DECEMBER 2025

A Foundational Study of Algorithmic Bias

A Policy Paper by
Data Science and Analytics Committee

American Academy of Actuaries
actuary.org

Data Science and Analytics Committee
Authors:

Maggie Ruzicka, MAAA, ASA
Robert Gomez, MAAA, FSA
Dorothy Andrews, MAAA, ASA, PhD
Mary Bahna-Nolan, MAAA, FSA
Robert Miccolis, MAAA, FCA, FCAS
Dave Sandberg, MAAA, FSA
David Schraub, MAAA, FSA
Paula Schwinn, MAAA, FSA
Alex Esche, MAAA, ASA
Shruti Gupta, MAAA, ASA

This paper was developed thanks to the work of the Academy's volunteers. To learn more about becoming a volunteer, please visit actuary.org/volunteer.

The American Academy of Actuaries is a 20,000-member professional association whose mission is to serve the public and the U.S. actuarial profession. For 60 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.

American Academy of Actuaries

AMERICAN ACADEMY OF ACTUARIES
1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036
202-223-8196 | WWW.ACTUARY.ORG

December 2025

Any references to current laws, regulations, or practice guidelines are correct as of the date of publication.

A Foundational Study of Algorithmic Bias

Introduction and Purpose

The average person, either knowingly or unknowingly, interacts with various algorithms in their personal and professional lives multiple times throughout a typical day. Algorithms—defined for the purpose of this discussion as processing systems utilized to search, select, extract, and use data in an automated routine—appear frequently in daily life. Examples include internet advertising, which utilizes browsing patterns to target specific buyers; GPS-suggested routes based on user travel preferences; and numerous entertainment recommendations based on prior viewing patterns.

These types of algorithms seem innocuous and are considered by many as valuable time-saving mechanisms for decision making and information gathering. However, there are numerous other algorithms used for the same purpose that without checks and balances may lead to unintended consequences to not just individual groups of people but to society as a whole. The widespread belief that algorithms embody pure objectivity and unbiased decision-making stems from their computational nature, implying a detached, logical process free from human emotion or prejudice. This ideal suggests that by operating on data and predefined rules, algorithms can offer consistent, fair, and efficient solutions on a scale, overcoming the cognitive biases inherent in human judgment. This aspiration often clashes with the complex realities of their design, deployment, and the data they are trained on.

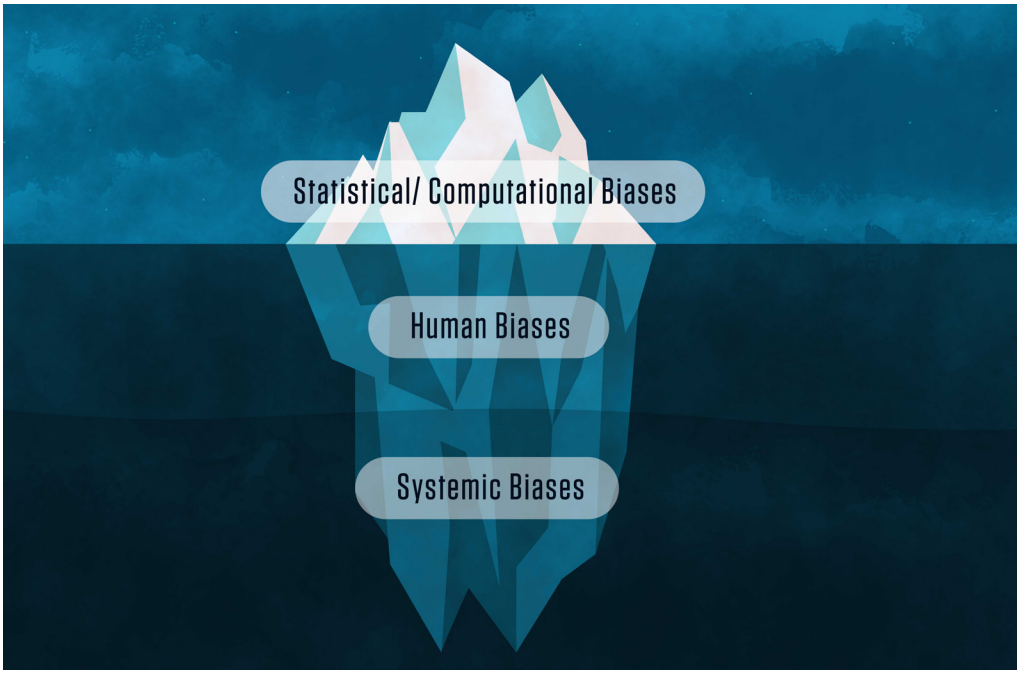
Algorithmic bias frequently arises from the historical and societal prejudices embedded within the vast datasets used for their training, or from the implicit biases of the human designers who make critical choices about features, objectives, and evaluation metrics. Biased algorithms can challenge efforts to reduce discrimination and inequality, limit opportunities for marginalized groups, and reinforce harmful stereotypes eroding trust in institutions and systems that rely on these algorithms.

Algorithmic bias refers to the inherent prejudices that can be unintentionally embedded within the algorithms that drive machine learning and artificial intelligence (AI) systems. This type of bias can stem from a variety of sources, including the data used to train the algorithms, the design of the algorithms themselves, and the societal context in which they operate. For actuaries, the impact of algorithmic bias potential continues to be a growing concern because at its core actuarial work revolves around accurate risk quantification and fair pricing.

