single model)	Model governance, which sets the stage for an organization's considerations of unfair discrimination, involves answering the following questions. a. **What is your guiding philosophy on unfair discrimination?** Is it to comply with regulation, or is there a more general approach that does not differ by jurisdiction? Do you have a mathematical approach to fairness, an outcomes-based approach, or both? b. **How do you plan to follow your philosophy?** What won't you do and how will you monitor compliance? c. **What models and data do you have now?** This is a good starting point to identify areas of potential unfair discrimination. d. **What models and data are you considering for the future?** Carriers should consider whether the model builders should be on higher or lower alert for potential unfair discrimination. e. **How will you monitor and correct for compliance?** What will you monitor, and if you do infer sensitive data such as race and ethnicity, who will have access to it?

(continued on next page)

Table 2. Five Steps of Model Development *(Continued)*

Step	Substeps
2. Project planning	Decisions at the project planning stage can have significant impacts on unfair discrimination. Questions at this stage include the following. a. **How is the problem formulated?** How a business problem is translated into an analytical problem can significantly affect unfair discrimination. b. **What unfair discrimination laws and regulations apply to this model?**
3. Data preparation and exploration	The data may lead to unfair discrimination because of (1) bias due to data completeness or diversity or (2) unfair discrimination in the underlying data. Carriers can seek to understand their data when viewed through a fairness lens. This process involves the following substeps. a. **Analyzing data sets for unfair discrimination.** b. **Improving data** if the testing shows areas of opportunity. c. **Taking data transformation considerations into account.** Insurance data is developed, capped, grouped, or otherwise transformed in some way. Carriers should be aware of the impact of these data adjustments on unfair discrimination. d. **Inferring race and ethnicity.** To test model outcomes for fairness, regulators such as the Colorado DOI are asking carriers to infer race using a method called Bayesian Improved First Name Surname Geocoding (BIFSG).
4. Modeling	In modeling, many choices are made, and – reflecting the idea of model multiplicity – carriers have several options to find a model that performs well and meets fairness criteria. Model builders can influence the fairness of the model in a variety of ways. a. **Choosing the model type.** All model types could lead to unfair discrimination; however, some types of models are more transparent than others, and this may help in fairness considerations. b. **Creating variables.** Carriers should consider fairness when they group their data and extrapolate their variables.

Table 2. Five Steps of Model Development *(Continued)*

Step	Substeps
	c. **Choosing variables.** Historically, regulators have tackled unfair discrimination by barring particular variables. d. **Penalizing the model for unfair discrimination.** During model training, carriers can set a loss function that can penalize the model for unfair discrimination. e. **Changing how the data set is partitioned into training and test data.** Partitioning the holdout test set and training set in different ways could result in differences in fairness. f. **Analyzing the accuracy of the model by groups of interest** to detect any bias in the algorithm.
5. Model implementation	A model itself is neither fair nor unfair, but how it is used and how it impacts people may result in unfair discrimination. Therefore, model implementation must be included in considerations on how to increase fairness. a. **Practical decisions**, such as determining what to do if production variables are missing, could impact fairness. b. **Tailoring model use**, such as putting a "human in the loop" for claims models, could also impact fairness. c. **Ongoing monitoring** of fairness metrics can take place in production. A model may be fair when implemented but drift over time as the carrier expands into new types of business or geographies.

1. Model Governance

Model governance[6] sets the stage for an organization's consideration of unfair discrimination. Establishing a broad philosophy and approach and cataloging the models and data the organization currently has is an essential starting point.

We present four questions for actuaries to consider as part of their model governance framework:

- What is your guiding philosophy on unfair discrimination?
- What models and data do you have now?
- What models and data are you considering for the future?
- How will you monitor and correct for compliance?

What is Your Guiding Philosophy on Unfair Discrimination?

The evolving regulations in Colorado recommend establishing a cross-functional team to provide governance on the issue of unfair discrimination. Questions to ask could include the following:

- What is your guiding philosophy on unfair discrimination? Is it to comply with regulation, or is there a more general approach that does not differ by jurisdiction? Do you have a mathematical approach to fairness, an outcomes-based approach, or both (see Box 1)?
- What dimensions are within scope? For example, will you consider race and ethnicity, religion, disability, etc.?
- What are the key areas of concern? For example, is it how much customers are being charged, who is being denied coverage, and/or how long it takes for customers to have their claims settled?
- What kinds of models and data are in scope? For example, will you consider machine learning models, generalized linear models (GLMs), and also traditional actuarial models?
- How will you determine whether differences in areas of concern are legitimate and fall within your guiding philosophy? For example, let's say that your rates for theft coverage are high in high-crime zip codes. However, high-crime zip codes correlate significantly with race. Is that OK? Regulators may opine on legitimate variables, but in the absence of such guidance, one framework outlined by O'Neil, Sargeant, and Appel (2024) for considering whether variables are legitimate is to perform a "balancing test," illustrated in Figure 1. You can consider where in the spectrum a variable falls, namely, (1) how strong is the predictive power, or how much it directly measures the thing you want to predict, and (2) how correlated the variable is to your group of interest.

Box 1. Do You Have a Mathematical Approach or an Outcomes-Based Approach?

An organization can take two approaches to unfair discrimination.

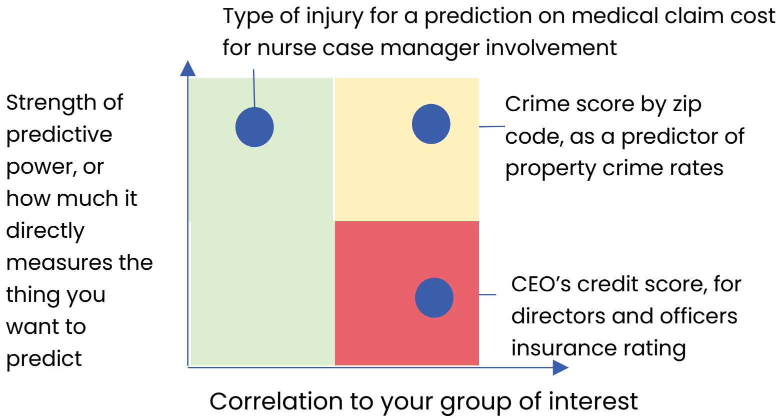
1. **A mathematical approach to fairness** looks at things like whether the rate aligns with loss cost, whether you have the same accuracy by groups of interest, and whether you have enough data by group to make good predictions.

