AMERICAN ACADEMY *of* ACTUARIES

Additional Considerations in Data Science

Discrimination: Considerations for Machine Learning, AI Models, and Underlying Data

AUGUST 2023

Key Points

- Unfair discrimination takes place when insurers consider factors that are unrelated to actuarial risk while determining whether to provide insurance to particular individuals or groups, and if so, at what price and with what terms.
- Insurance legislation has put in place measures to prevent unfair discrimination while still permitting actuarially justified risk selection. However, within insurance companies, various functions like marketing, rating, and underwriting have become more reliant on big data, algorithms, and machine learning. These processes might utilize variables that appear neutral on the surface but can lead to unequal impacts on different groups of people.
- Discrimination can originate from multiple sources, including the data, the algorithm, and the overall models used in these practices.

AMERICAN ACADEMY *of* ACTUARIES

1850 M Street NW, Suite 300
Washington, DC 20036
202-223-8196 | www.actuary.org

© 2023 American Academy of Actuaries. All rights reserved.

This issue brief explores the topic of discrimination in machine learning algorithms and artificial intelligence (AI) algorithms, and the underlying data of these models. It will define discrimination (including distinguishing between discrimination, unfair discrimination, and unjust discrimination); present practical methods for testing and monitoring algorithms; provide a regulatory overview of the issue; and identify considerations for actuaries, algorithm creators, and regulators.

The following topics are discussed in the issue brief:

I. Defining discrimination
Includes a high-level discussion of issues around unlawful, unfair, and discrimination, and several case studies highlighting the challenges with AI/machine learning models and their potential to discriminate.

II. Identifying discrimination through disparate impacts in models
Includes qualitative and quantitative testing options, a discussion of protected groups and proxy variables, monitoring activities, and suggestions for a company's framework around model governance.

III. Regulatory landscape and additional considerations
Includes an overview of the regulatory landscape surrounding this issue, resources for actuaries, and considerations for insurers.

I. Defining Unfair Discrimination[1]

Discrimination is a difference in treatment of different classes of people. Insurance pricing and underwriting has long been based on the principle of actuarial fairness or fair discrimination, which allows insurers to use particular facts about people to measure risk and set rates. Insurers use actuarial principles to determine appropriate rates by finding relationships between factors and risk of loss and then allocating costs accordingly.

Questions about the fairness of such discrimination are arising, especially as insurers gain access to more data and use this data in more complex models and algorithms. Unfair discrimination occurs when an insurer considers factors unrelated to actuarial risk when determining whether insurance will be sold to a particular person or class—and, if so, at what cost and on what terms. Consequently, actuarially supported risk classification is "fair discrimination."

However, state legislatures create discrete statutory exceptions to this core rule. When state legislatures determine that a factor's social unfairness exceeds the benefit of its predictive value, states limit the ability of insurers to price according to risk because the social costs of risk classification. States accomplish this by creating protected classes of applicants and policyholders who cannot be discriminated against regardless of actuarial support. Discriminating against these protected classes is called "unlawful discrimination."

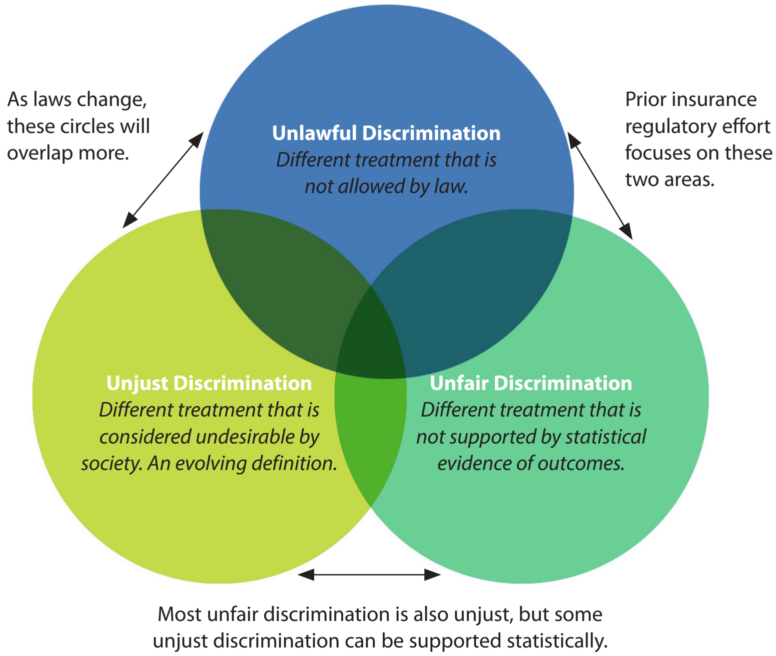
As society evolves, more categorizations are emerging as potential concerns. Differentiating by these characteristics may be "fair" and "lawful" but also considered undesirable or "unjust" by society.

See Figure 1 for a visual representation of these ideas.

1 For a glossary of terms used in this issue brief, see Appendix B.

Members of the Data Science and Analytics Committee, which authored this issue brief, include Jennifer Balester, MAAA, FCAS; Kirsten Pedersen, MAAA, FSA; and Andrea Rome, MAAA, FSA.

Figure 1: Connections Between Unlawful, Unfair, and Unjust Discrimination



According to ASOP No. 12,[2] the following should be considered by the actuary in the selection of rating characteristics:

- Relationship to expected outcome—the characteristic should differentiate between groups based on an expected outcome, such as cost.
- Causality—the characteristic should be related to the expected outcome, but it is not necessary to establish a cause-and-effect relationship.
- Objectivity—the characteristic should be based on readily verifiable, observable facts.
- Verifiability—the characteristic should not be able to be easily manipulated by the insured.
- Ease of administration—trade-offs should be weighed between the cost to use the variable and the benefit received from using it.
- Applicable law, industry practices, and business practices—the variables should comply with laws and be consistent with current practices.

Note, the ASOP No. 12 criteria above align with the definition of unfair discrimination. If an outcome is not based on sound actuarial principles[3] or is not related to actual or reasonably anticipated experience, it does not meet the required objectivity and is unfairly discriminatory. If an outcome violates the regulatory standards of fair practice, it does not meet applicable law, industry practices, and business practices in the ASOP No. 12 criteria.

2 Actuarial Standard of Practice No. 12, *Risk Classification (for All Practice Areas)*; Actuarial Standards Board. Currently being revised.
3 Sound actuarial principles require any underwriting or rating factor to accurately distinguish individuals on the basis of differences in expected costs associated with the transfer of risk.