As part of its efforts to examine actuarial practices and methods to assess the impact of algorithmic bias, the American Academy of Actuaries' Data Science and Analytics Committee is exploring issues regarding bias and algorithmic audits. This is the first of a three-paper series aimed at understanding how bias can impact and exacerbate unequal outcomes when developing algorithms used in critical decision making. Actuaries often rely on complex algorithms to aid in their work. The growth and utilization of AI to develop and process the data actuaries rely on requires a growing awareness of the potential for bias in these systems. Bias can have significant implications for both insurance companies and their customers. The consequences can be far-reaching, potentially leading to unfair treatment of individuals as well as resulting in financial and reputational losses. Biased algorithms could result in inaccurate risk assessments and mispricing affecting affordability for consumers or ultimately impact the profitability of the company. Algorithms do not create bias, as that requires a human element, but the underlying data used in the development of the algorithm can lead to unintended bias in the outcomes. Actuaries play a crucial role in identifying and mitigating algorithmic bias.

What Is Bias?

The National Institute of Standards and Technology (NIST) has defined three categories of bias using the Iceberg Principle. In the depiction[1] of the iceberg, the section above the waterline represents statistical bias. Just below the waterline is where human/cognitive biases are depicted, and systemic biases appear at the deepest depths of the iceberg.



This depiction is poignant for two primary reasons: visibility and ease of mitigation. Statistical bias is the easiest of the three to see and mitigate, while systemic bias is the most difficult of the three to identify and address.

Statistical bias is a mismatch between the results of a statistical study and the actual population it is intended to represent. It typically comes from flaws in data collection or design.

1 AI Bias Iceberg; National Institute of Standards and Technology; March 16, 2022.

Political polls are well-known examples of statistical bias. A historical example is the *Literary Digest* magazine nationwide poll ahead of the 1936 presidential election between Alfred Landon and Franklin Roosevelt. The magazine selected its population from subscriber lists and telephone directories at a time when telephones were considered a luxury. The resulting sample severely overrepresented middle- and upper-class voters in an election where economic policy was a major issue. The poll results showed Landon winning with 57% of the vote, while the actual election resulted in a 62% win for Roosevelt. This is an example of sampling bias, where the collected data doesn't represent the entire voting population, but instead reflects a convenient sample.

Sampling bias is still difficult to manage in modern exit polls, where a voting population can span urban and rural areas, income levels, ages, races, and ethnicities. Polls must be conducted across enough precincts to ensure that the sample size is representative of the diverse population. They should not be conducted only during part of the day, only offered in English, or limited to in-person voting. Additionally, polls need to consider selection bias where participants may not choose to participate; reporting bias, where participants rely on recall; and observer bias, where questions are framed to influence the answers.

These biases can be mitigated with random sampling or stratified sampling to ensure coverage of a wide range of subjects, and through careful design of questions and information gathering techniques.

Cognitive bias occurs when we interpret our world through the lens of our own experiences and beliefs. These biases provide shortcuts for us to take in new information, filter it for meaning, and choose our actions. These shortcuts, though, can be faulty.

There are over 180 cognitive biases that drive the way we make decisions and form opinions. The Cognitive Bias[2] Codex in Appendix A shows 188 cognitive biases, organized to show the many thought processes that are influenced by our cognitive biases—from our prioritization of information to conclusions we make when information is incomplete, and the thought processes that drive our actions.

2 The Cognitive Bias Codex; Wikimedia Commons; June 6, 2018.

Confirmation bias is one of the most common cognitive biases. It causes us to interpret data in a way that supports our already-formed beliefs. As a result, our own belief is strengthened while we ignore possible contrary interpretations. Many studies have been conducted on confirmation bias in medical diagnosis. One such study[3] in Germany, conducted in 2011 with psychiatrists, presented them with a patient with Alzheimer's disease who first showed symptoms of depression. After nearly all the psychiatrists diagnosed their patients with depression, they continued to question their patients, and only 13% of them followed up with investigations that would confirm their initial findings. As more symptoms were presented to show a clear case of Alzheimer's, 41% of the psychiatrists still did not change their diagnosis from their initial determination.

The framing effect is another common cognitive bias in which decisions are influenced by the way information is presented. It was especially noteworthy in combination with Simpson's Paradox during the COVID-19 pandemic. Simpson's Paradox is a phenomenon in which a trend can seemingly appear or disappear depending on how data is grouped.[4] In 2021, data was emerging from all around the world on COVID-19 severity as vaccinations were made available to everyone. Data released in August 2021 in Israel (see table below) showed that there were more severe cases of COVID-19 among vaccinated individuals (301 hospitalized at that time) than in unvaccinated individuals (214 people). As these reported statistics spread on social media, it led many to conclude that the vaccines were ineffective.

Those conclusions would have been hard to support if the same facts had been grouped into age bands—even as simple as "under age 50" and "50 and above"—and expressed as incidence rates so the sizes of the vaccinated and unvaccinated populations were reflected. In that case, we would have seen that unvaccinated people over 50 experienced severe cases of COVID-19 at nine times the rate of vaccinated people over 50. Framing the same information this way would have led people to a much more favorable conclusion about vaccine effectiveness.

COVID-19 Hospitalization Data, From Israel Dashboard, August 2021[5]						
	Population		**Severe Cases**		**Severe Case Incidence**	
Age	**Unvaccinated**	**Vaccinated**	**Unvaccinated**	**Vaccinated**	**Unvaccinated**	**Vaccinated**
< 50	1,116,834	3,501,118	43	11	0.004%	0.000%
> 50	186,078	2,133,516	171	290	0.092%	0.014%
Total	1,302,912	5,634,634	214	301	0.016%	0.005%

3 "Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses"; *Psychological Medicine*; 2011.
4 "Simpson's Paradox"; *Stanford Encyclopedia of Philosophy*; March 2021
5 "Israeli data: how can efficacy vs. severe disease be strong when 60% of hospitalized are vaccinated?"; Covid-19 Data Science; October 2021.

