ute

kling bias in artificial intelligence (and umans)

Article

By Jake Silberg and James Manyika

AI has the potential to help humans make fairer decisions—but only if we carefully work toward fairness in AI systems as well.

DOWNLOADS

Article (PDF-120 KB)

The growing use of artificial intelligence in sensitive areas, including for hiring, criminal justice, and healthcare, has stirred a debate about bias and fairness. Yet human decision making in these and other domains can also be flawed, shaped by individual and societal biases that are often unconscious. Will AI's decisions be less biased than human ones? Or will AI make these problems worse?

In, *Notes from the AI frontier: Tackling bias in AI (and in humans)* (PDF–120KB), we provide an overview of where algorithms can help reduce disparities caused by human biases, and of where more human vigilance is needed to critically analyze the unfair biases that can become baked in and scaled by AI systems. This article, a shorter version of that piece, also highlights some of the research underway to address the challenges of bias in AI and suggests six pragmatic ways forward.

"Will AI's decisions be less biased than human ones? Or will AI

make these problems worse?"

Two opportunities present themselves in the debate. The first is the opportunity to use AI to identify and reduce the effect of human biases. The second is the opportunity to improve AI systems themselves, from how they leverage data to how they are developed, deployed, and used, to prevent them from perpetuating human and societal biases or creating bias and related challenges of their own. Realizing these opportunities will require collaboration across disciplines to further develop and implement technical improvements, operational practices, and ethical standards.

AI can help reduce bias, but it can also bake in and scale bias

Biases in how humans make decisions are well documented. Some researchers have highlighted how judges' decisions can be unconsciously influenced by their own personal characteristics, while employers have been shown to grant interviews at different rates to candidates with identical resumes but with names considered to reflect different racial groups. Humans are also prone to misapplying information. For example, employers may review prospective employees' credit histories in ways that can hurt minority groups, even though a definitive link between credit history and on-the-job behavior has not been established. Human decisions are also difficult to probe or review: people may lie about the factors they considered, or may not understand the factors that influenced their thinking, leaving room for unconscious bias.

"In many cases, AI can reduce humans' subjective interpretation of data, because machine learning algorithms learn to consider only the variables that improve their predictive accuracy, based on the training data used."

In many cases, AI can reduce humans' subjective interpretation of data, because machine learning algorithms learn to consider only the variables that improve their predictive accuracy, based on the training data used. In addition, some evidence shows that algorithms can improve decision making, causing it to become fairer in the process. For example, Jon Kleinberg and others have shown that algorithms could help reduce racial disparities in the criminal justice system. Another study found that automated financial underwriting systems particularly benefit historically underserved applicants. Unlike human decisions, decisions made by AI could in principle (and increasingly in practice) be opened up, examined, and

interrogated. To quote Andrew McAfee of MIT, "If you want the bias out, get the algorithms in."

At the same time, extensive evidence suggests that AI models can embed human and societal biases and deploy them at scale. Julia Angwin and others at ProPublica have shown how COMPAS, used to predict recidivism in Broward County, Florida, incorrectly labeled African-American defendants as "high-risk" at nearly twice the rate it mislabeled white defendants. Recently, a technology company discontinued development of a hiring algorithm based on analyzing previous decisions after discovering that the algorithm penalized applicants from women's colleges. Work by Joy Buolamwini and Timnit Gebru found error rates in facial analysis technologies differed by race and gender. In the "CEO image search," only 11 percent of the top image results for "CEO" showed women, whereas women were 27 percent of US CEOs at the time.

Underlying data are often the source of bias

Underlying data rather than the algorithm itself are most often the main source of the issue. Models may be trained on data containing human decisions or on data that reflect second-order effects of societal or historical inequities. For example, word embeddings (a set of natural language processing techniques) trained on news articles may exhibit the gender stereotypes found in society.

> **"Models may be trained on data containing human decisions or on data that reflect second-order effects of societal or historical inequities. "**

Bias can also be introduced into the data through how they are collected or selected for use. In criminal justice models, oversampling certain neighborhoods because they are overpoliced can result in recording more crime, which results in more policing.

Data generated by users can also create a feedback loop that leads to bias. In Latanya Sweeney's research on racial differences in online ad targeting, searches for African-American-identifying names tended to result in more ads featuring the word "arrest" than searches for white-identifying names. Sweeney hypothesized that even if different versions of the ad copy—versions with and without "arrest"—were initially displayed equally, users may have clicked on different versions more frequently for different searches, leading the algorithm to display them more often.

A machine learning algorithm may also pick up on statistical correlations that are societally

unacceptable or illegal. For example, if a mortgage lending model finds that older individuals have a higher likelihood of defaulting and reduces lending based on age, society and legal institutions may consider this to be illegal age discrimination.

In order to minimize bias, how do we define and measure fairness?

