PERKINSCOIE

Artificial Intelligence, Machine Learning and Robotics

Combating Bias in Artificial Intelligence and Machine Learning Used in Healthcare

SEPTEMBER 2020

CHARLYN HO
COUNSEL & LEAD AUTHOR
+1.202.654.6341
CHo@perkinscoie.com

MARC MARTIN
PARTNER
+1.202.654.6351
MMartin@perkinscoie.com

SARI RATICAN
SENIOR COUNSEL
+1.310.788.3287
SRatican@perkinscoie.com

DIVYA TANEJA
ASSOCIATE
+1.332.223.3935
DTaneja@perkinscoie.com

D. SEAN WEST
COUNSEL
+1.206.359.3598
DWest@perkinscoie.com

Artificial Intelligence, Machine Learning and Robotics

Combating Bias in Artificial Intelligence and Machine Learning Used in Healthcare

Artificial intelligence (AI) has the promise to revolutionize healthcare through the use of machine learning (ML) techniques to predict patient outcomes and personalize patient care, but this use of AI carries the risk of what is described as "algorithmic bias" affecting such outcomes and care, even unintentionally.

AI seeks to enable computers to imitate intelligent human behavior, and ML, a subset of AI, involves systems that learn from data without relying on rules-based programming. AI has the power to unleash insights from large data sets to more accurately diagnose and evaluate patients and dramatically speed up the drug discovery process. However, despite its promise, the use of AI is not without legal risks that can generate significant liability for the developer and/or user of ML algorithms. Among other risks, the development and use of ML algorithms could result in discrimination against individuals on the basis of a protected class, a potential violation of antidiscrimination and other laws. Such algorithmic bias occurs when an ML algorithm makes decisions that treat similarly situated individuals differently where there is no justification for such differences, regardless of intent. Absent strong policies and procedures to prevent and mitigate bias throughout the life cycle of an ML algorithm, it is possible that existing human biases can be embedded into the ML algorithm with potentially serious consequences, particularly in the healthcare context, where life and death decisions are being made algorithmically.

In this article, we discuss the legal concerns arising from the potential bias manifestations in ML algorithms used in the healthcare context under U.S. law. We briefly describe the legal landscape governing algorithmic bias in the United States and offer some emerging tools that build on existing recommended best practices, such as adversarial debiasing and the use of synthetic data for detecting, avoiding, and mitigating algorithmic bias.

AI AND ML IN HEALTHCARE

ML techniques include supervised learning (a method of teaching ML algorithms to "learn" by example) as well as deep learning (a subset of ML that abstracts complex concepts through layers mimicking neural networks of biological systems). Trained ML algorithms can be used to identify causes of diseases by establishing relationships between a set of inputs, such as weight, height, and blood pressure, and an output, such as the likelihood of developing heart disease. As an example of the power of training ML algorithms on large amounts of data, a group of scientists trained a highly accurate AI system using electronic medical records of nearly 600,000 patients to extract clinically relevant data from huge data sets and associate common medical conditions with specific information. The ML algorithm was then able to use a patient's symptoms, history, lab results, and other clinical data to automatically diagnose common childhood conditions, enhancing physicians' ability to accurately diagnose illness as compared to human-only methods.

Unfortunately, ML algorithms developed in the healthcare context are not immune from algorithmic bias. In one pre-pandemic study, an algorithm used by UnitedHealth to predict which patients would require extra medical care favored white patients over black patients, bumping up the white patients in the queue for special treatments over sicker black patients that suffered from far more chronic illnesses. Race was not a factor in the algorithm's decision-making, but race correlated with other factors that affected the outcome. The lead researcher of this study—which motivated the New

York Department of Financial Services (DFS) and Department of Health (DOH) to write a letter to UnitedHealth inquiring about this alleged bias—stated that the "algorithm's skew sprang from the way it used health costs as a proxy for a person's care requirements, making its predictions reflect economic inequality as much as health needs." The DFS and DOH were particularly troubled by the algorithm's reliance on historical spending to evaluate future healthcare needs and stated that this dependence poses a significant risk of conflicts of interest and also unconscious bias. The DFS and DOH cited studies documenting the barriers to receiving healthcare that black patients suffer. Therefore, utilizing medical history in the algorithm, including healthcare expenditures, is unlikely to reflect the true medical needs of black patients, because they historically have had less access to, and therefore less opportunity to receive and pay for, medical treatment.