In addition to confirmation bias and the framing effect, we experience many other cognitive biases. Status quo bias is a preference for things to stay the same; it makes us hesitant to accept new findings, new approaches, and new conclusions. The fundamental attribution error is our tendency to excuse our own failures while assuming that failures of others are a result of their behaviors. Anchoring is a tendency to rely disproportionately on the first information we receive in a decision-making process, even as contradictory information is presented later. Finally, implicit stereotypes cause us to view people of certain professions, races, etc. negatively even against our explicit beliefs.

Cognitive biases can be difficult to overcome. Reducing the harm caused by cognitive biases requires acknowledging our own biases followed by constant, deliberate steps to test our attitudes and beliefs. This approach has shown to be effective in the psychiatric diagnosis case above[6], as the doctors who asked questions that challenged their initial diagnoses were significantly more likely to eventually arrive at the correct diagnosis.

Systemic bias results from procedures and practices that operate in ways which result in certain social groups being advantaged or favored and others being disadvantaged or devalued. Institutional racism and sexism are the most common examples.[7]

One of the most common examples of systemic bias in the U.S. is through racist practices such as de facto segregation of neighborhoods and unequal access to educational opportunities, food, medical care, transportation, and employment. For example, unequal funding for school districts creates separate and unequal educational systems.

An example[8] of this is geographic financial exclusion, commonly termed redlining, referring to the systematic withholding of lending opportunities from individuals solely based on their residential location, irrespective of the borrower's individual creditworthiness or ability to secure capital. For decades, housing finance institutions frequently applied these discriminatory policies, specifically targeting inner-city districts and severely restricting capital flows into neighborhoods predominantly inhabited by Black residents. The 1968 Fair Housing Act subsequently deemed racially motivated territorial lending illegal, assigning the critical oversight role—including compliance monitoring—to federal financial watchdogs such as the Federal Reserve.

6 Confirmation bias: why psychiatrists stick to wrong preliminary diagnoses"; *Psychological Medicine*; 2011.
7 "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence"; National Institute of Standards and Technology Special Publication 1270; March 2022.
8 "Redlining"; Federal Reserve History; June 2023.

When policies are established without recognizing the social norms and history of bias, inequity is allowed to persist. Deliberate effort is needed to recognize and reverse it. Without caution, algorithms may produce outcomes that reflect or perpetuate statistical, cognitive, and systemic biases.

Definitions

The current regulatory and industry focus on the discriminatory impact of big data and algorithms has increased the pressure to adopt precise, meaningful, credible, and actionable definitions. While there are many terms related to Artificial Information Systems (AIS) comprising big data and AI, this paper will focus on the five terms in the table below because of their materiality to this paper. The table below depicts three entities that have adopted definitions for the indicated terms.

AIS Term	NAIC	NIST	NCOIL
Algorithm	✓	✓	
Bias		✓	
Disparate Impact			
Unfair Discrimination	✓		
Proxy Discrimination	✓		✓

NAIC – National Association of Insurance Commissioners
NIST – National Institute of Standards and Technology
NCOIL – National Conference of Insurance Legislators

The National Association of Insurance Commissioners (NAIC), the National Council of Insurance Legislators (NCOIL), and the National Institute of Standards and Technology (NIST) play distinct but interconnected roles in shaping the country's regulatory and standardization landscape. The NAIC is the U.S. standard-setting and regulatory support organization, composed of the chief insurance regulators from all 50 states and U.S. territories. NCOIL works in conjunction with the NAIC, serving as the legislative clearinghouse comprising state legislators who review, recommend, and often sponsor statutory language necessary to implement model laws into actual state statute. NIST is a non-regulatory federal agency focused on measurement science, technology, and the development of standards across numerous industries.

The NAIC Model Bulletin[9], which will be discussed in more details later in the paper, defines an algorithm as "a clearly specified mathematical process for computation; a set of rules that, if followed, will give a prescribed result," which aligns with the NIST definition of the term. As of the writing of this paper, the NAIC has not formalized a definition of bias. This could be problematic if insurance companies are to be regulated for biased algorithmic outcomes. Because insurance regulation occurs at the state level, each state may adopt its own definition of bias. The NAIC, as the state member organization, can offer guidance on the adoption of definitions, but it is ultimately within the purview of states to frame and adopt definitions. The NIST categorizes bias, as discussed above, into three components: statistical, cognitive, and systemic.

While the term "disparate impact" has a specific legal definition, it has not been applied to insurance. In fact, in a 2009 policy briefing of the National Association of Mutual Insurance Companies (NAMIC), the term "disparate impact" was asserted to be in conflict with the term "unfair discrimination." Per the policy briefing, a neutral risk classification system's disproportionate impact on a protected class does not constitute unfair discrimination under any controlling state law.[10] The NAIC deems unfair discrimination as pricing individuals of the same rating class and hazard differently or when pricing differences are not justified by sound underwriting and actuarial principles and loss experience.

Finally, NCOIL defines proxy discrimination as "the intentional substitution of a neutral factor for a factor based on race, color, creed, national original or sexual orientation for the purpose of discriminating against a consumer to prevent that consumer from obtaining insurance or obtaining preferred or more advantageous pricing and insurance coverage due to that consumer's race, color, creed, national origin, or sexual orientation."[11] This definition asserts that discrimination by proxy can only occur through intentional acts. Algorithms have been shown to unintentionally result in proxy discrimination against protected classes even when protected class variables are not explicitly included in the model. Algorithms often find patterns of discrimination that are not always obvious by a survey of model variables.