While ASOP No. 12 does not require it, Precept 1 of the actuarial Code of Professional Conduct states: "An Actuary shall act honestly, with integrity and competence, and in a manner to fulfill the profession's responsibility to the public." Even if a variable fulfills ASOP No. 12 characteristics, and an actuary meets all applicable laws and regulations, unjust discrimination can still occur, and can be an important consideration for the actuary in their responsibility to the public.

Federally protected classes include race, color, religion, sex (including pregnancy, sexual orientation, or gender identity), national origin, age (40 or older), disability, and genetic information (including family medical history).[4] The classes are applicable, not only in employment, but in public funding, housing, and public accommodations. Each state can identify additional classes or remove classes from this list for protection under state laws. However, discriminating based on these classes is not always prohibited. For example, women were not allowed to serve in combat roles in the military for many decades. To this day, the military can still discriminate based on age. In insurance, protected and prohibited classes likewise vary by jurisdiction, and purpose (e.g., type of insurance). The regulatory, societal, and business landscape is changing rapidly.

Insurance rating systems were created to distinguish between risk classes. Insurers in ever more competitive markets strive to build models that better differentiate between insureds. Correctly identifying, segmenting, and pricing insureds can impact the solvency and financial strength of an organization. Some of the federally protected classes are deeply embedded in the insurance system, and few would argue that there is not a difference in propensity or severity of loss based on a difference in the characteristic. For example, a 75-year-old, on average, will have higher health care costs than a 35-year-old. Historically, risk classification using a given characteristic has been allowed if a relationship to loss propensity can be established.

Because fair and lawful discrimination is so deeply embedded in the insurance pricing process, the question becomes, when does this discrimination become "unfair" or "unjust"? Discrimination need not be overt to be illegal or run afoul of insurance regulators. Regulators are concerned with intentional discrimination, called disparate *treatment*, but also with disparate *impact*, whereby a protected group is disproportionately impacted by policies or practices that on the face appear to be neutral. Disparate impact may be caused by proxy discrimination, that is, when a facially neutral trait is used as a stand-in for a prohibited trait. See Figure 2.

4 U.S. Equal Employment Opportunity Commission.

Figure 2: Relationship of Disparate Impact and Proxy Discrimination

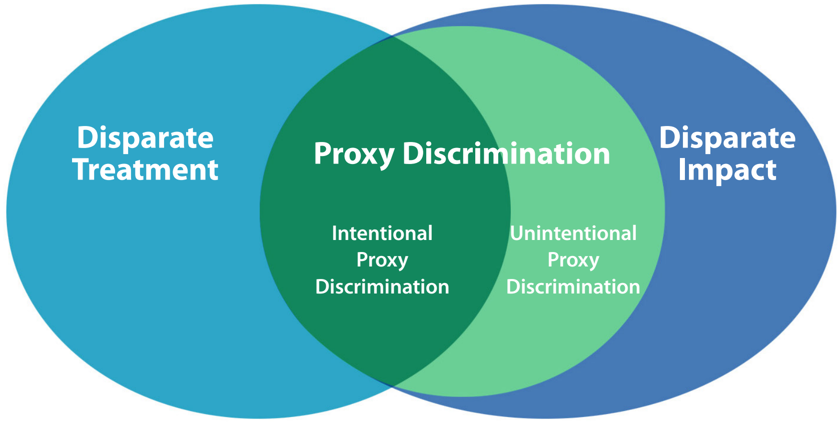


Diagram adapted from "Proxy Discrimination in the Age of Artificial Intelligence and Big Data", Section 2, by Prince & Schwarcz.

There are numerous examples of unintended discrimination based on the use of advanced algorithms or artificial intelligence. A study undertaken at the Massachusetts Institute of Technology showed that facial recognition software was much more accurate at identifying gender on photos of white males.[5] While gender was identified correctly in 99% of photos of white males, it was identified correctly in only 65% of photos of darker-skinned females. This represents a disparate impact, because (although unintended), the algorithm may have negatively affected the classes for which it performed poorly.

A ProPublica study[6] of the COMPAS recidivism algorithm showed that a risk assessment of those arrested for crimes was shown to incorrectly assess Black arrestees as higher risk of committing another crime while incorrectly assessing white arrestees as lower risk of committing another crime. White defendants assessed as low risk were arrested again 47.7% of the time as compared to 28.0% of the time for similarly assessed Black defendants. White defendants assessed as high risk were not arrested again 23.5% of the time as compared to 44.9% of the time for similarly assessed Black defendants. The inaccuracy of the algorithm had a disparate impact on Black arrestees.

It can be extremely difficult to discern whether variables used in predictive models or other artificial intelligence have a relationship with a protected class. Insurance companies have not historically attempted to measure disparate impact. They often do not collect or retain data on protected characteristics such as race, with some being forbidden to do so by statute and many others believing that to do so would be, in and of itself, unjustly discriminatory. While these data have not historically been collected, this approach is changing. Many are now advocating for "fair machine learning" algorithms, which have

5 "Facial Recognition Is Accurate, If You're a White Guy"; *The New York Times*; Feb. 9, 2018.
6 *"Un"Fair Machine Learning Algorithms*; Fu, et al.; June 10, 2020.

been proposed to adjust for social inequities. Fair machine learning requires collecting protected class data in order to better identify whether discrimination based on these features is occurring.[7]

Some characteristics historically used in insurance risk classification are linked to historical practices that may be unjustly discriminatory. Examples include motor vehicle reports (MVR) and credit-based insurance scores.

Motor vehicle reports containing details of a person's driving history have long been relied upon for automobile insurance underwriting and pricing and are used for life insurance underwriting as well. MVRs would seem to be an obvious choice for insurance underwriting and pricing given the seemingly objective and verifiable information contained in the report, as well as the demonstrable and understandable correlation between driving violations and the cost of insurance claims.

However, potential problems with MVRs arise when one digs deeper. Different states report different information on drivers. Lack of reciprocity between states can lead to inconsistencies in the MVR for two otherwise similar drivers, as can the ability in some jurisdictions for a driver to pay a fine at the time of the violation to avoid having it show up on the driving record.

Additionally, drivers from certain minority groups are more likely to be stopped and cited by police than white drivers. A 2016 study of 14 years of traffic data in North Carolina found that Black drivers were 63% more likely to be stopped by police even though they drove less.[8] Accounting for time on the road, the adjusted difference was 95%. Black drivers were also searched at double the rate of white drivers, with the searches of white drivers turning up more illegal activity. From this activity, we can extrapolate that Black drivers would have more violations on their MVR than a similar white driver and would therefore be charged a higher premium, despite similar levels of risk. Referring back to ASOP No. 12, this represents a case where the MVR's relationship to the outcome (driver risk and ultimate cost) is not clear.

