

LAW360

# AI's Baked-In Bias: What To Watch Out For

November 7, 2023

By Jonathon Talcott and Jonathan Hummel

On Oct. 30, the Biden administration announced a sweeping executive order on safe, secure and trustworthy artificial intelligence.

The order “establishes new standards for AI safety and security, protects Americans’ privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.”

Section 7 of the order is dedicated to advancing equity and civil rights. Specifically, the order grants broad powers to the attorney general to “address civil rights and civil liberties violations and discrimination related to AI.”

This is a direct acknowledgment of the risk of baked-in bias in AI systems. While the order only includes mandates governing the operation of the federal government, the order signals to legal practitioners the rising importance of using caution and careful consideration when choosing to implement an AI system to guard against such bias.


## **AI Bias: Historical Perspective**

One of the key drivers behind federal action is the recognition that AI systems are susceptible to various forms of bias.

At a very high level, an AI systems include data, training protocols, one or more models (e.g., a large language model or an image classification model), an inference process, and a monitoring and feedback process.

This is of course oversimplified, but it’s important to note that each of these components — and others not explicitly listed — are potential sources for bias. For example, AI systems are vulnerable to bias stemming from data, algorithms and human training.

Research has consistently demonstrated that AI can unintentionally perpetuate inequalities and reinforce harmful stereotypes, particularly when it is deployed in sectors rife with historical discrimination.[1]



For example, most systems are trained on huge amounts of historical data that may stretch decades or longer into the past, e.g., housing records, wage, employment and labor records, etc.

If there were discriminatory practices or bias in those sectors in the past, that bias may show up in the AI model trained on that historical data.

As highlighted by Olga Akselrod, senior staff attorney in the racial justice program at the American Civil Liberties Union:

[because] AI is built by humans and deployed in systems and institutions that have been marked by entrenched discrimination — from the criminal legal system, to housing, to the workplace, to our financial systems..., [b]ias is often baked into the outcomes the AI is asked to predict. Likewise, bias is in the data used to train the AI — data that is often discriminatory or unrepresentative for people of color, women, or other marginalized groups — and can rear its head throughout the AI’s design, development, implementation, and use.

For example, the Correctional Offender Management Profiling for Alternative Sanctions system for predicting risk of reoffending was found to predict higher risk values for black defendants — and lower for white ones — than their actual risk.

In another example, Google’s Ads tool was found to serve significantly fewer ads for high paying jobs to women than to men.[2] We know bias is baked-in, but holding algorithms accountable is difficult.[3]

## **What’s a Lawyer To Do?**

As AI becomes ubiquitous in business and government, legal professionals will find themselves at the forefront of ensuring compliance and risk management. This article discusses some key considerations for legal professionals.

In a nutshell, because AI is increasingly integral to businesses and institutions, legal professionals will need to thoroughly vet AI systems, including data and sources, data gathering and cleaning protocols, algorithms and AI training methods.

Understanding the potential biases and sources baked into these systems — and their various components — will be critical in advising clients.

One overarching theme for the practitioner is to understand that, due to the complexity of the computer science technologies and datasets at play, the practitioner should not shy away from working with or employing data scientists or other technical experts to better understand the subject matter.

## **What Are the Legal Implications?**

The lawyer’s first job is to understand the legal landscape in which the client is operating. The lawyer should identify and review any rules, regulations, laws or court cases related to artificial intelligence that might affect the client’s business or business partners.

This, of course requires an intimate understanding the client’s business and practices. Currently, the National Conference of State Legislatures reports that in 2023, “at least 25 states, Puerto Rico and the District of Columbia introduced artificial intelligence bills, and 15 states and Puerto Rico adopted resolutions or enacted legislation.”[4]

For example, California A.B. 1502 “prohibits a health care service plan or health insurer from discriminating on the basis of race, color, national origin, sex, age, or disability through the use of clinical algorithms in its decision-making.”[5]

In Illinois, the Anti-Click Gambling Data Analytics Collection Act “provides that no entity that operates a remote gambling platform or a subsidiary of the entity shall collect data from a participant with the intent to predict how the participant will gamble in a particular gambling or betting scenario.”[6]

## **Data**

One of the most significant contributors to AI bias is training data. AI systems learn from historical training data, which often reflects biases.[7]

For example, if historical housing data was generated during a period when discriminatory practices were prevalent in the market, e.g., systemic housing decisions disadvantaged one group, an AI trained on the historical housing data may perpetuate those biases in current housing decisions.