9 Use of Artificial Intelligence Systems by Insurers; National Association of Insurance Commissioners; December 2023.
10 "CreditBased Insurance Scoring: Separating Facts From Fallacies"; National Association of Mutual Insurance Companies; September 2006.
11 Property/Casualty Insurance Modernization Act; National Council of Insurance Legislators; April 2021.

Types of Algorithmic Harm

The process of creating algorithmic generated decisions relies on data and machine learning. Observed data is used to "train" the system, enabling it to learn and interpret various activities, allowing it to predict future outcomes. The quality and accuracy of the data directly impact the algorithm developed from this "training" process. The process focuses on designing an algorithm to interpret observed data for predicting future outcome. What happens if the data is flawed or embedded with bias? Actuaries rely on historical data to help predict the future. However, historical data may unintentionally reinforce stereotypes that impact underserved or disregarded segments of the population. These limitations in data may lead to what is defined as algorithmic harm.

Algorithmic harms are unintended negative consequences or loss of equitable outcomes that are driven by the design and data used to develop automated decision-making technological systems.

The Future of Privacy Forum[12] has grouped algorithmic harm into four main subcategories[13] defined as follows:

1. Loss of Opportunity: This category broadly refers to harms occurring within the domains of the workplace, housing, social support systems, health care, and education. For example, a study found that a widely used algorithm for predicting which patients would benefit from extra care was less accurate for Black patients compared to white patients. This led to a potential loss of opportunity for early detection and intervention for Black patients, who may not have received the necessary care and treatment they may have needed.[14]

2. Economic Loss: This category broadly refers to harms that primarily cause financial injury or discrimination in the marketplace for goods and services. One example is the use of algorithms in medical billing and coding. These algorithms are designed to streamline the submission and processing of insurance claims but can introduce errors and incorrect coding. Such mistakes can result in denied claims, delayed payments, and even fines for health care providers, as well as potential overbilling or underbilling that financially harms both patients and health care organizations.

12 Future of Privacy Forum (FPF); Future of Privacy Forum; 2025.
13 "Unfairness By Algorithm: Distilling the Harms of Automated DecisionMaking"; Future of Privacy Forum; December 2017.
14 "Dissecting racial bias in an algorithm used to manage the health of populations"; Federal Trade Commission; July 21, 2020.

3. Social Detriment: This category broadly refers to harms to one's sense of self, self-worth, or community standing relative to others. For example, algorithms may determine the types of insurance offered, and if these algorithms are biased or flawed, they can result in some consumers being denied insurance or having very limited insurance options. This can have a disproportionate impact on marginalized communities that may already face barriers to accessing affordable insurance.

4. Loss of Liberty: This category broadly refers to harms that constrain one's physical freedom and autonomy. The use of algorithms in health care can compromise patient privacy. As algorithms rely on vast amounts of personal data to make decisions, there is a risk of this data being accessed or shared without individual consent. This can result in a loss of an individual's right to control their own personal information and make decisions about who has access to it.

Algorithms are used in many different applications, and harms may manifest in any application or industry. Decision-makers utilizing algorithmic outcomes must be aware of the potential for algorithmic harm.

Measuring the impact and determining the appropriate course of action to limit or mitigate potential algorithmic harm is a complex challenge. Determining the extent of harm is difficult as there is always a trade-off between group and individual needs. While there is no clearly defined acceptable level of harm, understanding the frequency and severity of the potential impacts is the first step toward reducing unintended consequences driven by algorithms.

Case Studies in Algorithmic Harm

The following are examples of algorithmic harm. Specific harms, as defined by the Future of Privacy Forum, are ascribed to each example.

The Amazon Hiring Algorithm—Loss of Opportunity, Economic Loss

In 2014, Amazon engineers attempted to build an algorithm to automate hiring engineers. The algorithm was designed to review resumes and select candidates for potentially hiring.[15] Before the algorithm could be put into production, it was observed that it systematically discriminated against women applying for engineering positions. The algorithm eliminated candidates who attended women's colleges, played women's sports, and used female gendered terms in their resumes. A common denominator was the term "women." Amazon engineers, who were mostly men, likely did not attend women's colleges or play women's sports. Amazon also trained the model on its own resumes which reflected its hiring patterns. A second pattern the algorithm detected that distinguished males from females was the use of gender-nuanced words common to male versus female resumes. It was observed that men tended to use words like "executed" and "captured" in their resumes, while women tended to use words such as "offered," "provided," "served," "assisted," and "collaborated." Men were found to use these words less frequently.

The website theladders.com posted a study[16] that examined words that men versus women use on resumes across various industries such as financial services, IT, management consulting, and retail. In financial services, it found the following gendered terms prevalent in the resumes of males and females.

- Male: equity, portfolio, investment, capital, analyst, finance, market, stock, interests, technical
- Female: organize, event, volunteer, assistant, social, student, marketing, community, department, plan

The female terms tend to reflect more supportive or "soft" skills, while the male terms tend to reflect more technical or "hard" skills. Using gendered language was another way the algorithm was able to distinguish male applicants from female applicants. When combined with Amazon's historical male dominated hiring pattern, it is easy to see how the algorithm discriminated against women. Fortunately, the issues were identified before the algorithm went live. Amazon scrapped it before it entered production, where it could have caused real harm.