At the time of writing this article, AI is being utilized in the fight against the COVID-19 pandemic, including to triage patients and expedite the discovery of a vaccine. For example, researchers have developed an AI-powered tool that predicts with 70% to 80% accuracy which newly infected COVID-19 patients are likely to develop severe lung disease. The U.S. Centers for Disease Control and Prevention has leveraged Microsoft's AI-powered bot service to create its own COVID-19 assessment bot that can assess a user's symptoms and risk factors to suggest a next course of action, including whether to go to the hospital. While these advances will benefit many patients, they do not resolve the concerns relating to algorithmic bias and its repercussions for certain groups. Use of AI to triage COVID-19 patients based on symptoms and preexisting conditions can perpetuate well-researched and well-documented preexisting human prejudice against the pain and symptoms of people of color and women. Data on COVID-19 shows disparities based on race and socioeconomic status, including substantially higher mortality rates among certain racial groups. The risk of algorithmic bias in ML algorithms designed to combat COVID-19 is heightened because the data is not equally distributed across age groups, race, and other patient characteristics. Without representative data, there is a higher risk of bias. However, as discussed in more detail below, if such AI tools were debiased using adversarial debiasing, synthetic data, or some other solution, an AI tool could triage patients more appropriately based on their symptoms so that they receive the healthcare that they need.

OVERVIEW OF THE U.S. LEGAL LANDSCAPE RELATING TO ALGORITHMIC BIAS IN HEALTHCARE

While U.S. laws do not yet specifically address AI bias, existing federal and state laws may make algorithmic discrimination and bias unlawful, regardless of intent. For example, Section 1557 of the Affordable Care Act (ACA) prohibits any healthcare provider that is receiving federal funds to refuse to treat—or to otherwise discriminate against—an individual based on protected classifications such as race, national origin, or sex. In addition to healthcare-specific laws, several states specifically prohibit discrimination in hospitals or clinics based on protected classifications.

Although different laws apply different standards for when unlawful discrimination exists, generally, an antidiscrimination claim requires establishing either disparate treatment or disparate impact. Disparate treatment can be established by either facially disparate treatment (such as explicit race classifications) or intentional, but facially neutral, discrimination (such as zip code classification with the intent that zip codes serve as a rough proxy for race). Disparate

impact can be established by showing that a facially neutral policy or practice disproportionately affects a group sharing a protected trait, such as a religion.

We surmise that it is more likely that AI developers working on healthcare-related ML algorithms would design a facially neutral algorithm that contains undetected biases resulting in disparate impacts than an algorithm that intentionally treats people differently on an unlawful basis. In fact, a 2014 White House study on Big Data concluded that it is very rare for algorithmic bias to arise from intentional disparate treatment; instead, algorithmic bias is often caused by poor training data (data that is either inaccurate, out of date, or nonrepresentative of the population) and unintentional perpetuation of historical biases. Therefore, it is important for AI developers and users to keep abreast of developments in this complex area of law.

Continuing to use the ACA as an example of how algorithmic bias may be treated under U.S. law, at least some ACA Section 1557 claims can be established under a theory of disparate impact, but the limits on what types of such claims can rely on disparate impact alone is not yet well defined. Instead of providing its own rules on discrimination, Section 1557 applies four preexisting race, sex, age, and disability discrimination laws to federally subsidized health programs. Not all of these underlying laws allow plaintiffs to bring claims based on disparate impact, and courts are divided on whether a single standard should govern all Section 1557 claims or whether the differing standards from the underlying laws should be applied. If the differing standards from the underlying laws were applied, then different types of discrimination would be treated differently under Section 1557. Under such an interpretation, a Section 1557 race discrimination claim must allege disparate treatment, but a Section 1557 age, disability, or sex discrimination claim could allege disparate treatment or disparate impact. This reliance on differing standards appears to be the emerging majority position, but at least one federal court has ruled that Congress intended to create a new cause of action with a single standard passing Section 1557.