   How important is it for your organization to be mathematically fair?

2. **An outcomes-based approach to fairness** looks at how the results of a decision may be viewed as fair or unfair from the perspective of those who are most impacted and have the least power. For instance, an insurer may think that charging more for insuring a property in a high-crime area is justified, as the higher loss costs make it fair. However, a customer may not be able to afford to move out of a high-crime area, and therefore, to them, paying more for insurance because they live in this area may seem unfair.

   Are such views important for your organization to consider?

Figure 1. Balancing Test


Type of injury for a prediction on medical claim cost for nurse case manager involvement
Strength of predictive power, or how much it directly measures the thing you want to predict
Crime score by zip code, as a predictor of property crime rates
CEO’s credit score, for directors and officers insurance rating
Correlation to your group of interest

How Do You Plan to Follow Your Philosophy?

Typically, in a modeling process, roles and responsibilities are operational details that are kept within the analytics team; however, on topics of unfair discrimination, obtaining broad feedback from the governance team on how the organization will operationalize its fairness approach can be helpful. Following are some examples of relevant questions to seek feedback on.

What are some things you will and will not do? In particular, testing for unfair discrimination may require the organization to collect or infer customer race. However, customer race is a sensitive data set for carriers. Options on how to treat this data include the following:

- Allowing anyone in the organization to have access to this data
- Limiting who has access to this data
- Having no access to this data (Either no testing is performed, or, if testing is performed, it is “blinded” by using a consultant or an in-house firewall.)

If a carrier chooses the last option, it will be harder to achieve fair outcomes and avoid unfair discrimination. This paper assumes that data on inferred race is available throughout the modeling process.

Indeed, the most efficient way to create a model that is both predictive and fair is if the modeler considers metrics on unfair discrimination all through the modeling process, in the same way they assess metrics on predictive power and stability.

How will you monitor and correct for compliance? Carriers need to determine how they will measure and monitor unfair discrimination. As discussed, regulation has been evolving; in the past it focused on model inputs, and now there is a focus on model outcomes. So, it is considered best practice to monitor both the inputs to and outcomes of models.

The current proposed tests for unfair discrimination on model outcomes are complex. Companies could consider more basic metrics to measure unfair discrimination that can be updated easily throughout the modeling project and after model implementation, reducing the need to rerun complex tests.

For example, a carrier could monitor the difference in rate per exposure for its groups of interest, all legitimate variables being equal.

What Models and Data Do You Have Now?

Having a clear view of all the models and data in use across the organization is a good starting point for identifying potential areas of unfair discrimination.

It is helpful to maintain a central list of the following in-use items:

1. Models built in-house
2. Models that have been licensed
3. Data that the organization has licensed

   This applies to all models and data within the organization, not just those built or purchased by the analytics team. So, if marketing buys a model for targeted advertising, or a particular part of the business purchases some data to help with underwriting, this should be centrally documented.

Creating model cards for models in production provides transparency by succinctly summarizing each model in an easily understood way. This is an excellent practice to increase the transparency of models for a cross-functional team. The model card includes a summary of the goal of the model, the input data, its predictive variables, its limitations, and its performance. Google originally proposed this, and examples and more information are available on the company's website (Google, n.d.). An example of a model card for a claims fraud model is provided in Box 2.

Box 2. Example Model Card

Claims Fraud Model

The model flags the top 1% of claims that are likely to result in a fraud investigation at 30 days from first notice of loss. The goal is to reduce the cost of fraud.

Input: Claims features (age, gender, zip code, type of accident, etc.), loss cost details, claims notes, and external data on claims fraud.

Output: A flag with a reason code or codes for claims professionals, who assess and refer to the special investigations unit (SIU), which then assesses and may open a case.

Limitations: Our model is trained on whether a claim results in an opened fraud case. This has several limitations, including the following:

- Not all fraud cases turn out to be truly fraudulent.
- There will be cases of fraud that go undetected. To the extent that this is systematic, our model will also be unable to detect those cases.
- To the extent that our past practices had bias, then the model will exhibit that bias.
- Etc.

Performance: 1% of all claims result in an SIU fraud investigation. At 30 days, the claims flagged have an 8% chance of resulting in an SIU fraud investigation. . . .

What Models and Data are You Considering for the Future?

When choosing new models to build, the organization should consider whether the modelers should be on higher or lower alert for potential unfair discrimination. For example, companies may feel that creating a model to alert for claims fraud has a higher risk of unfair discrimination than building a model that identifies the construction type of a building from an image.

2. Project Planning

Decisions at the project planning stage can have significant impacts on unfair discrimination. Following are some important considerations.

How is the Problem Formulated?

How a business problem is translated into an analytical problem can significantly affect the potential for unfair discrimination and disparate impact. Elzayn et al. (2023) researched Internal Revenue Service models created to select targets for audits. Disparate impact was reduced by changing from a model that predicted compliance or noncompliance (a binary outcome) to a model that predicted the amount of money people failed to report.

For rating models, the analytical problem has traditionally been to predict the insured's loss cost over a single policy period. There may not be much room for an alternative problem formulation for these types of models.

However, claims and marketing models have significantly more space for alternative considerations. For example, a claims fraud model may currently be trained on a binary outcome of fraud or no fraud, where instead it could be trained to predict the cost of fraud. Changes like this have the potential to reduce unfair discrimination.