15 "Why Amazon's Automated Hiring Tool Discriminated Against Women"; American Civil Liberties Union; Oct. 12, 2018.
16 Qu Q, Liu QH, Gao J, Huang S, Feng W, Yue Z, Lu X, Zhou T, Lv J. Gender differences in resume language and gender gaps in salary expectations. J R Soc Interface. 2025 Jun;22(227):20240784. doi: 10.1098/rsif.2024.0784. Epub 2025 Jun 4. PMID: 40462711; PMCID: PMC12134937.

Bias in Health Care—Loss of Opportunity, Economic Loss, Social Detriment

- Bias in health care has existed since the arrival of African slaves on the shores of Point Comfort, Virginia in 1619.[17] Africans were believed to be inferior to whites and medical justification was sought to validate those beliefs. Some examples of "justified" prevalent medical biases included:

- The belief that Blacks can endure more pain than whites. This belief was reflected in gynecological experimentation on female slaves without anesthesia by Dr. James Marion Sims, known as the Father of Modern Gynecology. He is known for perfecting a procedure to repair vesicovaginal fistulas, a complication of childbirth. In his autobiography, Sims described the experimental surgeries on his Black slaves as "So painful, that none but a Black woman could have borne them."[18]

- Beliefs about differential pain thresholds by race continue to this day and influence the care of Black people. White laypersons and medical students have been documented in recent studies[19] to hold this false belief about pain threshold differences between Blacks and whites; these beliefs predict racial bias in pain perception and the accuracy of treatment recommendation. The net result is that Black patients experience unnecessarily more pain than white patients.

- The belief that kidney function is racially based led to discriminatory kidney function tests. Three formulas have used or currently use an adjustment for gender, race, or both: (1) the Cockcroft-Gault (CG) equation, (2) the Modified Diet Renal Disease (MDRD) Study equation, and (3) the Chronic Kidney Disease Epidemiology (CKD-EPI) equation. The adjustments in these formulas with respect to Blacks are based on a long-held belief that Blacks have higher muscle mass than whites. Several studies suggest that the equations used to calculate how well your kidneys filter blood,[20] tends to overestimate this metric in Black Americans, potentially leading to underdiagnosis of chronic kidney disease. The CKD-EPI equation was modified in 2012 to remove the race adjustment. Using race in formulas that test the Glomerular Filtration Rate (eGFR) stems from eugenic beliefs about the inferiority of Blacks compared to whites and serves to perpetuate and naturalize racial bias in medicine.

17 Washington, H. A. (2019). Medical apartheid. Random House.
18 Colleen Campbell, Medical Violence, Obstetric Racism, and the Limits of Informed Consent for Black Women, 26 MICH. J. RACE & L. 47 (2021). https://doi.org/10.36643/mjrl.26.sp.medical
19 Ibid.
20 Zhu, Y., Ye, X., Zhu, B., Pei, X., Wei, L., Wu, J., & Zhao, W. (2014). Comparisons between the 2012 new CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equations and other four approved equations. PloS one, 9(1), e84688. https://doi.org/10.1371/journal.pone.0084688

The CG equation has a gender adjustment, but not a race adjustment. It is noteworthy that the original study[21] conducted by Cockcroft and Gault included only a sample of 249 white male patients.

- Spirometers are used to detect respiratory diseases, and commercial units reflect corrections for race.[22] The corrections are built into the software. In the United States, correction factors of 10% to 15% have been applied to Blacks and factors of 4% to 6% to Asians.[23] Such practices erroneously reinforce race as a biological and medical category. History suggests Thomas Jefferson may have given life to the lung-function bias in his 1832 Notes on the State of Virginia where he noted deficiencies of the "pulmonary apparatus" of Blacks.[24] Jefferson's comment gave justification to the beliefs of plantation physicians that slavery "vitalized the blood" of slaves, making it necessary to improve their lung functioning. This is yet another example in history of attributing scientific importance to race.

  Research has debunked these beliefs and shown that lung capacity differences are independent of race and are instead related to disproportionate exposures to toxic environments, differential access to high-quality care, and the daily insults of racism[25] [26].

- While skin cancer is more prevalent among whites than Blacks, it is often detected in Blacks at later stages than in whites, resulting in lower survival rates for Blacks. Medical training and textbooks often do not depict images of skin cancer in Blacks and, as a result, medical professionals are not trained to effectively detect skin cancer in Blacks.[27] There are efforts underway to diversify the medical curricula to mitigate this bias in medical training.

- Health care workers have also observed biases in the care of patients with respect to race, ethnicity, and language. In a survey of more than 3,000 health care workers, more than half indicated that discrimination is a significant issue in health care and reported having personally witnessed patient discrimination on the basis of age, race, and language.[28] For example, medical workers observed doctors not providing the same level

21 Zhu, Y., Ye, X., Zhu, B., Pei, X., Wei, L., Wu, J., & Zhao, W. (2014). Comparisons between the 2012 new CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) equations and other four approved equations. PloS one, 9(1), e84688. https://doi.org/10.1371/journal.pone.0084688
22 Braun L. Race, ethnicity and lung function: A brief history. Can J Respir Ther. 2015 Fall;51(4):99-101. PMID: 26566381; PMCID: PMC4631137.
23 Ibid.
24 Ibid.
25 News; American Thoracic Society; 2025.
26 Lujan, H. L., & DiCarlo, S. E. (2024). Misunderstanding of race as biology has deep negative biological and social consequences. *Experimental physiology*, *109*(8), 1240–1243. https://doi.org/10.1113/EP091491
27 Brady J, Kashlan R, Ruterbusch J, Farshchian M, Moossavi M. Racial Disparities in Patients with Melanoma: A Multivariate Survival Analysis. Clin Cosmet Investig Dermatol. 2021 May 24;14:547-550. doi: 10.2147/CCID.S311694. PMID: 34079319; PMCID: PMC8163579.
28 "Revealing Disparities: Health Care Workers' Observations of Discrimination Against Patients"; The Commonwealth Fund; February 2024.