Further complicating the legal analysis, establishing either disparate treatment or disparate impact alone is rarely enough to establish liability. After a prima facie discrimination analysis under either theory, the inquiry generally shifts to whether there is enough justification for the discriminatory practice. The standards for when a practice will be considered justified vary for different statutory schemes and by discrimination type, but even facially discriminatory practices are permissible if a justification is found to be sufficient. Only when the justification is insufficient is discrimination unlawful. Medical decision-making, algorithmic or otherwise, will often be based on certain protected classifications. However, such classifications would not constitute unlawful discrimination when the classifications are relevant to medical outcomes, such as the ways COVID-19 may affect men and women differently.

Given the complexities of the antidiscrimination legal landscape and differing interpretations on how laws that predate the advent of AI and ML govern algorithmic bias, both AI developers and users are encouraged to work closely with their legal teams to evaluate the legal risks associated with their particular AI/ML use case.

Artificial Intelligence, Machine Learning and Robotics

Combating Bias in Artificial Intelligence and Machine Learning Used in Healthcare

CURRENT BEST PRACTICES FOR COMBATING ALGORITHMIC BIAS

Developing ML algorithms often requires vast quantities of training data for the ML algorithm to "learn." For example, if a health insurer wants to use an ML algorithm to estimate healthcare insurance costs, data scientists can train the ML algorithm on historical healthcare claims data for the ML algorithm to "learn" what variables affect healthcare insurance premiums so that it can predict healthcare insurance costs. In other words, ML algorithms use training data to learn how to recognize and apply patterns to make accurate predictions when presented with new data.

Current best practices for combating algorithmic bias center around avoiding the introduction of bias at each stage of the ML algorithm's development. For example, the Federal Trade Commission's 2016 report entitled "Big Data: A Tool for Inclusion or Exclusion?" encourages companies to, among other things, consider four questions regarding their algorithms:

1. How representative is the data set? If data sets are missing information from populations, take appropriate steps to address the problem. This is more simply known as the "garbage in, garbage out" problem.
2. Does your data model account for biases? Ensure that hidden bias is not having an unintended impact on certain populations.
3. How accurate are your predictions based on big data? Correlation is not causation. Balance the risk of using the results from big data, especially where policies could negatively affect certain populations. Consider human oversight for important decisions, such as those implicating health, credit, and employment.
4. Does your reliance on big data raise ethical or fairness concerns? Consider using big data to advance opportunities for underrepresented populations.

Although these recommended best practices seem obvious and straightforward, implementing them is easier said than done. Data scientists cannot easily remove biases that human beings infuse, often unconsciously, into training data and, therefore, the ML algorithm. There are few practical solutions to actually eliminate such unintended bias from an ML algorithm without impairing its efficacy.

Part of the challenge of eliminating algorithmic bias is that bias may be introduced into an ML algorithm at many different points in the development cycle, and eliminating bias requires constant vigilance throughout the development and deployment of the ML algorithm. For example, training data can be infected by historical bias (bias already existing in the world that is reflected in the data collection process, even with perfect sampling and feature selection), representation bias (bias resulting from how the relevant population is defined and sampled), measurement bias (bias resulting from the way features are selected and measured), and coded bias (bias introduced by the people developing the algorithms).

Even when care is taken to root out bias from data sets by not relying on protected characteristics (e.g., race, color, sex or gender, religion, age, disability status, national origin, marital status, or genetic information), sometimes the

algorithm uses variables that function as proxies for protected classes. For example, zip codes and/or language may correlate closely with race. Additionally, while it may seem logical to "blind" the ML algorithm to protected characteristics by omitting this variable from the training data, this mitigation technique may in and of itself result in bias. Two data scientists observed in their research on AI-based sentencing algorithms that women are less likely to reoffend than men in many jurisdictions. Therefore, blinding the ML algorithm to gender may result in judges being less likely to release female defendants before trial even though they have a lower chance of reoffending and may make it harder for companies to detect, prevent, and eliminate bias on exactly that criterion.