Credit-based insurance scores and other credit-related information have been widely adopted in insurance over the past 25 years. FICO estimates that 95% of personal lines (automobile and homeowners) insurers are using credit-based insurance scores.[9] While insurers only need to show that the predictive variable is correlated to claims costs (they do not need to explain why there is correlation), the industry has often offered the explanation that individuals who are careful with their credit will be careful in other

7 "Racial disparities revealed in massive traffic stop dataset"; University of South Carolina website; June 12, 2020.
8 Ibid.
9 "Credit-Based Auto Insurance Scores Explained"; *Forbes Advisor*; Nov. 23, 2022.

areas of their life, such as driving and taking care of preventive maintenance on their homes. Many argue, however, that the use of credit information in insurance adversely affects certain minority groups and lower-income customers. A 2007 report by the Federal Trade Commission found that while non-Hispanic whites and Asians were spread evenly over the range of credit insurance scores, Blacks and Hispanics were more heavily concentrated in the lowest scores that would be associated with the highest risk and highest prices.[10] The same report found that the credit-based insurance scores had a correlation with claims costs that was not simply a proxy for race. This would suggest that actuaries may use credit-based insurance scores and still follow the criteria of ASOP No. 12. The credit-based insurance scores differentiated between members of the same racial group (i.e., Hispanics with lower scores had higher estimated risk based on other factors than Hispanics with higher scores). If credit-based insurance scores were removed completely from automobile insurance pricing, some estimate that two-thirds of all customers would see price increases as the additional cost of high-risk insureds is spread across all insureds.[11]

As regulators grapple with what unfair and unjust discrimination look like, they might find it beneficial to keep in mind the role of regulation in the affordability and availability of insurance products. There are downstream consequences to removing the ability of insurers to price on certain characteristics. In a non-compulsory market, if insurance becomes too expensive for a low-cost group because the insurer is not allowed to price using a given characteristic, members of that group may decide to retain their own risk and leave the market. This leads the insurers to raise prices for the remaining insureds, driving more insureds to leave the market in a "death spiral" that leaves only those with the highest risk willing to purchase coverage at increasingly unaffordable prices. In a compulsory market, an inability to classify by risk factors may lead to rating, underwriting, marketing, and other business functions that do not align with actual risk. This may limit product availability due to profitability impacts.

The examples above demonstrate issues that must be considered carefully by regulators and algorithm creators. The concerns surrounding accurate rating and solvency should be balanced with concern over how insurance practices impact larger social goals and whether they exacerbate inequalities that exist outside of the insurance system. The key is proactively identifying, understanding, and preventing the use of characteristics that are socially suspect in algorithms.

10 *Credit-Based Insurance Scores: Impacts on Consumers of Automobile Insurance*; Federal Trade Commission; July 2007.
11 Ibid.

II. Identifying Disparate Impact in Models

Insurance legislation prohibits unfair discrimination while allowing actuarially justified selection of risk, or fair discrimination. However, many functions inside an insurance company—such as marketing, rating, and underwriting—increasingly use big data, algorithms, and machine learning. These processes may use variables that are facially neutral but produce a disparate impact to different classes of people. The result of some of these processes is unjust discrimination. Companies should assess these functions to determine where disparate impact exists.

Each company function has the risk of producing a disparate impact. Underwriting may be especially key because this function can determine where customers are first segregated into risk groups. Marketing models can determine who has access to be underwritten and become insured. If marketing models produce a disparate impact, certain groups may not have easy access to insurance. Policy and benefit changes may disproportionately affect one group over another.

Most insurance companies, predictive modelers, and regulators demonstrate an understanding of how to omit unlawful discrimination from these processes and models. However, unwanted disparate impacts may remain within proxy variables. This section provides a practical approach, including many techniques for testing and measuring that will support early detection of disparate impact in models. In addition, these tests could be used to identify potential bias in underlying data.[12]

A. Testing and Measurement

Although definitions of algorithmic fairness exist in literature, the following underlying criteria summarize fairness in algorithms:[13]

> *Statistical Parity*—Given two distinct classes—a protected and an unprotected class—the algorithm should produce outcomes in equal proportion for each class. For example, drivers being categorized by risk will form an equivalent distribution of results, regardless of race.
>
> *Conditional Statistical Parity*—Given two distinct classes, controlling for a small set of "legitimate" risk factors, the algorithm should produce outcomes in equal proportion for each class. For example, drivers with the same age and driving history should compute at a similar risk, regardless of race.

12 "Testing for Statistical Discrimination by Race/Ethnicity in Panel Data for Depression Treatment in Primary Care"; *Health Services Research.* April 2008.
13 "Algorithmic decision making and the cost of fairness"; *Proceedings of KDD '17*; August 2017.

Predictive Equality—The accuracy of the model is consistent across groups. For example, if a model predicts risk of an emergency room visit with 85% accuracy for men, it should also be 85% accurate for women.

The following are examples of disparate impact tests that may be used in practice.

The Four-Fifths Rule

In 1978, The Equal Employment Opportunity Commission (EEOC) released guidelines that have become the standard for hiring in the United States.[14] The guidelines state:

> *"The use of any selection procedure which has an adverse impact on the hiring, promotion, or other employment or membership opportunities of members of any race, sex, or ethnic group will be considered to be discriminatory…"*

The guidelines describe a measure known as the "four-fifths" rule, wherein any protected race, sex, or ethnic group should have a selection rate that is at least 80% of the selection rate of the group with the greatest selection rate. This is a common measure of disparate impact. An 80% threshold (and sometimes even 70%) is used by the IRS to determine if self-insured health reimbursement plans are nondiscriminatory.[15] The same is true for 401(k) contributions.[16]

The threshold is simple to apply. For example, if a company receives 10 applications from men, and 6 applications from women, if the company chooses to interview 3 men (a 30.0% "pass" rate) and 1 woman (a 16.7% "pass" rate), the disparity in the pass rates (16.7% ÷ 30.0% = 55.6%, less than 80%) would be a flag to the hiring manager and others observing the process that the hiring practices may contain bias in favor of men. After flagging this process for potential bias, more statistical tests and review are needed.

This established guideline, following the principle of statistical parity, can serve as a starting point for data scientists when evaluating algorithms for bias in many industries for many purposes. It can be applied both to underlying data and to model results. There is more detail on this method in Appendix A.

