Harvard Business Review

AI And Machine Learning

# What Do We Do About the Biases in AI?

by James Manyika, Jake Silberg, and Brittany Presten

October 25, 2019

Wsestend61/Getty Images

**Summary.** Over the past few years, society has started to wrestle with just how much human biases can make their way into artificial intelligence systems—with harmful results. At a time when many companies are looking to deploy AI systems across their operations, being... **more**

Human biases are well-documented, from implicit association tests that demonstrate biases we may not even be aware of, to field experiments that demonstrate how much these biases can affect outcomes. Over the past few years, society has started to wrestle with just how much these human biases can make their way into artificial intelligence systems — with harmful results. At a time

when many companies are looking to deploy AI systems across their operations, being acutely aware of those risks and working to reduce them is an urgent priority.

The problem is not entirely new. Back in 1988, the UK Commission for Racial Equality found a British medical school guilty of discrimination. The computer program it was using to determine which applicants would be invited for interviews was determined to be biased against women and those with non-European names. However, the program had been developed to match human admissions decisions, doing so with 90 to 95 percent accuracy. What's more, the school had a higher proportion of non-European students admitted than most other London medical schools. Using an algorithm didn't cure biased human decision-making. But simply returning to human decision-makers would not solve the problem either.

Thirty years later, algorithms have grown considerably more complex, but we continue to face the same challenge. AI can help identify and reduce the impact of human biases, but it can also make the problem worse by baking in and deploying biases at scale in sensitive application areas. For example, as the investigative news site ProPublica has found, a criminal justice algorithm used in Broward Country, Florida, mislabeled African-American defendants as "high risk" at nearly twice the rate it mislabeled white defendants. Other research has found that training natural language processing models on news articles can lead them to exhibit gender stereotypes.

**INSIGHT CENTER**

**AI and Bias**

Building fair and equitable machine learning systems.

Bias can creep into algorithms in several ways. AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender, race, or sexual orientation are removed. Amazon stopped using a hiring algorithm after finding it favored applicants based on words like "executed" or "captured" that were more commonly found on men's resumes, for example. Another

source of bias is flawed data sampling, in which groups are over- or underrepresented in the training data. For example, Joy Buolamwini at MIT working with Timnit Gebru found that facial analysis technologies had higher error rates for minorities and particularly minority women, potentially due to unrepresentative training data.

Bias is all of our responsibility. It hurts those discriminated against, of course, and it also hurts everyone by reducing people's ability to participate in the economy and society. It reduces the potential of AI for business and society by encouraging mistrust and producing distorted results. Business and organizational leaders need to ensure that the AI systems they use improve on human decision-making, and they have a responsibility to encourage progress on research and standards that will reduce bias in AI.

From the growing academic research into AI bias, two imperatives for action emerge. First, we must responsibly take advantage of the several ways that AI can improve on traditional human decision-making. Machine learning systems disregard variables that do not accurately predict outcomes (in the data available to them). This is in contrast to humans, who may lie about or not even realize the factors that led them to, say, hire or disregard a particular job candidate. It can also be easier to probe algorithms for bias, potentially revealing human biases that had gone unnoticed or unproven (inscrutable though deep learning models may be, a human brain is the ultimate "black box"). Finally, using AI to improve decision-making may benefit traditionally disadvantaged groups, as researchers Jon Kleinberg, Sendhil Mullainathan, and others call the "disparate benefits from improved prediction."

The second imperative is to accelerate the progress we have seen in addressing bias in AI. Here, there are no quick fixes. In fact, one of the most complex steps is also the most obvious — understanding and measuring "fairness." Researchers have developed technical ways of defining fairness, such as requiring that models have equal predictive value across groups or requiring that models have equal false positive and false negative

rates across groups. However, this leads to a significant challenge — different fairness definitions usually cannot be satisfied at the same time.

Still, even as fairness definitions and metrics evolve, researchers have also made progress on a wide variety of techniques that ensure AI systems can meet them, by processing data beforehand, altering the system's decisions afterwards, or incorporating fairness definitions into the training process itself. One promising technique is "counterfactual fairness," which ensures that a model's decisions are the same in a counterfactual world where attributes deemed sensitive, such as race, gender, or sexual orientation, were changed. Silvia Chiappa of DeepMind has even developed a path-specific approach to counterfactual fairness that can handle complicated cases where some paths by which the sensitive traits affect outcomes is considered fair, while other influences are considered unfair. For example, the model could be used to help ensure that admission to a specific department at a university was unaffected by the applicant's sex while potentially still allowing the university's overall admission rate to vary by sex if, say, female students tended to apply to more competitive departments.

These improvements will help, but other challenges require more than technical solutions, including how to determine when a system is fair enough to be released, and in which situations fully automated decision making should be permissible at all. These questions require multi-disciplinary perspectives, including from ethicists, social scientists, and other humanities thinkers.

What can CEOs and their top management teams do to lead the way on bias and fairness? Among others, we see six essential steps:

First, business leaders will need to stay up to-date on this fast-moving field of research. Several organizations provide resources to learn more, such as the AI Now Institute's annual reports, the Partnership on AI, and the Alan Turing Institute's Fairness, Transparency, Privacy group.

Second, when your business or organization is deploying AI, establish responsible processes that can mitigate bias. Consider using a portfolio of technical tools, as well as operational practices such as internal "red teams," or third-party audits. Tech companies are providing some help here. Among others, Google AI has published recommended practices, while IBM's "Fairness 360" framework pulls together common technical tools.

Third, engage in fact-based conversations around potential human biases. We've long relied on proxies such as procedural checks when deciding if human decisions were fair. Now, with more advanced tools to probe for bias in machines, we can raise the standards to which we hold humans. This could take the form of running algorithms alongside human decision makers, comparing results, and using "explainability techniques" that help pinpoint what led the model to reach a decision in order to understand why there may be differences. Importantly, when we do find bias, it is not enough to change an algorithm—business leaders should also improve the human-driven processes underlying it.

Fourth, consider how humans and machines can work together to mitigate bias. Some "human-in-the-loop" systems make recommendations or provide options that humans double-check or can choose from. Transparency about these algorithms' confidence in its recommendation can help humans understand how much weight to give it.

Fifth, invest more, provide more data, and take a multi-disciplinary approach in bias research (while respecting privacy) to continue advancing this field. Important efforts to make designers' choices more transparent and embed ethics into computer science curricula, among others, point the way forward on collaboration. More will be needed.

Finally, invest more in diversifying the AI field itself. A more diverse AI community would be better equipped to anticipate, review, and spot bias and engage communities affected. This will require investments in education and opportunities — work like that of AI4ALL, a nonprofit focused on developing a diverse and

inclusive pipeline of AI talent in under-represented communities through education and retraining.

AI has many potential benefits for business, the economy, and for tackling society's most pressing social challenges, including the impact of human biases. But that will only be possible if people trust these systems to provide unbiased results. Ultimately, we can't hold humans with bias — but only if humans are working together to tackle bias in AI.

**James Manyika** is chairman and director of McKinsey Global Institute (MGI), the business and economics research arm of McKinsey & Company.

**Jake Silberg** is a consultant in McKinsey & Company's San Francisco office.

**Brittany Presten** is a consultant in McKinsey & Company's Bay Area office.