How should we codify definitions of fairness? Arvind Narayanan identified at least 21 different definitions of fairness and said that even that was "non-exhaustive." Kate Crawford, co-director of the AI Now Institute at New York University, used the CEO image search mentioned earlier to highlight the complexities involved: how would we determine the "fair" percentage of women the algorithm should show? Is it the percentage of women CEOs we have today? Or might the "fair" number be 50 percent, even if the real world is not there yet? Much of the conversation about definitions has focused on individual fairness, or treating similar individuals similarly, and on group fairness—making the model's predictions or outcomes equitable across groups, particularly for potentially vulnerable groups.

Work to define fairness has also revealed potential trade-offs between different definitions, or between fairness and other objectives. For example, Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, as well as Alexandra Chouldechova and others, have demonstrated that a model cannot conform to more than a few group fairness metrics at the same time, except under very specific conditions. This explains why the company that developed COMPAS scores claimed its system was unbiased because it satisfied "predictive parity," but ProPublica found that it was biased because it did not demonstrate "balance for the false positives."

Experts disagree on the best way to resolve these trade-offs. For example, some have suggested that setting different decision thresholds for different groups (such as the predicted score required to receive a loan) may achieve the best balance, particularly if we believe some of the underlying variables in the model may be biased. Others contend that maintaining a single threshold is fairer to all groups. As a result of these complexities, crafting a single, universal definition of fairness or a metric to measure it will probably never be possible. Instead, different metrics and standards will likely be required, depending on the use case and circumstances.

Early technical progress is underway, but much more is needed

Several approaches to enforcing fairness constraints on AI models have emerged. The first consists of pre-processing the data to maintain as much accuracy as possible while reducing any relationship between outcomes and protected characteristics, or to produce representations of the data that do not contain information about sensitive attributes. This latter group includes "counterfactual fairness" approaches, which are based on the idea that a decision should remain the same in a counterfactual world in which a sensitive attribute is changed. Silvia Chiappa's path-specific counterfactual method can even consider different ways that sensitive attributes may affect outcomes—some influence might be considered fair and could be retained, while other influence might be considered unfair, and therefore should be discarded.

The second approach consists of post-processing techniques. These transform some of the model's predictions after they are made in order to satisfy a fairness constraint. The third approach either imposes fairness constraints on the optimization process itself or uses an adversary to minimize the system's ability to predict the sensitive attribute.

Researchers are also developing and testing other improvements. On the data side, researchers have made progress on text classification tasks by adding more data points to improve performance for protected groups. Innovative training techniques such as using transfer learning or decoupled classifiers for different groups have proven useful for reducing discrepancies in facial analysis technologies.

"Innovative training techniques such as using transfer learning or decoupled classifiers for different groups have proven useful for reducing discrepancies in facial analysis technologies."

Finally, techniques developed to address the adjacent issue of explainability in AI systems—the difficulty when using neural networks of explaining how a particular prediction or decision was reached and which features in the data or elsewhere led to the result—can also play a role in identifying and mitigating bias. Explainability techniques could help identify whether the factors considered in a decision reflect bias and could enable more accountability than in human decision making, which typically cannot be subjected to such rigorous probing.

Human judgment is still needed to ensure AI supported decision making is

fair

While definitions and statistical measures of fairness are certainly helpful, they cannot consider the nuances of the social contexts into which an AI system is deployed, nor the potential issues surrounding how the data were collected. Thus it is important to consider where human judgment is needed and in what form. Who decides when an AI system has sufficiently minimized bias so that it can be safely released for use? Furthermore, in which situations should fully automated decision making be permissible at all? No optimization algorithm can resolve such questions, and no machine can be left to determine the right answers; it requires human judgment and processes, drawing on disciplines including social sciences, law, and ethics, to develop standards so that humans can deploy AI with bias and fairness in mind. This work is just beginning.

Some of the emerging work has focused on processes and methods, such as "data sheets for data sets" and "model cards for model reporting" which create more transparency about the construction, testing, and intended uses of data sets and AI models. Other efforts have focused on encouraging impact assessments and audits to check for fairness before systems are deployed and to review them on an ongoing basis, as well as on fostering a better understanding of legal frameworks and tools that may improve fairness. Efforts such as the annual reports from the AI Now Institute, which cover many critical questions about AI, and Embedded EthiCS, which integrates ethics modules into standard computer science curricula, demonstrate how experts from across disciplines can collaborate.

One method for ensuring fairness focuses on encouraging impact assessments and audits to check for fairness before systems are deployed and to review them on an ongoing basis.

As we raise the bar for automated decision making, can we also hold human decision making to a higher standard?

Progress in identifying bias points to another opportunity: rethinking the standards we use to determine when human decisions are fair and when they reflect problematic bias. Reviewing the actual factors humans used (not what they say they used) when making a decision is much more difficult than evaluating algorithms. More often than not we rely on fairness proxies. For example, we often accept outcomes that derive from a process that is considered "fair." But is

procedural fairness the same as outcome fairness? Another proxy often used is compositional fairness, meaning that if the group making a decision contains a diversity of viewpoints, then what it decides is deemed fair. Perhaps these have traditionally been the best tools we had, but as we begin to apply tests of fairness to AI systems, can we start to hold humans more accountable as well?