Because the data scientists who develop ML algorithms may not be attuned to the legal considerations of algorithmic bias, both developers and users of ML algorithms should partner closely with their legal team to mitigate potential legal challenges arising from developing and/or using ML algorithms, particularly when data as sensitive as healthcare data is involved.

EMERGING TOOLS TO COMBAT ALGORITHMIC BIAS

1. Based on our experience counseling businesses seeking to eliminate algorithmic bias, it is often challenging to eliminate bias from the training data for a variety of reasons. First, companies may not have policies and procedures to detect and test for algorithmic bias and may be unaware of such bias, particularly if they did not develop the ML algorithm in-house. Second, given the vast amounts of data needed to train ML algorithms, companies developing ML algorithms may need to source training data from third-party sources and therefore may not have control or influence over the initial collection of that training data. As a result, even if a company identifies that its training data contains bias, it is not clear how it can rectify this issue to avoid algorithmic bias. Third, U.S. privacy and other laws may limit companies' ability to source representative data, whereas countries that have different privacy laws and norms (like China) may find it easier to develop and train ML algorithms on diverse training data. Finally, in the absence of the ability to rectify a biased ML algorithm, companies are faced with the unappealing choice of forging ahead with the ML algorithm (knowing that there is a risk of liability as a result of the bias) or starting over with a different ML algorithm.
2. AI developers can attempt to remove bias in the training data, in the trained model itself, or in the predictions (which are generally known as pre-processing, in-processing and post-processing bias mitigation techniques). Below, we describe nascent methods of mitigating, or even potentially eliminating, algorithmic bias that companies can consider deploying (a) in the trained model and (b) in the training data when it is impractical or impossible to remove bias from historical training data or inherent human bias introduced by the data scientists developing the ML algorithm.

Adversarial Debiasing

Adversarial debiasing is a supervised deep learning method whereby two algorithms are used to predict an output variable (e.g., organ transplant suitability) based on a given input (e.g., patient medical records) while remaining unbiased with respect to a particular protected variable (e.g., race). The first algorithm is known as the "predictor" and

simply refers to the ML algorithm that uses inputs ("X") to predict outcomes ("Y"), e.g., using an ML algorithm to predict a patient's suitability for an organ transplant based on medical records. As discussed above, predictor algorithms can perpetuate bias; for example, creditworthiness algorithms have been shown to deem men more creditworthy than women, even with all other factors being equal. This is the case even when the ML algorithm does not use gender as an input variable but relies on proxy variables (such as shopping habits) that correlate (even unintentionally) with gender.

In an ideal world, companies could still harness the power of AI trained on data that may contain bias but train the AI not to base decisions on protected variables such as race, age, and gender ("Z"). This is where the "adversary" algorithm comes in. In adversarial debiasing, the "adversary" is an algorithm used in conjunction with the predictor algorithm that is trained to predict the association of the protected variable, Z, with the output, Y. If the adversary is able to predict Z from Y (i.e., predict that an output such as a patient's need for healthcare is invalidly influenced by a protected variable, such as race, age, or gender), with everything else being equal, then there may be bias in the model. The ML model can then be trained to rely less and less on the protected variable and gradually become "debiased." When the predictor and adversarial algorithms are trained over multiple iterations, they have the potential to yield an unbiased predictive algorithm that does not significantly sacrifice accuracy. In one study on adversarial debiasing, Google and Stanford researchers demonstrated the ability to train a demonstrably less biased algorithm that still performed the task nearly as well as the original algorithm. Adversarial debiasing is still a fairly new technique, but it provides a glimmer of hope for a pragmatic tool to combat bias in AI.