of explanation to Spanish-speaking patients with low English language proficiencies as they did to white patients. Acceptance of self-advocacy was observed as greater among white patients than non-white patients. Perceptions of equal treatment were also observed at differential rates among whites and non-whites. The study also discusses the impact of observed differential care on the health care workers. Health care workers were found to experience stress because of the discrimination they observed. Several solutions were offered. Among them were: (1) providing anonymous ways workers could report discrimination in care, (2) ensure policies promote equitable treatment, (3) pay attention to the care of non-English speaking patients, and (4) train workers to spot discrimination in the administration of care.

- A health care referral algorithm was found to refer Black patients far less than white patients for specialized care, despite their having worse health risk scores.[29] The algorithm was biased in favor of white patients because the metric used to determine referrals was based on historical medical spending. Unfortunately, the metric did not recognize that health care spending is historically lower in communities of color due to a lack of access to health care and a general distrust of the medical establishment that discourages utilization, further contributing to lower medical spending among people of color.

These biases impact medical test results and statistical metrics based on them. It is important to recognize that third-party vendors providing metrics based on biased medical data may introduce bias into any actuarial analysis using that data. Further, it is critical to understand how third parties are constructing medical scores and how to assess those scores for embedded biases, otherwise actuaries may risk biasing actuarial model outcomes in ways that could adversely and unfairly affect certain segments of consumers.

The ProPublica Example—Loss of Opportunity, Economic Loss

Computational journalists are the data scientists of journalism. They are part journalist and part data scientist. A recent ad for a computational journalist issued by ProPublica described the role as one in which journalists working to uncover algorithms and social platforms that play an increasing role in our lives with the goal of detecting bias.[30] What was interesting about the ad was that journalism experience was touted as a plus and not a main requirement of the job. A background in quantitative studies such as statistics, data science, machine learning, deep learning, epidemiology, and coding was listed as the required

29 "Dissecting racial bias in an algorithm used to manage the health of populations"; *Science*; Oct. 25, 2019.
30 "Computational Journalist — ProPublica"; RemoteLi / ProPublica job board; 2025.

qualification for the position. Computational journalists are tasked with looking for bias wherever it hides, ferreting it out, and writing about it.

An auto insurance study led by a ProPublica[31] computational journalist examined neighborhoods in California, Illinois, Texas, and Missouri for differences in auto insurance being charged in white versus non-white communities, and the results revealed that members of non-white communities pay a disproportionately higher level of premiums than what is paid in white communities for the same level of risk. The ProPublica journalists aggregated risk data by zip code from the insurance departments of California, Illinois, Missouri, and Texas. They then compared that information with liability insurance premiums charged by some of the largest companies by market share in the four states. They created a metric called average loss, defined as average dollars paid out for liability claims by insurers divided by the number of cars insured by insurers.

Modeling data was obtained from Quadrant Information Services and S&P Global Inc. Quadrant provided premium quotes, and S&P provided rate filing manuals. The loss data was collected from insurance departments. For the analysis, the focus was on a single profile of consumer: a 30-year-old female safe driver with a bachelor's degree, excellent credit, no accidents or moving violations, and purchasing standard coverage. It was a single variable model with average loss as the only predictor; the target variable was the liability premium charged. They regressed average loss against liability premium using spline methodology. It is important to note that loss ratios do not appear to have been considered in this analysis, which may be a flaw, as loss ratios are a key measure of risk. Actuaries need to identify flaws in analyses such as these so that people do not draw the wrong conclusions from data. Insurance companies should ensure there is actuarial oversight of their models for this reason, which is a regulatory concern.

The 2019 General Accountability Report on the Benefits and Challenges Presented by Innovative Uses of Technology[32] provides evidence of regulatory concern regarding data scientists performing actuarial analyses. The authors of the report emphasized, "The use of AI to create underwriting models for determining premium rates can make it challenging for insurers to ensure that factors prohibited by regulation (such as race) are not used in models. Such models are often developed by data scientists who, unlike actuaries, may not fully understand insurance-specific requirements."[33]

31 "Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk"; ProPublica; April 5, 2017.
32 "Insurance Markets: Benefits and Challenges Presented by Innovative Uses of Technology"; U.S. Government Accountability Office; June 7, 2019.
33 Ibid.

It is important to have professionals who can be held accountable by professional standards working on modeling teams. Actuarial work is guided by the actuarial standards of practice (ASOPS), which provide guidance for the conduct of their work, and should be included on data science teams to ensure work products can withstand actuarial scrutiny.

NY Hiring Example—Loss of Opportunity, Economic Loss

The development of hiring platforms and online job applications has led to a significant increase in job applications, which has prompted hiring managers to lean on automated processes to sort candidates. These automated processes are predictive models, with AI automatically scanning resumes for key words that are predictive to the quality of future hires. These processes are continually being questioned as potentially flawed. Cathy O'Neil defines a model as a Weapon of Math Destruction[34] if it meets the following criteria: (1) the model is opaque, (2) the model is (potentially) unfair, and (3) the model is used at scale and has the possibility of repeating and amplifying unconscious human discriminations and biases. In chapter 6 of the book, she discusses how these models are used in hiring practices. This process could be considered an example of a weapon of math destruction. Local Law 144 attempts to address the second criteria and mitigate the risk of discrimination: It forces companies using these models to perform a bias audit.