"Much of the conversation about definitions has focused on individual fairness, or treating similar individuals similarly, and on group fairness—making the model's predictions or outcomes equitable across groups, particularly for potentially vulnerable groups."

Better data, analytics, and AI could become a powerful new tool for examining human biases. This could take the form of running algorithms alongside human decision makers, comparing results, and examining possible explanations for differences. Examples of this approach are starting to emerge in several organizations. Similarly, if an organization realizes an algorithm trained on its human decisions (or data based on prior human decisions) shows bias, it should not simply cease using the algorithm but should consider how the underlying human behaviors need to change. Perhaps organizations can benefit from the recent progress made on measuring fairness by applying the most relevant tests for bias to human decisions, too.

Six potential ways forward for AI practitioners and business and policy leaders to consider

Exhibit

Minimizing bias in AI is an important prerequisite for enabling people to trust these systems. This will be critical if AI is to reach its potential, shown by the research of MGI and others, to

drive benefits for businesses, for the economy through productivity growth, and for society through contributions to tackling pressing societal issues. Those striving to maximize fairness and minimize bias from AI could consider several paths forward:

1. Be aware of the contexts in which AI can help correct for bias as well as where there is a high risk that AI could exacerbate bias.

When deploying AI, it is important to anticipate domains potentially prone to unfair bias, such as those with previous examples of biased systems or with skewed data. Organizations will need to stay up to date to see how and where AI can improve fairness—and where AI systems have struggled.

2. Establish processes and practices to test for and mitigate bias in AI systems.

Tackling unfair bias will require drawing on a portfolio of tools and procedures. The technical tools described above can highlight potential sources of bias and reveal the traits in the data that most heavily influence the outputs. Operational strategies can include improving data collection through more cognizant sampling and using internal “red teams” or third parties to audit data and models. Finally, transparency about processes and metrics can help observers understand the steps taken to promote fairness and any associated trade-offs.

3. Engage in fact-based conversations about potential biases in human decisions.

As AI reveals more about human decision making, leaders can consider whether the proxies used in the past are adequate and how AI can help by surfacing long-standing biases that may have gone unnoticed. When models trained on recent human decisions or behavior show bias, organizations should consider how human-driven processes might be improved in the future.

4. Fully explore how humans and machines can work best together.

This includes considering situations and use-cases when automated decision making is acceptable (and indeed ready for the real world) vs. when humans should always be involved. Some promising systems use a combination of machines and humans to reduce bias. Techniques in this vein include "human-in-the-loop" decision making, where algorithms provide recommendations or options, which humans double-check or choose from. In such systems, transparency about the algorithm's confidence in its recommendation can help humans understand how much weight to give it.

5. Invest more in bias research, make more data available for research (while respecting privacy), and adopt a multidisciplinary approach.

While significant progress has been made in recent years in technical and multidisciplinary research, more investment in these efforts will be needed. Business leaders can also help support progress by making more data available to researchers and practitioners across organizations working on these issues, while being sensitive to privacy concerns and potential risks. More progress will require interdisciplinary engagement, including ethicists, social scientists, and experts who best understand the nuances of each application area in the process. A key part of the multidisciplinary approach will be to continually consider and evaluate the role of AI decision making, as the field progresses and practical experience in real applications grows.

6. Invest more in diversifying the AI field itself.

Many have pointed to the fact that the AI field itself does not encompass society's diversity, including on gender, race, geography, class, and physical disabilities. A more diverse AI community will be better equipped to anticipate, spot, and review issues of unfair bias and better able to engage communities likely affected by bias. This will require investments on multiple fronts, but especially in AI education and access to tools and opportunities.

ABOUT THE AUTHOR(S)

Jake Silberg is a fellow at the McKinsey Global Institute (MGI). **James Manyika** is the chairman of MGI and a senior partner at McKinsey & Company in the San Francisco office.

This article draws from remarks the authors prepared for a recent multidisciplinary symposium on ethics in AI hosted by DeepMind Ethics and Society. The authors wish to thank Dr. Silvia Chiappa, a research scientist at DeepMind, for her insights as well as for co-chairing the fairness and bias session at the symposium with James.

In addition, the authors would like to thank the following people for their input on the ideas in this article: Mustafa Suleyman and Haibo E at DeepMind; Margaret Mitchell at Google AI and Charina Chou at Google; Professor Barbara Grosz and Lily Hu at Harvard University; Mary L. Gray and Eric Horvitz at Microsoft Research; Professor Kate Crawford at New York University and Microsoft Research; and Professor Sendhil Mullainathan at the University of Chicago. They also wish to thank their McKinsey colleagues Tara Balakrishnan, Jacques Bughin, Michael Chui, Rita Chung, Daniel First, Peter Gumbel, Mehdi Miremadi, Brittany Presten, Vasiliki Stergiou, and Chris Wigley for their contributions.