Synthetic Data

Another potential tool to root out bias in AI is the use of synthetic data (i.e., artificially generated data replicating real-world statistical components). Synthetic data may hold promise as a technique for augmenting, replacing, or correcting for biases in training data. Data synthesis is an emerging data augmentation technique that creates and enables access to realistic (albeit not real) data that retains the properties of an original, real data set. To create synthetic data, an artificial neural network or other ML process learns the characteristics and relationships of the real data to generate the synthetic, yet realistic, data. To date, the most common purposes of using synthetic data have been situations where real data is expensive to collect or unavailable or there is a need to preserve/protect data subject privacy. But the process of creating a synthetic data set can also be leveraged for mitigating bias in the original, real data set. A team from IBM Research presented an illustration of how synthetic data could be employed to reduce bias in AI: when bias in AI is the result of a data set involving a privileged and an unprivileged group (for example, men and women), for every data point, a new synthetic data point can be created that has the same features, except the synthetic data point would be labeled with the other gender. These new synthetic data points together with the original data points would make up the new data set, which ideally would equally weight men and women. So, rather than trying to remove discrimination from the data set, unlike some other approaches, synthetic data instead generates a new data set that is similar to the real one and aims to be debiased while preserving data utility. Although there may be risks associated with its use, such as inaccurate or ineffectual algorithms, synthetic data holds promise as a tool to mitigate bias in AI.

Other Practical Bias Mitigation Techniques

As these solutions continue to be tested and refined, developers of AI solutions could consider other practical backstops, such as conducting regular audits of algorithms to check for bias, increasing human involvement in the design and monitoring of algorithms, and relying on cross-functional teams to pressure-test an algorithm from different perspectives. Users of AI solutions that do not have the means to change the design or development of the ML algorithm could consider contractually mitigating their liability by requiring the AI developers from whom they purchase the ML algorithm to implement antibias techniques and bias mitigation best practices and indemnify for liability arising out any unlawful discrimination or other claim caused by bias in the ML algorithm.

ACHIEVING "FAIRNESS" IN ARTIFICIAL INTELLIGENCE

The problem of how to combat unfair algorithmic decision-making has received much attention in computer science literature. Despite this academic scrutiny, little progress has been made on how to mitigate against the risk of unfairness by computational decisionmakers. In part, this lack of progress is a result of there being no universal definition of the term "fairness." Fairness is a multifaceted societal construct and often depends on individual perspective. Computer scientists have created antibias toolkits that arguably do not allow ML algorithms to achieve societal "fairness," given that term's amorphous definition, but rather aim to optimize accuracy of prediction while achieving statistical or mathematical "sameness." Despite the distinction between societal fairness and mathematical sameness, the term "fairness" is still used in computer science literature to describe antibias mitigation techniques and metrics. Several metrics for measuring computational fairness have been advanced, including demographic parity (also known as statistical parity), equality of odds, and equality of opportunity.

Under demographic parity, the likelihood of a positive outcome should be the same regardless of whether a person is in a protected group or not. With this antibias method, the ML algorithm will make predictions that are not dependent on a protected variable ("Z" in our previous example). Under equality of odds, the likelihood of true positives and false positives should be the same regardless of whether a person represents a protected variable or not. Therefore, equality of odds is satisfied if the accuracy of the ML algorithm is constant across all groups. Under equality of opportunity, the likelihood of true positives should be the same regardless of whether a person is in a protected group or not. Therefore, under equality of opportunity, individuals that represent different protected variables should have an equal chance of being classified by the ML algorithm for a positive outcome. The following examples apply these concepts to a hypothetical scenario where a race-blind ML algorithm seeks to classify individuals' need for healthcare:

- If the ML algorithm achieves demographic parity, then the percentage of white people and black people deemed to need healthcare is equal, regardless of whether one group on average needs more healthcare than the other group.
- If the ML algorithm achieves equality of odds, then no matter whether a patient is white or black, if they are sick, they have equal odds of being deemed to need healthcare, and if they are not sick, they have equal odds of being deemed not in need of healthcare. Therefore, if a higher percentage of the black patient population is sick, then a

higher percentage of black patients will be deemed to need healthcare. Note that if equality of odds is satisfied, demographic parity may not be satisfied because white patients and black patients will be deemed to need healthcare at different levels.