Automated Employment Decision Tools (AEDT)[35] [36]

"Local Law 144 of 2021 regarding automated employment decision tools ("AEDT") prohibits employers and employment agencies from using an automated employment decision tool unless the tool has been subject to a bias audit within one year of the use of the tool, information about the bias audit is publicly available, and certain notices have been provided to employees or job candidates."

New York passed this law[37] [38] to prevent discrimination in the use of AI in HR decisions (e.g., hiring, promotion). The law is based on statistical parity metrics. The Academy provides definitions of the statistical parity metrics in the August 2023 issue brief, *Discrimination: Considerations for Machine Learning, AI Models, and Underlying Data*.[39] The law also defines notification and external audit requirements and provides for penalties of up to $500 per violation.

34 O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
35 Automated Employment Decision Tools (AEDT); New York City Department of Consumer & Worker Protection; 2025.
36 Automated Employment Decision Tools (Int. 18942020); New York City Council; Dec. 11, 2021.
37 Automated Employment Decision Tools (Int. 18942020); New York City Council; Dec. 11, 2021.
38 Ibid.
39 *Considerations for Machine Learning, AI Models, and Unfair Discrimination;* American Academy of Actuaries; August 2023.

Regulatory Responses to Algorithmic Bias

The oversight of AI systems is a regulatory priority, and regulators are devising tools to aid them in this quest. Colorado is the first state to enact a law requiring insurance companies to demonstrate that their use of External Consumer Data and Information Sources (ECDIS) in algorithms and predictive models is not harming consumers relative to access and affordability.[40] On July 6, 2021, Governor Polis signed Senate Bill (SB) 21-169 into law, and insurers are required to comply with the law. The law prohibits unfair discrimination on the basis of race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression in any insurance practice, including marketing, underwriting, pricing, utilization management, reimbursement methodologies, and claims management related to the transaction of insurance. It is important to note that New York, Texas, and California have also developed regulatory tools to regulate the use of AI systems by insurers. Other states are following suit.

The NAIC Model Bulletin

In December 2023, the NAIC adopted the NAIC Model Bulletin: Use of Artificial Intelligence Systems by Insurers.[41] The purpose of the bulletin is to provide state regulators with a tool to evaluate how insurers will govern their use of AI systems. The bulletin defines an AI system as "a machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, content (such as text, images, videos, or sounds), or other output influencing decisions made in real or virtual environments. AI Systems are designed to operate with varying levels of autonomy."[42] The bulletin addresses AI principles, definitions, board of directors oversight, senior management accountability, governance, risk management and internal controls, third-party AI systems and data, documentation and reporting, internal audit procedures, and testing. As of this writing, twenty-four states have adopted the bulletin. The NAIC tracks state adoptions on its website.[43]

40 SB21-169: Restrict Insurers' Use Of External Consumer Data; Colorado General Assembly; July 6, 2021.
41 Use of Artificial Intelligence Systems by Insurers; National Association of Insurance Commissioners; Dec. 4, 2023.
42 Ibid.
43 Use of Artificial Intelligence Systems by Insurers Map; National Association of Insurance Commissioners; Oct. 31, 2025.

CO Law and the NAIC Model Bulletin

Colorado's law and the NAIC model bulletin share common themes across purpose, scope, applicability, enforcement, approach, and governance structure. Both regulatory tools have the goal of addressing unfair discrimination using a risk-based approach. (SB) 21-169 is specific to unfair discrimination with respect to race. The bulletin applies to all protected classes and applies to a broader array of issues. Both address governance and risk management.

Colorado's law and the bulletin are focused on preventing discriminatory practices across all insurance operations, including underwriting, marketing, and claims administration. The Colorado law defines insurance practices as marketing, underwriting, pricing, utilization management, reimbursement methodologies, and claims management. The bulletin references the insurance product life cycle and defines it as including product development, marketing, sales, and distribution, underwriting and pricing, policy servicing, claim management, and fraud detection.

The Colorado law carves out a specific component of AI, namely ECDIS and models using ECDIS, while the bulletin uses the term "AI Systems" and defines them as machine-based systems that can generate predictions, recommendations, content, and other output that may be used in decision-making.

The bulletin serves as general guidance for states in their development of an AI governance and risk management framework. States may adopt the bulletin in whole or in part and make modifications as appropriate. The Colorado law requires companies operating in Colorado to comply with it as it is written.

The Colorado law and the bulletin require documentation regarding the use of data and models for compliance. Under the bulletin, documentation is suggested for governance, risk management controls, and internal audit to ensure the accountability, fairness, and transparency of AI systems throughout its life cycle.

Both the Colorado law and the bulletin aim to protect consumers from unfair discrimination. The bulletin does not define unfair discrimination, but the Colorado law defines and uses the term "unfairly discriminate" with respect to protected classes, including race. The bulletin more generally discusses that an AI system should be designed to mitigate adverse outcomes, which would presumably include adverse outcomes affecting protected classes. They differ with respect to explicitly mentioning the NAIC AI principles, enforcement, and a testing regimen. The bulletin does not refer to enforcement, because the states would have authority over enforcement, and enforcement could differ by state. States would also have authority over the type of testing that could reasonably be implemented within their jurisdictions. While the Colorado law does not explicitly mention the NAIC AI principles, these principles were adopted by the Executive Committee in August 2020.

Conclusion

Algorithms have become a part of almost every aspect of life in the 21st century, creating a necessity to not only understand them but also to develop ways to mitigate the impact of algorithmic bias. The consequences of algorithmic bias can be far-reaching and profoundly impactful. For example, biased algorithms can lead to discriminatory outcomes in areas such as hiring, criminal justice, and health care, perpetuating existing inequalities and exacerbating social divides. Moreover, algorithmic bias can undermine the trust and credibility of these algorithms, hindering their adoption and limiting their potential to drive innovation and progress.