Tests of Statistical Significance

After identifying potential disparate impact issues, the modeler should also determine whether their calculations are statistically significant. In other words, if the four-fifths test says there is a difference in results between one class and another, then we must measure the likelihood that this is not due to random chance. Two types of tests could be performed here:

14 Code of Federal Regulations Title 29, Part 1607.
15 "Nondiscrimination Testing of Self-Insured Health Benefits: An Overview"; Buck; May 3, 2018.
16 "401(k) Nondiscrimination Testing—Basics and Deadlines"; Employee Fiduciary; Jan. 5, 2023.

- Z Test, also known as a 2 Standard Deviations test.
  [*The U.S. Department of Labor uses an absolute value Z-Score of 2 and above as a statistically significant value for its disparity-checking procedures.*[17]]
- Fisher's exact test (better for sample sizes below 30).
  [*The U.S. Department of Labor uses a p-value of less than 0.025 as a statistically significant value for its disparity checking procedures.*]

There are more details on these methods in Appendix A.

Tests of Practical Significance

Beyond statistical tests, there are some more practical tests that can be applied to identify disparate impact in algorithms:

- *Back-testing*—Show performance of an updated model compared to previous models and previous methods.
- *Scenario Testing*—Establish a set of finite scenarios or cases to run through the model. For example, if testing two identical populations that differ by a protected class, would the model produce the same results? If the company expands the model to a new market, will the model perform differently?
- *Modeler Bias*—Even if exhaustive testing is done, not every case will be measured or even considered, as many of the test decisions are reliant on the modeler's own experience. Consider implicit bias testing, a measure of underlying biases a person may hold, to identify any possible blind spots and keep those in mind as predictive models are created.
- *User Testing*—Similar to market research, test out the models on a small but diverse group of users. Receive feedback from these users on results and determine whether there are any holes in the algorithm.
- *Ethical Matrix*—There are many examples of this technique of applying a set of ethical considerations by stakeholder into the model design process. One specific example comes from the book *Ethics of Artificial Intelligence* by S. Matthew Liao, where an ethical matrix that incorporates the language of data science is presented for consideration.

Examples of practical testing are provided in Appendix A for further reading.

17 "Validation of Employee Selection Procedures"; U.S. Department of Labor, Office of Federal Contract Compliance Programs.

There are several prebuilt tools[18] available from Google, IBM, ORCAA, Microsoft, Python, and others that can aid in detecting and mitigating disparate impacts. These can be helpful but may not catch all instances of discrimination and should not replace the experience and judgment of quality data scientists, actuaries, and industry experts.

The above processes can be followed to decrease any immediate risks of disparate impact as a model is being developed. An additional resource regarding model best practices is the National Association of Insurance Commissioners' (NAIC's) Principles on Artificial Intelligence.[19]

In instances where protected classes are not immediately available in the data for testing disparate impact, approximations can be made using U.S. Census data or other public sources of demographic information. In cases such as these, it is not appropriate to measure discrimination on a person-by-person basis. Instead, these approximations can show general harm to protected classes that is still useful for modifying business practices or the algorithm.

B. Other Considerations While Identifying Disparate Impact

The EEOC lists several protected classes whose members are protected from employment discrimination that occurs because of that person's membership in a protected class.[20] In employment-related algorithms, it is important to keep these classes in mind and to perform the appropriate testing. For non-employment algorithms, the list of classes to watch for is not as straightforward; it depends greatly on regulations from state to state and what product is being sold.

In insurance risk classification algorithms (and any secondary models), the list of classes to test is not as clear-cut. At a basic level, the model being used for risk classification must exclude prohibited rating variables applicable to the product and location (unlawful discrimination). Further, the process should review any significant variables not supported by evidence and/or are being applied at the individual level instead of a class level (unfair discrimination).[21] In addition, certain variables may be problematic based on social and ethical considerations of fairness (unjust discrimination).

18 "5 Tools to Detect and Eliminate Bias in Your Machine Learning Models"; *Towards Data* Science; March 2, 2021.
19 "National Association of Insurance Commissioners (NAIC) Principles on Artificial Intelligence (AI)"; NAIC; Aug. 14, 2020.
20 "3. Who is protected from employment discrimination?" U.S. Equal Employment Opportunity Commission.
21 "Unfair Discrimination," Sec. G(3), pp. 880-884; *Unfair Trade Practices Act*; NAIC Model Regulation Service; January 1993.

In algorithms where the consumer is not directly impacted financially and the model is being built for financial insight, for marketing, or for other predictive exercises, there is less regulation to prohibit potential disparate impacts. Nevertheless, it is prudent to examine the algorithm for unlawful, unfair, and unjust discrimination for the following reasons:

- *Company Reputation*—An algorithm that disproportionately impacts a protected class could introduce some reputational risk if it is discovered by an outside party.
- *Correctness*—If an algorithm or underlying data contains bias that is not identified or addressed, the results of the algorithm may not be interpreted or applied correctly. Depending on the application, this could result in severe consequences for the company and affected customers.
- *Inclusive Design*—Even if a company's current customer base is homogenous, a company looking to grow will want to make sure its algorithms can expand and apply to the wider population without major rebuilding of the model. This may not always be possible, but it is important to consider the purpose of the algorithm and how it might be expanded in use.
- *Modeling Decisions*—
    - *Bias / variance trade-off:* An understanding of disparate impacts that are occurring in the model might inform the modeler's decision to adjust the complexity of the model to produce more reasonable results. For example, if a model that incorporates a proxy variable produces disparate impacts of an undesirable nature, the modeler can examine whether the effect is lessened with more variables, or with the removal of the proxy.
    - *False positives and true negatives:* An understanding of how the algorithm impacts various classes could inform the modeler's decision on whether it is more beneficial to err on the side of more false positives or more true negatives.
- *Regulatory Considerations*—There may be requirements in certain states regarding fair algorithms that extend to many insurance activities. (See Colorado SB 21-169 in Section III of this issue brief.)

C. Identifying Problematic Proxy Variables

Antidiscrimination laws limit actuarially justified insurance discrimination for socially suspect characteristics like race and income. However, removing disallowed characteristics from algorithms does not necessarily address the problem. In many cases, data points attached to a particular person are not independent of each other. In the absence of disallowed variables, the algorithm may learn to identify and rely upon seemingly facially neutral variables that have a close correlation to protected characteristics or traits. A few examples of problematic proxy variables might include:

- Name (through text recognition) as a proxy for gender, race, religion, or age
- ZIP code as a proxy for race, ethnicity, political affiliation, or socioeconomic status

These variables may cause protected classes to be disparately impacted, even when the underlying factor is removed. In fact, there are many scenarios where removing a disallowed factor and all variables that may proxy for a prohibited characteristic may cause a bias in another direction or produce poor results. Therefore, instead of simply removing all problematic variables, the modeler and algorithm owners should understand what problematic variables exist in the data and how they are driving results. The impact of each variable should be clearly communicated, and a business and ethical decision made on which variables should be permitted. Results may have to be modified to achieve equal outcomes while balancing insurance principles of rating.

