Professional Perspective

Avoiding Human Bias in Artificial Intelligence

Heather J. Meeker and Amit Itai, O'Melveny & Myers

Bloomberg Law

Avoiding Human Bias in Artificial Intelligence

Contributed by Heather J. Meeker and Amit Itai, O'Melveny & Myers

Today, AI systems help create efficiencies and business opportunities in financial services, marketing, health care, and government. However, AI systems are only as good as they are designed to be, and lately, we have begun to understand how our human biases can creep into AI. Accordingly, lawmakers, regulators, and civil activists have begun to focus on AI biases and how they might affect our society. As they do so, they have demanded that businesses be held accountable for their use of AI. Put simply, bias in AI has now become a legal issue that companies must address.

What is Bias?

While human bias is a well-documented and widely understood concept, what exactly is AI bias?

Bias, when applied to statistical analysis, means an error introduced into sampling or testing by selecting or encouraging one outcome or answer over others. In statistics, bias does not imply human bigotry, or even bad intent–just faulty experimental design. For example, if you want to measure whether people prefer coffee or tea, it is probably not best to do your polling in a coffee shop.

But bias can be subtle and difficult to avoid. You could poll people on the street, but you might get a skewed result if you polled near a coffee shop or in a neighborhood populated with people from a culture that traditionally drinks coffee. That bias may not be evident until you understand more about the population you are measuring.

Recently, various stakeholders have started to acknowledge how much of our human biases make their way into AI. AI systems are only as good as the data we put into them. For example, the data that we use to train AI may be selected in a biased way, or the "ground truth" against which AI models are measured might have been created in a biased way. Moreover, we often do not recognize that the data is biased until after it is used to train an AI system, which makes AI bias an ongoing problem that can be difficult to undo.

Even if companies building AI systems do not intend to discriminate, the tools they use can still have discriminatory outcomes. And because software controls so much of our day-to-day lives, the result is systemic bias that can be challenging to eradicate. Being aware of the risk of bias and working to mitigate it should be a top priority for anyone designing automated systems.

As companies work to develop AI systems that we can trust, it is critical to ensure that AI algorithms and systems can be easily audited. If the AI is a "black box" that cannot be re-engineered, the only solution may be to throw it away and start over, discarding useful insights along with the biases, or avoiding the technology altogether.

But improving technological design is not enough. Companies that implement and rely on AI systems need to take proactive measures to avoid bias–whether intentional or not. Many companies today are adopting corporate AI compliance policies with a view to bias prevention and the proper use of AI. Companies should also consider preparing an AI incident plan to address and mitigate any biases as soon as they are uncovered by internal or external stakeholders.

AI Tutorial

Since renowned computer scientist John McCarthy first coined the term AI in 1955, many products and services have been touted as using AI. But what exactly is AI?

Not All 'AI' Is AI

Lately, AI has become a buzzword that is often misused. According to a survey conducted last year by London-based venture capital firm MMC Ventures, 40% of European startups that are classified as AI companies do not actually use AI in a way that is "material" to their businesses.

AI is a subfield of computer science that aims to build systems capable of gathering data and using that data to make decisions and solve problems. However, in practice, the term is used in ways that are often imprecise or even misleading. AI comes in two forms–"simple" or non-machine learning AI, and machine learning AI.

Simple AI is capable of solving a specific problem, but cannot learn by itself, and requires explicit human intervention to learn. Simple AI systems may mimic human interaction, such as troubleshooting guides or automated help desk systems, and are sometimes called expert systems.

While these systems may perform some kinds of natural language processing, they usually do not have the ability to learn and improve from the accuracy of past answers. Bias in simple AI is relatively easy to fix, because humans control each step of the process. A simple AI also will not change due to a biased data set. Discard the faulty data set and the system will work as initially planned.

On the other hand, machine learning AI means any computer-based system that observes, analyzes, and learns without human intervention. These systems can learn and improve from experience without human intervention. The key feature of machine learning AI systems is that they are iterative—they get better and more powerful as they collect and analyze additional data.

Using them at different time points will often yield different results. But as a result, it is not always possible to explain how they made their decisions. This is especially true of a specific subset of machine learning AI called deep learning. Providing data to a machine learning system usually cannot be undone, as the AI system literally re-codes itself following the analysis of the new data, and so all data provided to it must be vetted for potential bias.

AI Bias

Over the past few years, AI experts and other stakeholders have started to explore the different ways that AI bias can occur.

For example, Amazon famously pulled the plug on its AI recruitment tool because it could not stop it from discriminating against women. Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a decade. However, due to male dominance in the tech industry, most of these resumes came from men. Amazon tried to edit the programs to make them gender-neutral, but they could not ensure that the AI would not devise other ways of sorting candidates based on the existing data that could prove discriminatory. Amazon eventually disbanded the team as executives lost hope for the project.