Addressing algorithmic bias requires a multifaceted approach, beginning with raising awareness of the problem and its potential consequences among stakeholders, including developers, policymakers, and the general public. This can be achieved through education, training, and public discourse, fostering a shared understanding of the challenges and opportunities associated with algorithmic decision-making.

Efforts focused on improving the quality and representativeness of the data used to train algorithms are another step toward limiting algorithmic bias. This can involve collecting more diverse and inclusive datasets, as well as employing techniques such as data augmentation and synthetic data generation to mitigate the effects of underrepresentation and imbalance.

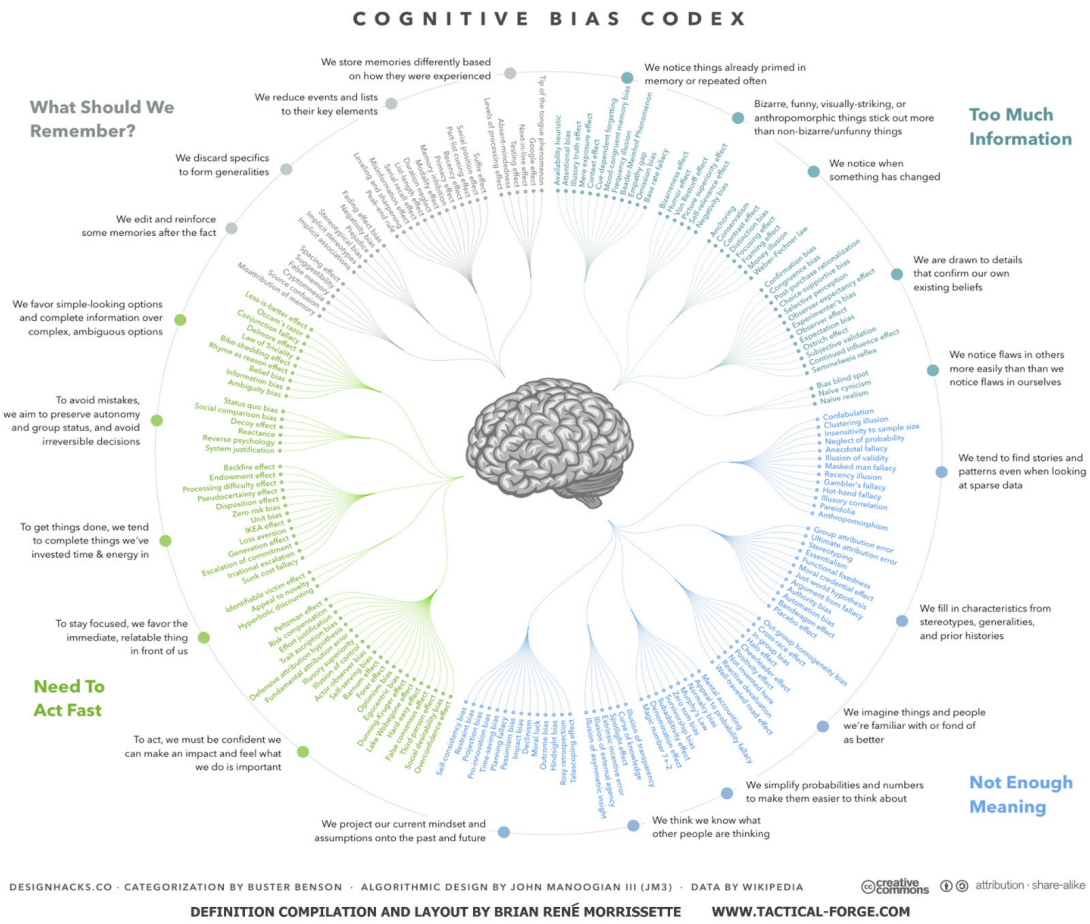
The design of algorithms requires careful and thoughtful audit processes to identify and mitigate potential sources of bias. This can involve the use of fairness metrics and evaluation frameworks, as well as the adoption of best practices in algorithmic development, such as transparency, explainability, and accountability. Actuaries can also develop and use algorithms to help create societal benefits by exploring how to employ this technology for preventing accidents, mitigating losses, encouraging healthier and safer habits, and investigating ways to avoid or overcome hazardous conditions, with the goal of creating safer and healthier communities.

Establishing robust regulatory frameworks and standards to govern the development and deployment of algorithmic driven systems will be a crucial step in reducing the impact of algorithmic bias. This can involve the creation of independent oversight bodies, the establishment of legal safeguards, and the implementation of auditing and certification processes to ensure that systems are designed, developed, and operated in a manner that is fair, transparent, and accountable.

Algorithmic bias is a complex and multifaceted issue that requires a concerted and collaborative effort to address. By its nature, algorithmic bias may never be entirely eliminated or be able to be discerned. However, by raising awareness, improving data quality, refining algorithmic design, and establishing robust regulatory frameworks, actuaries can work toward limiting the impact of algorithmic bias and developing systems that are fairer and more equitable across diverse populations.

Appendix A: Cognitive Bias Codex

More information on the codex can be found at [The Cognitive Biases List: A Visual Of 180+ Heuristics](#) and [Cognitive Bias Codex](#) (Wikimedia Commons).



1850 M STREET NW, SUITE 300, WASHINGTON, D.C. 20036

202-223-8196 | **ACTUARY.ORG**