Identifying proxies for prohibited characteristics can be done through measures of correlation between variables in the model. For example, a principal component analysis shows which variables are similarly impacting results. Any number of other analytic approaches can be used to understand the connection between the protected variable and all other features. The American Academy of Actuaries Data Science and Analytics Committee plans to discuss proxy variables in further detail in a future paper.

D. Monitoring Disparate Impacts

The producers of an algorithm ought to regularly check for bias. Some machine learning algorithms change and evolve. Others may not perform as well on new data over time. All such algorithms have the potential to produce disparate impacts. It is recommended that the algorithms, the underlying data, and the results be scanned for bias on a regular, scheduled basis—at least annually.

Third-party monitoring and scanning (by, for example, external consultants and auditors) may identify issues that were not caught by an internal review and could be considered when planning monitoring protocols.

If issues are identified, it is recommended they be addressed as soon as possible. This may include making changes to the data, the algorithm, prospective modifications to results, or retrospective adjustments to correct results for affected parties.

All algorithms are recommended to have a sunset clause: established timelines when the algorithm will be fully reviewed or fully updated or retired. The algorithms that operate in perpetuity are those that are most likely to develop issues over time.

E. Governance Structure

Any company using machine learning and AI algorithms ought to have a risk-based governance structure in place for implementation, monitoring, and addressing problems. Along with establishing a cross-functional model governance committee, that structure might include the following:

- *Company's AI Ethical Statement*—Identifies what impacts will not be allowed in an AI algorithm, the goal for machine learning, and how the company intends to address any wrongs or complaints that occur from its algorithms. This may include a customer "AI bill of rights."
- *Modeling Team*—The initial work of creating the model, performed by analysts, data scientists, actuaries, and other technical individuals is the first line of defense in identifying and removing bias from their models, as well as communicating any issues. If the modeling team is not diverse, there might be potential discrimination issues that have been overlooked.
- *Review Committee*—Company leaders and any stakeholders in use of the algorithm that have regular meetings to discuss impacts of the algorithm, and resolve any current issues around the model, including potential unfair discrimination issues. This review also includes a legal review of modeling and variables.
- *Audit Team*—Independent from the Modeling Team, this group will track the results of the model, identify how it has changed over time, and report back to the Review Committee. It may identify potential unfair discrimination that is occurring in the model and utilize third-party auditing software or consultants to check the algorithm regularly.
- *Risk Assessment*—The algorithm is thoroughly reviewed and evaluated for any potential risks, including those connected to protected classes. This is a natural part of any company's robust enterprise risk management function. Further, any potential risks of the model are vetted against the company's risk appetite statement. A procedure is established for how to report any emerging risks stemming from use of the model.

- *End-User Meetings*—Areas of the company and individuals who are using the algorithm or benefiting from its results meet regularly to discuss any unusual results or concerns.
- *Documentation*—The company documents detailed information about the governance structure, roles and responsibilities and the items above. It will also include written policies and procedures along with describing any testing that has been performed, outcomes, and business decisions (for example, decisions made regarding unfair discrimination). Additionally, a full inventory of models and algorithms will be part of it.

III. Regulatory Landscape and Additional Considerations

Discrimination may arise from the data itself, the algorithm, and the whole models used in the above practices. There is also additional risk in implementation, due to how end-users are using the model and the resulting information that is being produced. Sometimes, changes in the environment may render the algorithm unusable. The risk of all these elements has increased with the use of big data, more sophisticated algorithms, and machine learning. Given the rapid pace of change in the underlying technology, there is concern over how to oversee and regulate these algorithms.

The following sections will provide an overview of the regulatory landscape, resources for actuaries, and considerations for insurers regarding unfair discrimination.

A. Regulatory / Legislative Landscape

Historically, regulations have sometimes prohibited the use of certain protected and proxy variables—for example, in the individual commercial health marketplace. However, prohibition may not be the best way to deal with machine learning algorithms and AI, as an increasing number of variables may become proxy variables for the prohibited classes. As more variables and/or proxy variables are prohibited, there is an increased risk of insolvency, especially for non-compulsory coverages. Fair discrimination promotes insurer solvency by appropriate actuarial risk classification.

Current Regulatory Landscape

The following laws and regulatory actions are currently in place in the U.S. to protect against unfair discrimination:

- McCarran-Ferguson Act of 1945—cost-based pricing requirements. Discrimination is unfair if it not actuarially justified.
- FCRA – Fair Credit Reporting Act
- Model Unfair Trade Practices Act
- Market Conduct Exams
- Rate reviews and financial exams

All state laws prohibit unfair discrimination in rates, coverages, benefits, terms, and conditions of insurance policies. Under these laws, unfair discrimination occurs when similar risks are treated differently. States generally allow life and disability insurance to differentiate by gender and disability in their risk classification. Companies generally have compliance, audit and enterprise risk management departments that create governance and processes to verify compliance with these regulations and rules.

Newer Trends in the Regulatory Landscape

As concerns of unfair and unjust discrimination increase, states are considering how to address the issue. Some states are passing or considering new laws that broaden the definition of unlawful discrimination. Other states are updating guidance and issuing statements pointing to existing law. Below are several examples of state actions on discrimination:

- Colorado SB 21-169 (2021)—This is a key new state law that is intended to protect consumers from unfair[22] discrimination when insurers use external consumer data. This law closely aligns unfair discrimination with disparate impact. It prohibits insurer use of external data and information sources along with their algorithms and models from unfairly discriminating based on certain protected classes. These classes are race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, and gender expression. This prohibition applies to any insurance practice including marketing, underwriting, pricing, utilization management, reimbursement methodology, and claims management. After the commissioner engages in a stakeholder process, rules on how to comply will be adopted by insurance type and practice. The stakeholder process began with life and personal auto insurance underwriting. These rules will also include a provision allowing a reasonable period for remedies. Insurers must demonstrate to the state that they

22 Colorado's definition of "unfair" does not align with traditional insurance understanding of "unfair" and aligns more with this issue brief's definition of "disparate impact." The definition from SB21-169 is: "'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."

comply. The rules will require insurance companies to provide information on external consumer data and how it is used. A reasonable risk management framework must be created and maintained. On May 26, 2023, Colorado released a second draft proposed *Algorithm and Predictive Model Governance Regulation* for life insurers. The Academy responded with feedback on the original draft and will continue to monitor and provide assistance.