What Unfair Discrimination Laws and Regulations Apply to this Model?

Carriers need to consider what unfair discrimination laws and regulations apply to their models. Teams also need to take care when building models for one purpose or jurisdiction that are then repurposed for another use case or jurisdiction.

3. Data Preparation and Exploration

Evaluating a data set for its potential to result in unfair discrimination is a practical step that may save time later in the model development process (see next section, "4. Modeling"). Several possible options to improve the underlying data set exist, depending on the particular challenges and specifics of the project.

Before preparing, testing, improving, and/or transforming the data, actuaries should review Actuarial Standard of Practice 23: Data Quality (Actuarial Standards Board 2016). This standard provides data definitions, analyses, communications and disclosures, and recommended practices.

Analyzing Data Sets for Unfair Discrimination

The first step in analyzing data sets is understanding potential pitfalls and areas where the data set may be unfairly discriminatory. As mentioned earlier, this necessitates handling data on inferred race and ethnicity.

Data quantity: There may be insufficient data for an algorithm to make inferences about less well-represented groups. (see Box 3 for an example). Running a basic set of checks to identify the raw number of records by protected class, percentage of total records by class, and percentage by class and groups of interest can identify whether there will be credibility issues due to lack of data. If credibility issues are identified, teams can discuss internally whether to (a) try to supplement the data, (b) change the analytical approach, and/or (c) change how the algorithm results are implemented

Data quality: Having significantly lower-quality data – as in incomplete or potentially inaccurate entries – for certain groups (see example in Box 4) can lead to inaccurate conclusions about those groups. Running a set of data quality checks by groups of interest can identify whether the data is of different quality by class. Recommended checks to run include determining the percentage of duplicate records, incomplete records, nulls or missing values, high or low outliers, impossible values (e.g., zip code 99999 or model year 1900), and recency/staleness of records (e.g., percentage of records by month/year) by protected class to identify whether there are more significant quality issues with some groups than others.

Box 3. Data Quantity Example for Gender Nonbinary Customers

A carrier may write personal auto insurance in states that require rates for customers whose gender identity is nonbinary (not specifically male only or female only). However, if the company lacks credible data on nonbinary risks, how will it create rates for these customers? These rates could be a weighted average of the male and female rates, the maximum, the minimum, or something else. These resulting outcomes could differ substantially, and decisions should be made thoughtfully, keeping in mind actuarial professional standards, the company's philosophy with regard to unfair discrimination, and the governing laws and regulations.

Box 4. Data Quality Example for Missing Values

If a check on missing values by inferred race reveals significantly fewer missing values for "last annual mileage" for White customers than for Black customers, the manner of dealing with those missing values may impact the overall assessment of the risk of those Black customers. If the decision is to be conservative and impute a high value (e.g., the 75th or 90th percentile) instead of the average or median value, the comparative risk of Black customers may be artificially inflated.

Underlying bias or unacceptable differences in the data: The collection or codification of data may be biased (Box 5) or have differences that are considered unacceptable by regulators. This bias could be in the independent variable data used to assess risk or in the target data itself. Understanding data sources, collection and codification practices, and the data quality standards applicable to that collection and codification can help point out variables that might warrant further investigation, which can then be run through the tests laid out below (in the discussion of early indicators of unfair discrimination) to assess their potential risk.

Box 5. Underlying Bias in Data Example for Medical Costs

A 2019 study by Obermeyer et al. showed that an algorithm used by healthcare providers to assess medical risk and suggest treatment underestimated both the need for treatment and the expected costs for Black patients compared to white patients with the same health and comorbidity profiles. This is because the model relied on healthcare cost data as a proxy for treatment need, and Black patients have lower healthcare costs for the same health profiles.

While the study focused on the healthcare space, this is also a relevant consideration for P&C insurers. Medical costs are part of loss costs in many P&C pricing and claims models.

<u>Early indicators of unfair discrimination in the data:</u> Calculating the target variable by protected class will identify the raw differences in the data set. Statistical tests such as *t*-tests, chi-square tests, or analysis of variance will also identify statistical differences between and among groups. These analyses can determine whether the differences in the target variable are what is expected or allowed. Depending on the specific regulations of a given state, finding differences between groups of interest in a simple univariate analysis does not necessarily indicate that an algorithm built on that data set will fail – as, for example, rates by groups of interest may be allowed to differ as long as those differences are driven by other, legitimate variables (such as age, gender, and tobacco use for the Colorado life insurance draft guidelines). However, the presence of these univariate differences can indicate that additional awareness and analysis are needed.

Analytics teams can take this approach one step further and modify it to the specific regulation they want to comply with. For example, take Colorado's draft testing guidelines. For personal auto, the DOI will likely specify what variables are "allowable." Segmenting the data by these allowable variables and reexamining the target variable will give a better sense of whether the spread is unacceptable. For example, suppose gender, age, and prior claims are specified as allowable variables. In that case, looking at the loss costs by race for groups of customers of the same gender, age, and prior claims will further highlight whether the model is likely to pass or fail the test.

Additionally, studying the correlations between the protected class and other variables can reveal which variables may be more or less likely to impact the group of interest in the data.

Suppose the analytics team sees differences in rates by a particular protected class. In that case, the team can further investigate how that class differs from the rest of the population using some of the tests described below. These investigations should provide a path forward in developing a model that avoids unfair discrimination.

A list of higher-risk data: Having a checklist of higher-risk data types is helpful in ensuring that practitioners are aware that these types of data may be more likely to lead to unfair discrimination. The following checklist has been compiled by reviewing research papers, such as Members of the 2021 CAS Race and Insurance Research Task Force (2022) and statements from regulators on the types of data that may lead to unfair discrimination:

- Medical cost data
- Law enforcement data (for example, motor vehicle records)
- Social media
- Location tracking
- Retail purchase history
- Internet usage
- Mobile activity
- Biometric data, including facial recognition
- Geographical data
- Homeownership data
- Credit information
- Education level
- Civil judgments
- Court records
- Condition or type of an applicant's electronic devices

The fact that a data element is on this list does not mean that it cannot or should not be used in a model, but it may warrant greater internal validation for potential unfair discrimination or disparate impact than some other factors.