For the lawyer, mitigating risks related to biased data starts with thoroughly understanding your client’s business, industry and data practices.

What data does the client collect and use? How is it processed and stored? Is the data cleaned or otherwise processed before being used? What policies and procedures are in place that govern the collection and use of data?

In order to avoid biased data the lawyer should help their client draft and implement policies that, for example, require documentation and evaluation of any proposed data sources, and require that any data used for training an AI system is gathered from diverse and representative datasets while also reflecting the intended application’s target population.

The policies should also govern data preprocessing procedures to remove or mitigate biases present in training data, including biases related to race, gender, age and other sensitive attributes. Further, AI data policies should establish clear guidelines for handling missing data and outliers to prevent or reduce their influence on model training.

Lawyers should advise their clients to maintain clear documentation of the entire data pipeline, including data sources, preprocessing steps, model architecture and evaluation results. Best practices call for documentation that includes information on how fairness and bias considerations were integrated into data gathering and preprocessing.

## **Algorithms**


Further, the algorithms themselves can be sources of bias.

As discussed above, the algorithm can inherit bias found in data, but the algorithm can also exhibit bias by virtue of design choices made during development.

For example, suppose a financial institution uses an AI-driven credit scoring algorithm to determine loan eligibility for applicants. While the training data is carefully curated to avoid bias and includes a diverse sample of applicants, the algorithm on its own, may by intentional or inadvertent design choice, place too much emphasis on an applicant’s age when making credit decisions.

While age is a legally protected characteristic under anti-discrimination laws, the algorithm has not been programmed to treat age as a neutral factor. Consequently, the algorithm may systematically favor younger applicants over older ones, even when all other factors, such as income, credit history and employment status, are equal. In this case, the bias is not a result of biased training data but rather a flaw in how the algorithm assigns importance to different features.

Such inherent biases can lead to discrimination and must be identified and rectified through careful examination of the algorithm’s rules and decision criteria.



Lawyers should work with clients to draft policies and procedures around algorithmic fairness including developing fairness metrics that identify bias in the model. A simple example of a fairness metric is the disparate impact metric.

The metric compares the proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group. The calculation is the proportion of the unprivileged group that received the positive outcome divided by the proportion of the privileged group that received the positive outcome.

The industry standard is a four-fifths rule: If the unprivileged group receives a positive outcome less than 80% of their proportion of the privileged group, this is a disparate impact violation.[8] However, you may decide to increase this for your business, and a “Disparate Impact Remover” has been proposed to address just this issue.[9]

## Humans

Humans are another source of bias in AI systems.

Currently, many of the large and powerful AI systems are still designed by humans, though that may be changing. Human can introduce bias into the AI system in almost every stage of development and training from data collection and data cleaning to actual training parameters.

For example, when gathering and curating the datasets used to train AI models, individuals involved may inadvertently introduce bias by selecting data that reflects their own perspectives and biases. For instance, if the dataset used to train a housing decision-making AI system predominantly includes historical housing decisions based on biased criteria, such as racial or socioeconomic factors, the AI model is likely to learn and perpetuate these biases in its recommendations.

Bias can also get baked into AI systems during the labeling process, which involves annotating data for supervised learning. Data labeling involves labeling the data that will be used to train the model. For example, a human may review photos of cars, bicycles, planes, and buses and label the images as containing those objects — think CAPTCHA.[10]


During model development, humans make decisions about the architecture and hyperparameters, which can inadvertently magnify existing biases.[11][12] For example, the choice of certain training objectives or optimization functions can amplify biases present in the data. Moreover, the evaluation metrics used to assess the model’s performance may not fully capture the fairness and ethical considerations, leading developers to prioritize performance over fairness, thereby reinforcing biases.

Further, annotators — who are humans — may have their own subjective viewpoints and interpretations, leading to inconsistencies or biases in the labeled data. Additionally, if labeling guidelines are unclear or biased themselves, annotators may unintentionally perpetuate such biases.