- If the ML algorithm achieves equality of opportunity, individuals of different races should have an equal chance of being classified as needing healthcare. Note that in contrast to equality of odds, equality of opportunity requires nondiscrimination for positive predictions (one that yields a benefit to a person) but does not require nondiscrimination for negative outcomes (one that is disadvantageous to a person). As a result, the ML algorithm does not need to be nondiscriminatory in determining whether people that are not sick are deemed not to need healthcare at an equal rate.

These are all normative concepts of fairness seeking to define standards of fairness that ML algorithms should achieve. As such, they can exist in tension with antidiscrimination laws that do not utilize fairness as a legal standard.

UNFAIRNESS VERSUS UNLAWFUL DISCRIMINATION

Apart from the problem of there being no widely held normative or legal concept of computational fairness, decisionmakers must recognize that the various prohibitions on discrimination in the United States are not aimed at promoting notions of fairness. Instead of seeking to optimize fairness, antidiscrimination laws operate as "side-constraints"—rules that limit the means by which other goals can be pursued. As side-constraints, antidiscrimination laws do not require or permit decisions based on protected classifications in the pursuit of fairness; rather, they require decisionmakers to provide adequate reasons for certain decisions that either involve disparate treatment of individuals based on protected classifications or have a disparate impact on groups sharing a protected trait.

In the United States, discrimination claims are typically similarly evaluated regardless of the direction in which benefits flow, which means that courts will not typically grant decisionmakers deference when they are making decisions based on protected classifications in an attempt to correct for past discrimination. As a result, a healthcare provider who employs techniques such as adversarial debiasing to influence an AI system in an attempt to correct for historical, representation, or measurement bias affecting a group sharing a protected trait could be subject to a discrimination claim under a disparate treatment theory because the adversarial model employed by the provider would be making decisions on the basis of a protected classification.

Therefore, before engaging in debiasing activities that could give rise to a discrimination claim, an organization will need to carefully consider and document why the debiasing activities are necessary to accurately train an algorithm.

CONCLUSION

Healthcare is just one industry among many that is being transformed by the power of AI and ML. In the context of life-and-death medical decisions, AI holds great potential to improve the quality and consistency of healthcare. This promise, however, needs to be tempered with an understanding and evaluation of the potential risks and emerging best practices to eliminate algorithmic bias.

PERKINSCOIE

Artificial Intelligence, Machine Learning and Robotics

Combating Bias in Artificial Intelligence and Machine Learning Used in Healthcare

About Perkins Coie's Artificial Intelligence, Machine Learning and Robotics (AI) Group

From concept to launch, our AI attorneys provide guidance on the development of products and services that merge digital presence, physical hardware and human-inspired intelligence. We represent providers and purchasers of AI and we routinely advise on data use, including data ownership, data privacy compliance and data security compliance, intellectual property issues associated with AI projects, portability and re-usability of AI, risks of algorithmic discrimination, and other legal issues unique to AI.

Our team is frequently recognized for our capabilities and experience that intersect with AI, such as startup company financings, commercial contracts, cloud and edge computing, litigation risk mitigation, intellectual property, labor and employment, real estate and investment management as well as the guidance we provide in connection with mapping the regulatory landscape governing the development and use of AI, IoT and blockchain technologies.

Many of our attorneys have advanced scientific or computer programming degrees that include a focus on artificial intelligence, or have previously worked as researchers or in-house counsel for companies focused on AI before it made the transition from purely scientific theory to actual, real world application. Our multidisciplinary capabilities and experiences span the wide range of machine learning and sensing technologies, including engineering, biometrics, optics, microelectronics, systems engineering, signal processing and neural networks.

CO-AUTHORS

Charlyn Ho
Counsel

Marc Martin
Partner

Sari Ratican
Senior Counsel

Divya Taneja
Associate

D. Sean West
Associate