Improving Data

If these initial tests indicate that the data may have some areas of concern, several options exist to improve it.

Obtain additional data: If the quantity of data on protected classes is insufficient, the team can consider expanding the data set by

- using more of the organization's own data and increasing the scope (years, geographies, etc.),
- using submission data (for example, in commercial lines, credible exposure and prior loss data could supplement the data from the bound book of business), and/or
- purchasing additional data.

Adding data, where possible and practical, can avoid some of the pitfalls of other techniques. However, purchasing data is not always possible or practical. Expanding the scope to include additional internal data points may introduce inaccuracies, such as older data having less relevance or submission data being incomplete or inaccurate.

Improve data quality: For data with quality issues, analyzing different methods for improvement and implementing methods that reduce that discrimination can avoid the potential exacerbating effect of data quality issues.

There are three ways to tackle data quality issues:

- Short term: changing the rules for data quality adjustments, such as changing the imputation method or rules for when data is discarded
- Medium term: building models to impute data or discard it when missing
- Longer term: working to improve the data collection systems and train the users to improve data at the source

Relabel the data: To remove significant differences in outcomes in the data, the data itself – and, specifically, the target variable – can be changed.

For example, imagine that a carrier declined all applicants who did not provide their last annual mileage to manage and reduce uncertainty in risk assessment. The company finds that this has caused different acceptance rates by inferred race and ethnicity. Going forward, it removes this trigger for declination. In this case, if the carrier builds a model with declinations as its target, it could see the impact of restating the data so that those who would no longer be declined are marked as “accepted.”

If this relabeling can be done in a clear, quantifiable, and systematic way, it is one of the simplest ways to obtain unbiased data and may alleviate the need for many of the techniques discussed for adjusting modeling and implementation.

However, changing the target variable for a model can be risky and has traditionally been an acceptable part of model building in only a limited way. Without clear guidelines on how and to what extent the target can be changed, relabeling the data can create less accurate models and even introduce additional bias. Such an approach may be more suitable for allowing the analytics team to review scenarios and develop hypotheses as to the source of data deficiencies.

Taking Data Transformation Considerations into Account

In addition to cleaning data, considerable time is spent transforming the data for subsequent modeling. Data is grouped, capped, or otherwise transformed in some way. Actuaries should be aware of the impact of these data adjustments on unfair discrimination. In actuarial rating models, these adjustments may include the following:

- Loss development: How losses are developed to their ultimate value can significantly impact the final model. The team can look at the severity of closed losses by protected class and the paid and incurred loss at various development periods to see if there are differences in development. Unfair discrimination could exist if meaningful differences in the type of loss exist.
- Capping losses: Capping losses may impact some protected classes more than others. For example, perhaps after looking at capped and uncapped losses by the group of interest, the actuaries find that one group has a higher severity before capping than others, but this is significantly muted after capping. This may inform the team's decision on how to cap losses.

Inferring Race and Ethnicity

To test models on fairness of outcomes, regulators such as the Colorado DOI have suggested that carriers infer the race of their customers using Bayesian Improved First Name Surname Geocoding (BIFSG). First developed by the RAND Corporation, this method uses a customer's first name, surname, and address, as well as census data, to provide a probability that the customer belongs to a particular race/ethnicity. Statistical Methods for Imputing Race and Ethnicity by Baeder, et al (2024), provides a history of BIFSG, its predecessors and variations and their relative levels of accuracy.

This methodology is widely used. For example, the Consumer Financial Protection Bureau relies on Bayesian Improved Surname Geocoding (BISG) to conduct fair lending analysis of providers of non-mortgage credit products, such as auto lenders, who are not allowed to collect consumers' demographic information. More details on the bureau's approach can be found in its 2014 paper (CFPB 2014).

The D.C. DISB study *Report on Market Conduct Examination: Evaluating Unintentional Bias in Private Passenger Automobile Insurance* (D.C. DISB 2024) used BIFSG and, more specifically, the algorithm detailed in Voicu (2018). More details on DISB's use of BIFSG can be found in Appendix A of the department's draft report.

BIFSG provides a probability that a customer is of a particular race, and in the DISB study, the customer is assigned the race with the highest probability. There are other approaches; for example, race may be assigned only if a certain threshold is met, and race is otherwise unknown, or the full vector of probabilities may be used, so that when aggregating results by race, the vector can be used as weights.

Depending on the organization's approach to inferring race, the fairness testing will show differing results. For more details on the methodology, including its accuracy, see the 2023 summary paper from the 2023 BISG conference (Appel et al. 2023).

4. Modeling

Modeling is an iterative process, with many decisions made at each juncture. There are many possible reasonable models, and carriers have several options for finding a model that will meet fairness criteria. Indeed, the concept of model multiplicity holds that for any given problem, several predictive models with the same level of accuracy can be found (Box 6).

This occurs because, for a given error rate, there are different ways to distribute this error over a population. A hypothetical example is provided in Black, Raghavan, and Barocas (2022). They consider two models predicting outcomes for a data set with equal proportions of men and women. While both models have 80% accuracy, one is 80% accurate for both men and women, while the other is 100% accurate for men and 60% accurate for women.

Box 6. Example of Model Multiplicity

In one research setting, Coston, Rambachan, and Chouldechova (2021) tested accuracy and disparate impact over a range of models aimed at criminal risk assessment. They show that alternative models exist to the one deployed that have equivalent accuracy (within 1%) yet have a more than 10% lower selection rate disparity across racial groups.

This is encouraging and shows that, among all the reasonable options that modelers currently consider, there may be several that meet fairness criteria as well as accuracy criteria and other business concerns. Below we detail some key decisions model builders can make with a focus on fairness.