To address the risk of human introducing bias into the AI systems, the lawyer should start by thoroughly understanding how the AI system works, including its data sources, algorithms and decision-making processes. This understanding will be crucial in identifying potential sources of bias and assessing the extent to which human bias may have been introduced.

Lawyers should also help their clients assess their legal and ethical obligations concerning bias in AI systems by reviewing relevant laws and regulations, industry standards and ethical guidelines. It’s essential to determine whether the AI system’s bias may lead to legal liabilities or ethical concerns.

Lawyers can advise their clients to conduct bias audits and assessments of the AI system.



This involves examining the data used for training, the labeling process, and the decision-making criteria to identify and quantify any bias. External auditors or experts may be engaged for an impartial evaluation.

Lawyers should encourage clients to maintain comprehensive records of their AI system's development and decision-making processes. Transparency and documentation can help demonstrate due diligence and compliance with legal and ethical standards.

## **Ongoing Efforts**

Bias in AI is not static; it can evolve over time as the AI system interacts with users and adapts to changing data.

Continuous monitoring and evaluation are necessary to detect and mitigate evolving biases. Given the interdisciplinary nature of the field, legal professionals may need to collaborate closely with technology experts, data scientists and AI specialists.

Effective communication between legal and technical teams will be essential to ensure that AI systems promote equality and social justice, rather than perpetuate historical biases. Lawyers should stay up to date on the latest trends and also educate themselves about the basics of artificial intelligence.

The field of artificial intelligence is particularly difficult to pin down because it evolves faster than rules and regulations can be drafted and implemented. Thus, it will be essential to stay updated on evolving case law and legal interpretations related to AI bias.

## **Conclusion**

Addressing baked-in bias in AI systems requires increased awareness, transparency and ethical considerations at every stage of development and deployment.

Developers must strive for diverse and representative datasets, clear labeling guidelines, and ongoing bias audits to mitigate the inadvertent introduction of human bias into AI systems, ultimately working toward more fair and equitable AI technologies.

This involves evaluating the data sources, algorithms and training processes to identify and rectify bias.

Legal experts will need to help organizations establish guidelines that align with both legal requirements and broader ethical considerations. Training programs and seminars on AI ethics, bias mitigation, and compliance with civil rights laws will be essential for organizations to navigate the evolving legal landscape.

**Jonathon A. Talcott** is a partner and **Jonathan P. Hummel** is an associate at Ballard Spahr LLP.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] Wach, et al., 2023.

[2] Ntoutsis, et al “Bias in data-driven artificial intelligence systems — An introductory survey” Wiley Interdisciplinary Reviews (2/3/2020).

[3] Janssen, Marijn; Kuk, George; “The challenges and limits of big data algorithms in technocratic governance” Government Information Quarterly. 2016, 371-377.

[4] <https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation>.

[5] <https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation>.

[6] <https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation>.

[7] Leavy S, O’Sullivan B, Siapera E. Data, Power and Bias in Artificial Intelligence, in AI for Social Good Workshop. 2020.

[8] <https://towardsdatascience.com/ai-fairness-explanation-of-disparate-impact-remover-ce0da59451f1>.

[9] Disparate Impact is a metric to evaluate fairness. It compares the proportion of individuals that receive a positive output for two groups. Disparate Impact Remover is a pre-processing technique that edits values, which will be used as features, to increase fairness between the groups. (<https://arxiv.org/abs/1412.3756>)

[10] CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) is a type of security measure known as challenge-response authentication

[11] The architecture of an AI model refers to its fundamental structure and design. It outlines the way in which the model is organized, including the arrangement of its neural network layers (in the case of deep learning models) and the connections between them.

[12] Hyperparameters are settings or configurations that are not learned from the data but are set by the developer before training the model. These parameters control various aspects of the training process, affecting how the model learns. Examples of hyperparameters include the learning rate (which controls how quickly the model adapts to the data), the batch size (how many data samples are processed in each training iteration), the number of training epochs (how many times the model sees the entire training dataset), and regularization strength (to prevent overfitting).

Reprinted with permission from Law360. © 2023 ALM Media Properties, LLC.  
Further duplication without permission is prohibited. All rights reserved.