- Rhode Island introduced HB6236, the *Rhode Island Data Transparency And Privacy Protection Act* on March 30, 2023. This bill allows a consumer to opt out of their data being used for Algorithms. The bill also provides protections around discrimination. A similar bill was introduced in Pennsylvania.[23]
- Connecticut Notice Concerning the Usage of Big Data and Avoidance of Discriminatory Practices (2021—amended and updated April 2022). This notice reminds insurance companies using internal and external data that they must comply with state and federal antidiscrimination laws. This includes reviewing the technologies used with this data in every "facet of the insurance life cycle." The 2022 update also added an annual data certification.
- New York State Insurance Circular Letter No. 1 (2019)—While this is also not a new regulation, it does help clarify the department's interpretation of existing laws regarding unfair discrimination and transparency requirements to avoid unfair trade practices such as Insurance Law Articles 24, 25 and 42. This letter applies to life insurers using external data sources, algorithms, or models related to rate calculations or life insurance underwriting. The two key areas of concern addressed are related to unlawful discrimination and transparency for consumers. It requires an independent analysis of external data tools and sources to prove they don't collect or use prohibited data. The insurer must also establish that using external data or tools doesn't result in unlawful discriminatory underwriting or rating guidelines. The letter provides two general questions for insurance companies to consider during this evaluation. Additionally, insurers must openly communicate to consumers about the process and sources used when adverse underwriting decisions are made.

23 "2023 State-by-State AI Legislation Snapshot"; BCLP; April 13, 2013.

B. Types of Regulations Regarding Unjust Discrimination

When considering regulation to address unjust discrimination, regulators have either approached the problem by prohibiting a few problematic variables, by allowing only a few agreed-upon variables, or any range of solutions in between. Each approach has benefits and drawbacks associated with it that should be considered carefully. The range of results of these regulations is summarized in Table 1.

Table 1: Types of Regulation Regarding Unfair Discrimination

Categories of Regulation	Examples	Anti-selection Impact	Discrimination Factor	Price
No Risk Classification Allowed	+ Health care in certain countries is available to all citizens and ultimately funded through taxes.	High-risk enrollees benefit and enroll in large numbers. Unless compulsory, low-risk enrollees much less likely to participate.	Low	Equal cost for all enrollees. Higher due to selection.
Only Specified Variables Allowed	+ Health insurance—Affordable Care Act (ACA) only allows four variables (age, area, tobacco use, and family composition) to rate individuals. + California's Proposition 103, passed in 1998, limits auto insurers to only using certain factors in calculating premiums.	High-risk enrollees benefit. Low-risk enrollees less likely to participate.	Mostly Low	Minimal variation in cost. Higher due to selection.
Only Specified Variables Prohibited	+ Prohibition on classification by ethnicity, or income levels + GINA (Genetics Information Nondiscrimination Act) prohibits discrimination based on genetic information in health insurance.	Classification mostly aligns with risk. Limitations may lead to increased selection in a few areas.	Moderate	Some variation in cost, based on risk.
All Risk Classification Allowed	+ Many types of non-compulsory commercial insurance have no restriction on variables. One example is general liability insurance.	Classification mostly aligns with risk.	High	High variation in cost, aligned with risk

C. Resources for Actuaries

The following resources may provide guidance when dealing with unfair discrimination.

Actuarial Standards of Practice

- ASOP No. 12, *Risk Classification (for All Practice Areas)*—"[R]isk classification is allowed so long as it is 'based on sound actuarial principles' and 'related to actual or reasonably anticipated experience.'" This ASOP is currently being updated.
- ASOP No. 23, *Data Quality*—This ASOP provides guidance to the actuary when performing actuarial services involving data. This includes guidance on selecting, reviewing, and using data.

- ASOP No. 53, *Estimating Future Costs for Prospective Property/Casualty Risk Transfer and Risk Retention*—"This ASOP provides guidance to the actuary when performing actuarial services with respect to developing or reviewing future cost estimates for prospective property/casualty risk transfer and risk retention."
- ASOP No. 56, *Modeling*—"This ASOP provides guidance to the actuary when performing actuarial services with respect to designing, developing, selecting, modifying, using, reviewing, or evaluating models."

Casualty Actuarial Society's Statement of Principles Regarding Insurance Ratemaking (Principle 4)

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

Actuarial Code of Professional Conduct

- Precept 1: "An Actuary shall act honestly, with integrity and competence, and in a manner to fulfill the profession's responsibility to the public and to uphold the reputation of the actuarial profession."

American Academy of Actuaries Resources

- *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*; November 2021.
- *On Risk Classification*; November 2011.
- Issue Brief on *Sourcing Protected Class Information in P&C Insurance*; June 2022.
- DC DISB Testimony and Presentation; June 2022.
- *An Actuarial View of Correlation and Causation—From Interpretation to Practice to Implications*; July 2022
- *Approaches to Identify and/or Mitigate Bias in Property and Casualty Insurance*; February 2023
- *An Actuarial View of Data Bias: Definitions, Impacts and Considerations*; July 2023
- A Public Policy Practice Note: *Model Governance—Some Considerations for Practicing Life Actuaries*; April 2017
- A Public Policy Practice Note: *Model Risk Management*; May 2019

Other Actuarial Guardrails

- The *Qualification Standards for Actuaries Issuing Statements of Actuarial Opinion in the United States* requires annual continuing education on bias topics.

D. Considerations for Insurers

The following contains considerations for insurance companies to help identify and guard against unfair and unjust discrimination in addition to the testing considerations mentioned previously:

Data

- Data sources should be analyzed for objectivity, applicability, statistical credibility, and accuracy. This includes traditional and nontraditional data sources.
- Certain proxy variables should be carefully reviewed in models due to the potential adverse impact to protected classes. However, it may be a viable option to use some of these in testing in place of variables that are not accessible. Additionally, including the protected class variables in the models in order to identify and then remove the effects of the variables on protected classes is being researched and recommended more.
- There is an open issue of how to approach data correction on nontraditional data sources. Using data subject to the Fair Credit Reporting Act allows customers to contest questionable data. Using other types of data does not currently have this protection for consumers. This will need to be addressed in order to maintain consumer trust.