Choosing the Model Type

Will the organization use a simple linear model or a more complex machine learning model? Usually, this decision is based on considerations other than fairness. However, fairness can also be considered, particularly in higher-risk applications. For example, for a claims model, machine learning could be used. However, a GLM might provide increased transparency without a significant loss in accuracy. Because GLMs are more transparent, identifying and correcting the inequity will be easier if unfairness is in fact detected.

Creation of Variables

In creating variables, actuaries should consider fairness when grouping and extrapolating data for their models

Grouping data. How do the protected classes fall by group? For example, model builders may use their judgment to group zip codes into territories. They can study the territories by inferred race and ethnicity to see whether there are alternative territorial groupings that achieve a similar goal while avoiding segmentation along racial lines.

Extrapolation. Model builders may have very little data on the edges of their variables. For example, for younger or older drivers, they may have little data and need to make assumptions about their performance based on the performance of drivers at other ages. How do these choices impact each group of interest? Instead of assuming that variables behave in a particular way, model builders should be thoughtful in reviewing their data, identifying what other information they can benchmark to, and considering how their assumptions impact the groups of interest.

Choice of Variables

Traditionally, actuaries have complied with regulations prohibiting certain variables or proxies for those variables in certain jurisdictions.

Remove variables. Model builders can consider removing variables most highly correlated with a group of interest. It may also be important to understand how removing a variable could impact other outcomes, such as the potential for adverse selection and competitive advantage.

Replacing variables. Actuaries can consider other variables that may improve the predictive power of their models. For example, instead of using a variable closely correlated to a group of interest, they could consider an alternative variable that may deliver the same predictive power.

Penalizing the Model for Unfair Discrimination

During model training, analytics teams can set a loss function that can penalize the model for differences in the distribution of outcomes for protected class applicants relative to control class applicants.

Adding penalties is a general technique that can be applied to many methods, including machine learning models. Instead of transforming the data, the machine learning models and their learning objectives can be modified. Most models are trained to optimize some measure of accuracy, but they can also be adjusted to consider fairness. This can be done by changing the loss function to consider fairness or by imposing constraints on model predictions. An example of this approach in the area of fair lending can be found in Zhang, Lemoine, and Mitchell (2018).

For GLMs, one approach would be to add a penalty parameter to the loss function. This works similarly to regularization, where parameter values are penalized to reduce complexity and overfitting. However, here a penalty parameter is introduced to reduce unfairness. This parameter would need to satisfy a specific mathematical definition of fairness. For example, it could be forced to have equal true positive rates for both protected and other classes. In researching how this approach could be practically applied, we concluded that further study is needed.

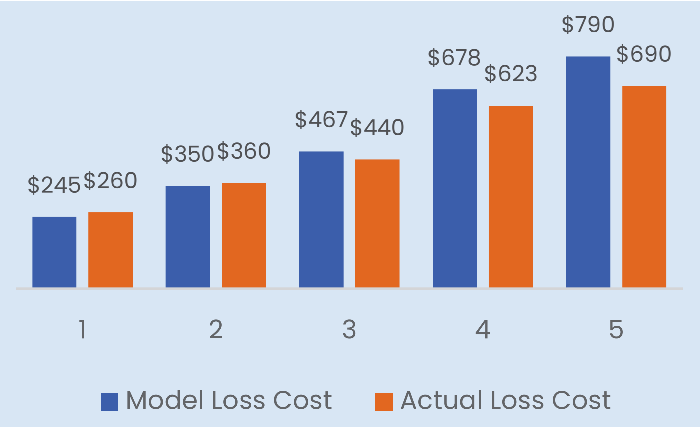
Changing How the Data Set is Partitioned into Training and Test Data

With all models, the builders decide how to partition the holdout test set and training set. Their choices can also result in significant differences in fairness. In particular, if an organization's book of business is changing significantly, and the carrier only recently started writing customers in a particular group of interest, out-of-time holdout testing[7] may show significant issues in performance for that class. In this case, the analytics team can also try a more random holdout test set.

Analyzing the Accuracy of the Model by Groups of Interest

To detect any bias in their algorithms, actuaries can view typical model accuracy metrics such as *p*-values, *R*-squared values, mean squared error, deviance, and lift charts of the model performance by groups of interest (see Box 7 for an example). These metrics may highlight opportunities to improve the model's fairness by improving the model's accuracy by protected class.

Box 7. Example of a Lift Chart



Above is a mock lift chart showing the accuracy performance of a model that predicts the loss cost per vehicle for a particular group of interest.

The chart shows five equal buckets of exposures, with the model loss cost (shown in blue) ranging from lowest in the first bucket to highest in the fifth bucket. The actual loss costs are also shown, in orange. Comparing the actual losses to those estimated by the model shows that those exposures expected to have the highest loss costs were overestimated. The actual loss costs were lower than expected. The model builders can revisit the model to understand and, if needed, remedy this result.

5. Model Implementation

A model itself is neither fair nor unfair, but how it is used and how it impacts people may result in unfair discrimination. Therefore, model implementation must be included in considerations on how to increase fairness.

Practical Decisions

When a model is implemented, many additional practical decisions could cause unfair discrimination. Some examples include the following:

- Capping the discount and surcharge in a model, which may be done to reduce volatility or comply with regulation. Modelers can analyze the impact this has on groups of interest.
- Setting minimum premiums.
- Decisions on what to do if a particular variable is missing in production.

Tailoring Model Use

Claims and marketing models can be used in many ways, and specific decisions on how to use a given model can significantly impact unfair discrimination. For example, when a customer is flagged for potential fraud, are they immediately routed to a process that requires them to provide additional documentation, or must someone first verify the information that led to the trigger and agree that there is likely fraud? This is called "putting a human into the loop."

Stopping Use of the Model

If a model is problematic, carriers should consider pausing its use while evaluating how it could be improved or replaced with a less problematic alternative.