But Amazon is not alone. There have been numerous cases of AI causing, or perpetuating, racial biases, from police profiling to where your kids go to school. Recent studies showed that AI also led to biases in financial services where credit algorithms may violate anti-discrimination laws even when they are designed not to, as well as other consumer lending-related biases. In health care, concerns have been growing regarding potential AI biases, such as misdiagnosis of melanoma in African Americans and disparities in health care based on socioeconomic factors.

Data Bias

The first and most common type of AI bias arises from the data itself. AI systems are only as good as the data we put into them.

The clearest example of AI bias stemming from data is when the data contains implicit racial or gender biases. This type of bias usually occurs when a pre-existing database is being fed to an AI in its entirety to be distinguished from situations where data is collected for the purpose of training the AI. For example, if we were to train an AI system with data from police records and prior arrests, the relevant AI system may very well learn and adopt biases that already exist in the data, such as discrimination against minorities.

In other words, even if the AI itself was designed properly but was fed with information that reflects existing societal biases, e.g., that the police in a certain city arrested more African Americans than White Americans—even when the alleged offense was identical—then the AI will become biased as well.

Similarly, if, in most tech companies, certain roles are predominantly filled by men, this gender bias will creep into any AI recruiting model that uses existing employee data. From the AI perspective, the data that it is being fed to it is the "truth," and it will try to extract insights from it. If the data is flawed, i.e., biased, the insights will be flawed as well.

Data Collection Bias

Bias can also be introduced into AI systems through data collected for the purposes of training the AI. For example, if a facial recognition algorithm is trained with data sets that mostly consist of photos of light-skinned faces rather than dark-

skinned faces, the resulting facial recognition system would unsurprisingly be worse at recognizing darker-skinned faces. This can not only include AI that produces biased results, but products that systemically underserve certain groups or communities.

In criminal justice models, oversampling certain neighborhoods because they are over-policed can lead to recording more crime, which may then result in the AI recommending more policing in those exact neighborhoods. For example, if officers patrol areas that they believe are "suspected areas" and observe new criminal acts that confirm their prior beliefs regarding the suspected neighborhoods, the newly observed criminal acts that police document as a result of these targeted patrols then feed into the predictive AI system.

This, in turn, creates a feedback loop where the model becomes increasingly confident that the locations most likely to experience further criminal activity are exactly the locations that police officers had previously believed to be high in crime.

Design Bias

It is also possible for humans to introduce bias into AI systems at the design stage. AI design includes selecting which variables and attributes the AI system will consider. For example, a credit card company may want to predict a consumer's "creditworthiness." But in order to translate such a vague concept into something that can be modeled and computed, the company must define creditworthiness within the context of a specific goal—for example, minimizing loan defaults or maximizing profit margins. In modeling creditworthiness, a company could use as inputs the customer's age, income, or number of paid-off loans.

On the other hand, in the case of a health-care model or an insurance company, inputs could include how many visits a patient makes to the ER or the pharmacy in the past year. These choices may track the implicit bias of the designers. Choosing which variables to consider and which to ignore can significantly influence a model's accuracy.

While the impact of design choices on accuracy is easy to measure, its impact on the model's bias is not. Often, such bias can only be detected and measured in hindsight, after the model has been deployed and scaled, and the "tainted" data has already been integrated into the AI system.

Legal Risks Due to AI Bias

As lawmakers, regulators, and civil rights activists have begun to focus on potential AI-related biases and how they may affect our society, AI bias can no longer be considered as a mere academic or technological issue. AI bias has now become a legal issue that companies and investors must address or deal with significant consequences later.

State and federal legislators have already started introducing laws regarding the regulation of AI. For example, in early 2018, New York City introduced its first algorithm accountability law, and similar bills were later introduced in Washington State and California. On the federal level, in 2019, new legislation was introduced in Congress that would require companies to audit their AI for potential biases and submit such assessments to the FTC.

As AI-related legislation becomes prevalent and regulators become more active in this field, companies should expect increased scrutiny with respect to how they are deploying their AI as well as the direct and indirect effects of their AI systems.

Moreover, due to the broad implications of AI-related biases, companies should expect class actions to be filed as these laws become widely available across states. Class action and civil rights lawyers may also attempt to use existing anti-discrimination laws to sue for AI biases. In fact, a class action against YouTube alleging that it uses AI, algorithms, and other tools to profile, target, and censor users based "wholly or in part" on race has already been filed.

Finally, in addition to the legal consequences, companies should be aware that AI bias incidents could have far-reaching PR implications.

Best Practices and Solutions

Companies deploying AI should take proactive measures and adopt tools to prevent and address potential biases. For larger companies, this usually includes adopting a written AI policy that assigns responsibility for ethical use of AI to specific stakeholders within the organization, establishing protocols to avoid use of biased AI, and to respond to any claims of bias.

Read more about bias in AI on Bloomberg Law