Modeling

- Modeling teams involved in all aspects of creation, development, review, and monitoring should be diverse to protect against biases and model drift over time.
- Teams will want to choose best practices for the specific line of insurance and approach.
    - Model creation/selection
    - Model testing and auditing, including continuous maintenance
    - Model governance and risk management
- Creating an "Insurance Modeler's Hippocratic Oath" will help modelers frame their approach to predictive models and consider any downstream effects from their work. There is one in the "Financial Modeler's Manifesto"[24] that could be used as a framework.

24 "Financial Modeler's Manifesto"; Derman and Wilmott; Jan. 7, 2009.

Disclosure and Customer Service

- Algorithm Bill of Rights[25]
    - · Companies should be transparent about data, models, and algorithms.
    - · Customers have the right to understand what is being used and why.
    - · There should be a path of recourse for complaints and unfair impacts.
    - · Customers should feel secure that there is no unfair discrimination.
- Make customers whole—If unfair discrimination affects consumers, there should be methods for companies to restore fairness and compensate customers for adverse outcomes, if applicable. Similar discussions can be explored around unjust discrimination as it impacts consumers.
- A safe harbor could be defined for unintentional unfair discrimination if testing and all other model guidelines are followed. Transparency and communication with disclosures between all parties would need to be defined as well as processes to make customers whole. If there are too many roadblocks and penalties for unintentional errors, innovation could ultimately be stifled. Creating a safe harbor will help provide a balance between the benefits provided by new technology, data and processes, and their unintended risks. Similar discussions can be explored around unjust discrimination as it impacts consumers.

25 *A Human's Guide to Machine Intelligence*; Kartik Hosanager; 2019.

Appendix A: Technical Detail for Testing

The following information is provided for reference, detailing some of the potential techniques for testing, monitoring, and adjusting algorithms with potential discrimination concerns. Model owners should only create algorithms that they are qualified to produce. Model owners should fully understand the function of the model and be able to explain it to others. When potential discrimination issues are unearthed, these issues should be disclosed to stakeholders and dealt with in a way that promotes solvency yet does not unfairly harm the public.

80% Rule or "Four-Fifths" Rule

This rule assumes a "privileged" group and "unprivileged" group with which to compare ratios of a favorable outcome, like so:

This rule becomes more difficult when being applied to multiple groups and multiple favorable outcomes. Therefore, it is recommended that testing be completed using the four-fifths rule by modifying the test like so:

$$\frac{P(X = favorable\ outcome)|(Y = unprivileged\ group)}{P(X = favorable\ outcome)|(Y = privileged\ group)} \geq 80\%$$

Where each outcome is measured across each type of group. The results can be reviewed graphically by group or outcome with the 80% threshold displayed on the graph.

$$\frac{P(X = outcome_i)|(Y = unprivileged\ group_j)}{P(X = outcome_i)|(Y = privileged\ group)} \geq 80\%$$

This measure should be applied both to the underlying data, with respect to the target variable, and to the model results, with respect to the predicted variable.

Confusion Matrices and F1 Scores

This method uses the measures of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) to calculate the accuracy of the model as a whole and each class within the model.

An F1 score is a measure of how correctly the model predicts results, where a score of 100% is a perfectly accurate model.

$$F_1 score = 2 \times \frac{2TP}{2TP + FP + FN}$$

This score can be calculated for each class, as well as the entire model (either Micro, Macro, or weighted). Then a comparison can be made between the class and the whole. Potential unfair discrimination issues can be uncovered by identifying performance gaps between classes.[26]

26 For more information on this method, please see "[Confusion Matrix for Your Multi-Class Machine Learning Model]"; *Towards Data Science*; May 29, 2020.

Disparate Impact Remover

When reviewing initial data, the modeler may discover that one group has a significantly different profile from another, based on features that are being used as predictors. One method for adjusting the model to remove the disparate impacts is to modify that feature so that the algorithm can no longer distinguish between groups using it.[27]

Model Drift Testing

Model drift is the tendency, either over time or suddenly, for the relationship between model features and results to degrade. This drift negatively impacts model performance and may lead to unfair discrimination in algorithms. Model drift is usually observed by end users when the algorithm doesn't work well.

The best way to detect this is to monitor the algorithm. This can be done by setting thresholds on model measurement that include:

- *Average model runtime:* +/- runtime change threshold can help you determine whether something has changed "under the hood" in the algorithm.
- *Model metrics on accuracy:* If the number of false positives changes beyond a threshold number, or the accuracy changes more than +/- the original number, this may be a signal that the model has drifted.
- *Data metrics showing unbalanced or new classes:* If the starting data is significantly different than the data the model was trained on, you may need to adjust.

It may also be prudent to consider rolling out new versions of the model to a smaller population for a fixed period, to determine that all appropriate thresholds are met. These business rules and the above thresholds should be a documented part of a company's algorithm workflow.

There are many other methods and independent software[28] that can help analyze model drift. It is recommended that these tools be used along with other techniques, as out-of-the box solutions may not always identify issues specific to each algorithm.

Bias Analysis Focused on Model Error

Much has been written about the trade-off between bias and variance in machine learning algorithms. If a model has high bias, it has been oversimplified and may not do a good job distinguishing between classes. High variance indicates the opposite—that the model is very precise and may be overfitted to the training data. Therefore, it is important to use techniques to quantify bias and measure the direction, magnitude, and uncertainty of systemic errors.

27 This technique is introduced in *Certifying and Removing Disparate Impact*; by M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian; July 16, 2015.
28 "Drift detection overview"; IBM; April 21, 2023.

The following resources will be helpful when understanding and applying these techniques:

> "Balancing Bias and Variance to Control Errors in Machine Learning"; *Towards Data Science*; May 5, 2017.
>
> "Good Practices for Quantitative Bias Analysis"; *International Journal of Epidemiology*; December 2014.

Linear Discriminant Analysis

This method is closely related to analysis of variance (ANOVA) and regression analysis, where a dependent variable is expressed using a linear combination of other features. This method is different in that the dependent variable is categorical. It is also closely related to principal components analysis (PCA), where combinations of variables are analyzed and selected that best explain the data.

In machine learning, this approach, much like PCA, can be used as a dimensionality reduction technique. It can capture important information while limiting the number of variables.

In determining whether a model contains discrimination concerns, a modeler can use this technique to find the driving components of each class and identify potential proxy variables and other problematic variables that may benefit from a closer examination.

Tests of Statistical Significance:

Z-Test, or 2 Standard Deviations Test

A z-test is performed to determine whether results are statistically significant. In other words, is the potential discrimination identified in the model large enough/credible enough to warrant further action?