Ongoing Monitoring

Analytics teams should continue to monitor their models for unfair discrimination after they are deployed and review them periodically. Shifting demographics over time in the regions in which the carrier operates can result in changes in the outcomes of fairness tests. This type of ongoing monitoring is also important if the organization is expanding into new geographies or types of business.

6. Conclusion

The regulations relating to unfair discrimination are changing, and are likely to vary across jurisdictions, which may leave carriers uncertain about how to proceed. Analytical professionals such as actuaries cannot tackle this issue alone. Collaboration across the organization is necessary to address the problem effectively. By taking careful and calculated steps, remaining open to learning, and adapting as needed, actuaries can achieve success in this new paradigm.

7. Appendix: Colorado's Life Insurance Underwriting Regulation Detail

Colorado's draft regulation, Quantitative Testing for Unfairly Discriminatory Outcomes for Algorithms and Predictive Models Used for Life Insurance Underwriting, was released by the Department of Insurance (DOI) on September 28, 2023. It mandates that insurers infer the race and ethnicity of life insurance applicants and test algorithms and models that use external consumer data and information sources (ECDIS) for potential unfair discrimination. This draft regulation requires that model outcomes be tested annually for unfair discrimination and mandates corporate governance via a risk management framework. This testing of model outcomes is new for insurers, as the historical regulatory focus has been on inputs and justification of rates based on loss costs. The draft requirements for the risk management framework include detailed inventories and descriptions of ECDIS, algorithms, and models, as well as the establishment of cross-functional governance committees.

The draft requires two tests. The first assesses whether application approval rates differ significantly by race and ethnicity (in other words, are there statistically significant approval-rate differences for Black, Hispanic, and Asian or Pacific Islander applicants compared to white applicants?). The second considers whether premium rates differ significantly by race and ethnicity.

Both tests require data on a customer's race and ethnicity. The draft requires this to be estimated using Bayesian Improved First Name Surname Geocoding (BIFSG), a statistical method developed by the RAND Corporation. This approach uses a customer's first name, surname, and address to estimate their race and ethnicity using census data.

To assess whether there are statistically significant differences in approval rates by race, the company is required to build a logistic regression model. The response variable is approval/disapproval, and only five specific predictors are allowed (policy type, face amount, age, gender, and tobacco use). Race and ethnicity form the sixth dummy variable.

The test is whether there is a statistically significant difference in approval rates by race, that is:

1. If the dummy race and ethnicity variable is significant, i.e., has a *p*-value of less than 0.05
2. If the difference in approval rates between white applicants and applicants in each of the other race and ethnicity groups is 5 percentage points or greater

If the company's model outcome for approval rates by race fails both tests, then a third is run. For this test, the company must build two more logistic regression models. In the first, all variables must be incorporated, including ECDIS and traditional underwriting factors. The second model needs to include a dummy variable representing the estimated race and ethnicity in conjunction with the rating variables from the first model. The test is then:

3. If there is any difference in the coefficients in the ECDIS between the two models, one with race and ethnicity and one without

If the model outcome fails this test, the variable and the model are deemed unfairly discriminatory.

The test for statistically significant differences in premium rates by race and ethnicity is similar to the test for differences in acceptance rates by race and ethnicity.

The Colorado DOI has been working with stakeholders to improve and finalize the life insurance regulation draft. Feedback from stakeholders led the DOI to consider a few key points regarding the proposed testing regulation: (1) the DOI is examining whether a 5% difference threshold for application approvals and premium rates can effectively flag unfair discrimination; (2) it is assessing the practicality of conducting unfair discrimination tests on premiums, due to the quantity and complexity of factors involved in computing premiums; and (3) it is discussing the possibility of establishing a safe harbor to protect insurers from liability when they genuinely try to follow the regulations.

While the current draft regulation for testing for unfair discrimination applies only to life insurers, the Colorado DOI is expected to propose a similar draft regulation for property and casualty insurers, starting with private passenger auto as the first line to be subjected to such regulation. If the draft regulation is enacted, carriers would need to perform annual quantitative assessments of their data and models that use ECDIS to detect potential unfair discrimination.

There are still two important unknowns in Colorado's testing regulation:

1. What will the allowable specific predictors be for the first two tests? For example, will territory be an allowable predictor?
2. For the third test, how much of a difference in ECDIS coefficients is enough to deem it unfairly discriminatory? Simply introducing any additional factor into a model will likely change the coefficients of all the other factors.

We trust that Colorado will continue to provide clarity on the testing framework.

The American Academy of Actuaries and the American Property Casualty Insurance Association raised concerns about the suitability of using the life insurance testing draft regulation for personal auto insurance. They pointed out several ways in which personal auto insurance differs significantly from life insurance.

First, personal auto insurers use numerous variables traditionally considered ECDIS in life insurance, such as previous accidents, traffic violations, and credit-based insurance scores. These factors have long been accepted as nondiscriminatory in personal auto rating and underwriting. Second, personal auto insurance encompasses a broader range of coverages than life insurance. These include bodily injury liability, property damage liability, collision, medical payments, and other coverages, while life insurance is solely focused on a single coverage.

Additionally, other commenters have highlighted that the rate filing and approval process for personal auto insurance is notably slower than that for life insurance, which would potentially result in further delays in the testing process.

8. References

Actuarial Standards Board. 2016. "Actuarial Standard of Practice 23: Data Quality." American Academy of Actuaries. Last modified December 2016. http://www.actuarialstandardsboard.org/asops/data-quality/.

Actuarial Standards Board. 2019. "Actuarial Standard of Practice 56: Modeling." American Academy of Actuaries. Last modified December 2019. http://www.actuarialstandardsboard.org/asops/modeling-3/.

Appel, Jacob, Bennett Borden, Cathy O'Neil, Dan Svirksy, and Sam Tyer-Monroe. 2023. *Promises and Limits of Inferring Protected-Class Data for Disparate Impact Testing of AI Systems: Conference Report*. New York: O'Neil Risk Consulting and Algorithmic Auditing. https://orcaarisk.com/in-the-news/2023/12/20/bisg-conference-summary-paper-now-available.

Black, Emily, Manish Raghavan, and Solon Barocas. 2022. "Model Multiplicity: Opportunities, Concerns, and Solutions." In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–63. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533149.

CDI (California Department of Insurance). 2022. "Allegations of Racial Bias and Unfair Discrimination in Marketing, Rating, Underwriting, and Claims Practices by the Insurance Industry." Bulletin 2022-5. June 30, 2022. Sacramento: CDI. https://www.insurance.ca.gov/0250-insurers/0300-insurers/0200-bulletins/bulletin-notices-commiss-opinion/upload/BULLETIN-2022-5-Allegations-of-Racial-Bias-and-Unfair-Discrimination-in-Marketing-Rating-Underwriting-and-Claims-Practices-by-the-Insurance-Industry.pdf.

CFPB (Consumer Financial Protection Bureau). 2014. *Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity*. Washington, DC: CFPB. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

CID (Connecticut Insurance Department). 2022. "Notice to All Entities and Persons Licensed by the Connecticut Insurance Department Concerning the Usage of Big Data and Avoidance of Discriminatory Practices." April 20, 2022. Hartford: CID. https://portal.ct.gov/-/media/cid/1_notices/technologie-and-big-data-use-notice.pdf.

Coston, Amanda, Ashesh Rambachan, and Alexandra Chouldechova. 2021. "Characterizing Fairness over the Set of Good Models under Selective Labels." *Proceedings of the 38th International Conference on Machine Learning* 139: 2144–5. https://proceedings.mlr.press/v139/coston21a.html.

D.C. DISB (District of Columbia Department of Insurance, Securities and Banking). 2023. "Data Call Q&A." Updated April 12, 2023. https://disb.dc.gov/page/data-call-qa.

D.C. DISB. 2024. *Report on Market Conduct Examination: Evaluating Unintentional Bias in Private Passenger Automobile Insurance.* Washington, D.C.: Government of the District of Columbia. https://disb.dc.gov/sites/default/files/dc/sites/disb/page_content/attachments/Unintentional%20Bias%20report%20-%20v.2%20draft.pdf.

Elzayn, Hadi, Evelyn Smith, Thomas Hertz, Arun Ramesh, Robin Fisher, Daniel E. Ho, and Jacob Goldin. 2023. "Measuring and Mitigating Racial Disparities in Tax Audits." Working paper. Stanford, CA: Stanford Institute for Economic Policy Research. https://siepr.stanford.edu/publications/working-paper/measuring-and-mitigating-racial-disparities-tax-audits.

ColFDIC (Federal Deposit Insurance Corporation). 2021. "Fair Lending Laws and Regulations." In *FDIC Consumer Compliance Examination Manual*. Arlington, VA: FDIC. https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/4/iv-1-1.pdf.

FSRAO (Financial Services Regulatory Authority of Ontario). 2022. *Operational Risk Management Framework in Rating and Underwriting of Automobile Insurance*. Information Guidance No. AU0137INF. Toronto: FSRAO. https://www.fsrao.ca/industry/auto-insurance/regulatory-framework/guidance-auto-insurance/operational-risk-management-framework-rating-and-underwriting-automobile-insurance.

Google. n.d. “Model Cards: The Value of a Shared Understanding of AI Models.” Accessed July 22, 2024. https://modelcards.withgoogle.com/about.

NYDFS (New York Department of Financial Services). 2024. “Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing.” Proposed Insurance Circular Letter. Published January 17, 2024. https://www.dfs.ny.gov/industry_guidance/circular_letters/cl2024_nn_proposed#_ftn2.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations.” *Science* 366 (6464): 447–53. https://www.science.org/doi/10.1126/science.aax2342.

O’Neil, Cathy, Holli Sargeant, and Jacob Appel. 2024. “Explainable Fairness in Regulatory Algorithmic Auditing.” *West Virginia Law Review* (forthcoming). Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4598305.

Voicu, I. 2018. “Using First Name Information to Improve Race and Ethnicity Classification.” *Statistics and Public Policy* 5 (1): 1–13. https://doi.org/10.1080/2330443X.2018.1427012.

White House. 2022. “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People.” Accessed July 22, 2024. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

White House. 2023. “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” Published October 30, 3023. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

Wright, Lucas, Roxana Muenster, Briana Vecchione, Tianyao Qu, Senhuang (Pika) Cai, Alan Smith, Jacob Metcalf, and J. Nathan Matias. 2024. “Null Compliance: NYC Local Law 144 and the Challenges of Algorithm Accountability.” Available on OSF: https://osf.io/upfdk/.

Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. *Mitigating Unwanted Biases with Adversarial Learning*. Washington, DC: Association for the Advancement of Artificial Intelligence. https://arxiv.org/pdf/1801.07593.

9. Additional Resources

Australian Human Rights Commission. 2022. *Guidance Resource: Artificial Intelligence and Discrimination in Insurance Pricing and Underwriting*. Sydney: Australian Human Rights Commission. https://humanrights.gov.au/sites/default/files/document/publication/ai_guidance_resource_december_2022.pdf.

Baeder, Larry, Erica Baerd, Peggy Brinkmann, et. Al. 2024. "Statistical Methods for Imputing Race and Ethnicity." https://www.soa.org/498d64/globalassets/assets/files/resources/research-report/2024/stat-methods-imputing-race-ethnicity.pdf.

Black, Emily, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. 2023. "Less Discriminatory Algorithms." *Georgetown Law Journal* 113 (1). SSRN: http://dx.doi.org/10.2139/ssrn.4590481.