A few conditions need to be met when performing a z-test:

- Sample size must be larger than 30.
- Data points should be independent.
- Ideally, data should be normally distributed, but not strictly necessary with large samples.
- The sample should be randomly selected from a larger population where all data points have an equal chance of being selected.

An example of this test:
There are 100 people in our sample—60 in Class A and 40 in Class B. Class A positive results are at 75% and Class B positive results are at 50%.

Because Class B results are less than 80% of Class A positive results, Class B fails the four-fifths rule, and there might be some disparate impacts occurring in the model.

To understand the likelihood of this being a valid result, the next step is to perform a 2-proportion z-test.

The null hypothesis is that the difference in class results is statistical noise.

The alternate hypothesis is that the difference in class results is significant.

Class A proportion = P_A = 45/60 = 75%
Class B proportion = P_B = 20/40 = 50%
Total Sample proportion = P = (45+20)/(60+40) = 65%

$$Z = \frac{(P_A - P_B) - 0}{\sqrt{P(1-P)(\frac{1}{n_A} + \frac{1}{n_B})}}$$

$$= \frac{(75\% - 50\%) - 0}{\sqrt{65\%(1-65\%)(\frac{1}{60} + \frac{1}{40})}}$$

Z = 2.568

The U.S. Department of Labor[29] uses an absolute value Z-Score of 2 and above as a statistically significant value for their disparity checking procedures (approximately a +/- 2.5% likelihood of null hypothesis). Therefore, this is considered a statistically significant result, and further action should be taken to review the disparate results between Class A and Class B.

Fisher's Exact Test or Chi-Squared Test

Fisher's Exact Test is based on a hypergeometric distribution and provides an exact p-value representing the likelihood of the null hypothesis, that the variation occurred randomly and is statically insignificant. This test is better for sample sizes below 30, while sample sizes above 30 follow a similar method, called the Chi-Squared Test.

Below is an example using Class A and Class B, but with smaller sample sizes:
There are 30 people in our sample. 20 in Class A and 10 in Class B. Class A positive results are at 75% and Class B positive results are at 50%.

29 "Validation of Employee Selection Procedures"; op. cit.

Because Class B results are less than 80% of Class A positive results (50% ÷ 75% = 66%), Class B fails the four-fifths rule, and there might be some disparate impacts occurring in the model. Failing this initial test, the next steps are to determine whether those differences are statistically significant. To do so, the modeler must test the current case and anything more extreme (in this case, where the proportion in Class B is even lower).

This Case:	Positive Result	Negative Result	Total
Class A	15	5	**20**
Class B	5	5	**10**
Total	**20**	**10**	**30**

$$Probability = \frac{Positive_{total}!\ Negative_{total}!\ A_{total}!\ B_{total}!}{Total!\ Positive_A!\ Negative_A!\, Positive_B!\ Negative_B!}$$

$$Probability = \frac{20!\ 10!\ 20!\ 10!}{30!\ 15!\ 5!\ 5!\ 5!} = 0.130$$

More Extreme Case:	Positive Result	Negative Result	Total
Class A	16	4	**20**
Class B	4	6	**10**
Total	**20**	**10**	**30**

$$Probability = \frac{20!\ 10!\ 20!\ 10!}{30!\ 16!\ 4!\ 4!\ 6!} = 0.034$$

And so on, down to 20 Class A positives and 20 Class B negatives. Adding up, we get a resulting probability = 0.169 for this current sampling and all "more extreme" cases.

The U.S. Department of Labor[30] uses a p-value of less than 0.025 as a statistically significant value for its disparity checking procedures. Using that measuring stick, this particular example indicates a statistically insignificant result, and the null hypothesis is supported.

30 ibid.

Appendix B: Glossary

The following terms are discussed in this issue brief:[31]

Algorithm—A set of instructions to describe how to implement a process.

Artificial Intelligence (AI)—A program that performs a task that one would normally think of as being calculated or determined by humans. It is discussed in the context of automation and human augmentation. It can also be a group of algorithms that can modify and create new algorithms as it processes data.

Conditional Statistical Parity (CSP)—CSP is closely related to the idea of fairness through blindness, in which one attempts to create fair algorithms by prohibiting the use of protected attributes, such as race. However, as frequently noted, traditional approaches to achieve fairness find it difficult to restrict to legitimate features that do not at least partially correlate with race and other protected attributes, and so one cannot be completely "blind" to the sensitive information and so do not necessarily reduce disparities. Conditional statistical parity mitigates these limitations of the blindness approach. CSP is discussed in a later section of this issue brief.

Discrimination—A difference in treatment of different classes of people. Can be intentional or unintentional.

> ***Unfair Discrimination***—In insurance, a difference in treatment of different classes of people that is not based on solid statistical evidence of differing risk. Most actuarial certifications require a statement that rates are not "unfairly discriminatory."
>
> ***Unlawful Discrimination***—Discrimination that is in violation of regulatory standards or statutes.
>
> ***Proxy Discrimination***—A facially neutral practice that results in a difference in treatment of different classes of people.
>
> ***Unjust Discrimination***—Socially undesirable inequities. Discrimination that is in violation of societal or ethical understanding of equal opportunity. This discrimination may be lawful, and may be considered fair in an insurance context, but is still undesirable from a societal standpoint.

31 Definitions consistent with *Big Data and Algorithms in Actuarial Modeling and Consumer Impacts*, American Academy of Actuaries, November 2021, except for "discrimination," "statistical parity," and "unfair discrimination."

Disparate Impact—Unintended discriminatory effects on classes of persons, from using practices of neutral intent.

Disparate Treatment—Intentional unlawful discrimination, where members in a protected class are deliberately treated differently than members in non-protected classes.

Insurance—As used in this issue brief, an economic device transferring risk from an individual (or groups of individuals) to a company (or government body) to reduce the uncertainty of risk via pooling.

Machine Learning—An evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. It is a subset of artificial intelligence.

Protected Class—A group of people who share a common characteristic, for whom federal or state laws have created protections that make it unlawful to discriminate against members of the group on the basis of that characteristic.

Proxy Variable—A variable that serves as a substitute for another. In cases of risk classification, a proxy variable may be stand-in data for a disallowed variable and may drive similar results as the disallowed variable.

Statistical Parity—An algorithm satisfies this definition if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class. Also referred to as group fairness, demographic parity, acceptance rate, and benchmarking.

The American Academy of Actuaries is a 19,500-member professional association whose mission is to serve the public and the U.S. actuarial profession. For more than 50 years, the Academy has assisted public policymakers on all levels by providing leadership, objective expertise, and actuarial advice on risk and financial security issues. The Academy also sets qualification, practice, and professionalism standards for actuaries in the United States.