Cavanaugh, Lauren, on behalf of American Academy of Actuaries Casualty Practice. 2023. Letter to District of Columbia Department of Insurance, Securities and Banking Associate Commissioner Philip Barlow on Draft Data Call on Unintentional Bias in Automobile Insurance. https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/comment_letter_to_disb_unintentiontal_bias.pdf.

Chibanda, Kudakwashe F. *Defining Discrimination in Insurance*. Arlington, VA: Casualty Actuarial Society. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Defining_Discrimination_In_Insurance.pdf.

Chiu, Keli. 2020. "Discrimination in Ontario Automobile Insurance." Medium. https://medium.com/swlh/discrimination-in-ontario-automobile-insurance-d68e6013c798.

Consumer Federation of America. 2023. Letter to District of Columbia Department of Insurance, Securities and Banking Associate Commissioner Philip Barlow on Draft Data Call on Unintentional Bias in Automobile Insurance. https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/unintentional_bias_data-call_cfa_comments.pdf.

D.C. DISB. 2023a. "DISB Data Call: Qualitative Questions." https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/disb_data_call_questions.pdf.

D.C. DISB. 2023b. "District of Columbia Request for Data – Private Passenger Auto." Data call template. https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/disb_data_call_template.xlsx.

D.C. DISB. 2023c. "Examination Warrant." https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/warrant-market-conduct-private-passenger_auto.pdf.

Deloitte Center for Regulatory Strategy. 2023. *Artificial Intelligence (AI) State of Play in Insurance Regulation: Developments as of March 2023*. New York: Deloitte. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/Advisory/us-advisory-deloitte-ai-state-of-play-in-insurance-regulation-march-2023.pdf.

Egan, Nancy J., David F Snyder, and Robert C. Passmore, on behalf of American Property Casualty Insurance Association. 2023. Letter to District of Columbia Department of Insurance, Securities and Banking Associate Commissioner Philip Barlow on Draft Data Call Related to Unintentional Bias in Personal Automobile Insurance. https://disb.dc.gov/sites/default/files/dc/sites/disb/publication/attachments/apcia_comments_dc_draft-data-call_final_01202023x.pdf.

Members of the 2021 CAS Race and Insurance Research Task Force. 2022a. *Approaches to Address Racial Bias in Financial Services: Lessons for the Insurance Industry*. Arlington, VA: Casualty Actuarial Society. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Approaches-to-Address-Racial-Bias_0.pdf.

Members of the 2021 CAS Race and Insurance Research Task Force. 2022b. *Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored*. Arlington, VA: Casualty Actuarial Society. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Understanding_Potential_Influences.pdf?utm_source=Website&utm_medium=Press+Release&utm_campaign=RIP+Series.

Mosley, Roosevelt, and Radost Wenman. 2022. *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance*. Arlington, VA: Casualty Actuarial Society. https://www.casact.org/sites/default/files/2022-03/Research-Paper_Methods-for-Quantifying-Discriminatory-Effects.pdf.

Pederson, Kirsten, on behalf of American Academy of Actuaries Life Practice Council and Casualty Practice Council. 2023. Letter to Colorado Division of Insurance Commissioner Michael Conway on Colorado draft regulation, Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes. https://www.actuary.org/sites/default/files/2023-10/life-letter-colorado-draft-regulation.pdf.

Policy and Education Branch, Ontario Human Rights Commission. 1999. “Discussion Paper: Human Rights Issues in Insurance.” Toronto: Ontario Human Rights Commission. https://www3.ohrc.on.ca/sites/default/files/attachments/Discussion_paper%3A_Human_rights_issues_in_insurance.pdf?_gl=1*t0b9f1*_ga*MTgyMTgwNzI3Mi4xNzE4MDM4NjU2*_ga_K3JBNZ5N4P*MTcxODAzODY1Ni4xLjEuMTcxODAzOTA4Mi4wLjAuMA..&_ga=2.210817232.475831965.1718039082-1821807272.1718038656.

Wang, Gary C. 2023. “Predictive Data Modeling: Understanding the Value and Limitations of Imputing Race and Ethnicity.” Pinnacle Actuarial Resources. https://www.pinnacleactuaries.com/article/predictive-data-modeling-understanding-value-and-limitations-imputing-race-and-ethnicity.

Wenman, Radost Roumenova. 2023. “Colorado’s Division of Insurance Proposes a Draft for Testing the Fairness of Algorithms and Predictive Models in Life Insurance.” Pinnacle Actuarial Resources. https://www.pinnacleactuaries.com/article/colorados-division-insurance-proposes-draft-testing-fairness-algorithms-and-predictive.

10. Endnotes

1. Concerning Protecting Consumers from Unfair Discrimination in Insurance Practices, S.B. 21-169 (2021), https://leg.colorado.gov/sites/default/files/2021a_169_signed.pdf.
2. Concerning Quantitative Testing of External Consumer Data and Information Sources, Algorithms, and Predictive Models Used for Life Insurance Underwriting for Unfairly Discriminatory Outcomes, 3 C.C.R. 702-10 (2023) (draft proposed new regulation), https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view.
3. At the time of writing, the most recent updated draft of the EU AI Act is Regulation (EU) 2024/. . . of the European Parliament and of the Council of laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.
4. Fair Housing Act, 42 U.S.C. §§ 3601-3619, 3631 (1968), https://www.justice.gov/crt/fair-housing-act-1#:~:text=The%20Fair%20Housing%20Act%20prohibits%20discrimination%20on%20the%20basis%20of,or%20more%20major%20life%20activities.
5. Automated Employment Decision Tools, New York City Local Law 2021/144, https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=.
6. The definition of “governance and controls” from Actuarial Standard of Practice 56: Modeling is “the application of a set of procedures and an organizational structure designed to reduce the risk that the model output is not reliably calculated or not utilized as intended” (Actuarial Standards Board 2019).
7. Out-of-time holdout testing is when the training data is from a set period of time, for example, 2018-2021, and it is tested on a holdout data set of data from a different period, for example, 2022.