McKinsey Global Institute

Notes from the AI frontier: Tackling bias in AI (and in humans)

Article

By Jake Silberg and James Manyika

June 2019

The growing use of artificial intelligence in sensitive areas, including 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?

These complex challenges relating to computer algorithms are not new. In 1988, the UK Commission for Racial Equality found a British medical school guilty of discrimination because the computer program it was using to determine which applicants would be invited for interviews was determined to be biased against women and applicants with non-European names.[1] 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—meaning the "biased" algorithm still produced less biased results than humans at other schools. Finally, the human admissions officers' biases would probably never have been demonstrated but for the use of the program.

Today, the use of AI is projected to grow across a wide range of use cases and application arenas as a result of progress in machine learning, particularly deep learning. In coming decades, AI applications are likely to affect individuals, businesses and other organizations across sectors, and economies everywhere, touching billions of lives. As a result, understanding the complex intersections between AI and bias will only become more important. Defining what counts as bias is itself still a challenge (see Box 1, "Defining bias and fairness"). But definitional challenges notwithstanding, many experts tend to welcome algorithms as a refreshing antidote to human biases that have always existed. At the same time, many worry that algorithms may bake in and scale human and societal biases.

The evidence suggests that both claims are valid and must be taken seriously. However, the claims also suggest two opportunities. 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. To realize these opportunities will require collaboration across disciplines to further develop and implement technical improvements, operational practices, and ethical standards. Lessons learned from these experiences will also need to be continuously integrated to

Box 1

Defining bias and fairness

Bias and fairness are complex human notions. While "bias" can refer to any form of preference, fair or unfair, this article uses the term to mean "unfair," "unwanted," or "undesirable" bias—that is, systematic discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics.[2] Perhaps the most discussed forms of "unfair bias" in the literature relate to particular attributes or groups such as disabilities, race, gender, and sexual orientation. This is certainly not a complete list, which is why this article will sometimes broadly refer to "sensitive attributes" or "protected characteristics" (clearly somewhat legalistic terms) that could be the basis of unfair discrimination.

The absence of unwanted bias is not sufficient to conclude that a system is "fair." Literature about this topic is extensive, particularly concerning the ethical need to understand the historical and social contexts into which these systems are being deployed.[3] This paper discusses some methods that seek to measure bias and fairness with respect to AI systems. More comprehensive views on the principles and values for the fair and ethical use of AI can be found in the AI Now Institute's annual reports and the Asilomar AI principles, among many other important discussions.

Another important strand of discussion has to do with the issue of how AI is used. One argument suggests that to the extent AI is used for decision making, prediction, and allocative efficiency, it will always be subject to challenges of bias and fairness. Such arguments suggest that thinking of AI's use more from the point of view of a tool to enable equity and inclusion is a more fruitful approach for wider social benefit. This is clearly an important avenue for further research.

minimize bias in both AI systems and human decision making.

Given the wide range of potential applications in the commercial and public sectors, it is important that leaders in these arenas understand the complications and challenges related to AI and bias, as well as the work under way to tackle these issues and the limitations of that effort. Drawing on the work of many in the field, the goal of this article is to:

— 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 biases that can become baked in and scaled in AI systems.

— Highlight some of the research under way across disciplines to address the challenges of bias in AI and point to some emerging resources and practical approaches for users and others wishing to delve deeper.

— Suggest some pragmatic ways forward and some of the work needed.

These are still early days when it comes to AI and bias and being wary of promises of quick fixes or silver-bullet solutions is appropriate. The work highlighted here is by no means comprehensive nor is it the end of the story with respect to AI and bias. Rather, it is the start of a journey to ensure that AI lives up to its potential.

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.[4] 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.[5] Finally, human decisions are difficult to probe or review. Humans may lie about the factors they considered, but they also may not understand the factors that influenced their thinking, leaving room for unconscious bias.[6]

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.[7] In addition, some evidence shows that algorithms can improve decision making, causing it to become fairer in the process.[8] For example, Jon Kleinberg and others have shown that algorithms could help reduce racial disparities in the criminal justice system.[9] Similarly, a study

found that automated financial underwriting systems particularly benefit historically underserved applicants.[10] Finally, unlike human decisions, decisions made by AI could in principle (and increasingly in practice) be opened up, examined, and interrogated. These advantages over humans involve caveats, but they point to exciting possibilities. 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.[11] 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. Research has also highlighted "harms of representation," meaning discrepancies in how different groups experience technology.[12] Work by Joy Buolamwini and Timnit Gebru found error rates in facial analysis technologies differed by race and gender.[13] 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.[14] These examples highlight how some algorithms, adopted and implemented for their efficiency and efficacy benefits, can also deploy unnoticed or unchecked biases at scale in the process.

Underlying data are often the source of bias

While this form of bias has often been called "algorithmic bias," the underlying data rather than the algorithm itself are most often the main source of the issue. Here some researchers make a useful distinction and separate the model into two different algorithms—the trainer, which can be biased by the underlying data and training process, and the screener, which simply makes predictions based on the trainer.[15] 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.[16] Such biases are often encoded by other variables even when algorithms are prevented from considering protected characteristics directly. For example, in the hiring algorithm discussed above, the system learned to favor words that were more commonly found on men's applications, such as "executed" or "captured."[17]

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 more recorded crime, which results in more policing.[18] In financial decision making, undersampling certain groups could lead to models that approve groups of applicants at lower rates. The choice of variables can also introduce bias. For example, Ziad Obermeyer and Sendhil Mullainathan analyzed a prominent healthcare algorithm that quantifies how sick patients are by measuring their cost of care. Although the variables are highly correlated, because African-American patients in the data set tended to have lower treatment costs for the same level of sickness, the choice of variables led the algorithm to enroll African-American patients in supplemental programs at a much lower rate than white patients with the same level of sickness.[19]

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.[20] 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. Given the number of algorithms reacting to billions of user actions every day, this is an increasingly important potential source of bias.

Finally, a machine learning algorithm may 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.[21]

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

How should we codify definitions of fairness? 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?[22] 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.[23] However, deciding on the best metric or combination of metrics to determine if a system demonstrates group fairness, individual fairness, or other notions of fairness is complex. Arvind Narayanan identified at least 21 different definitions of fairness and said even that was "non-exhaustive."[24]

Work to define fairness has also revealed potential trade-offs between different definitions, or between fairness and other objectives. 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.[25] 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." Other research has shown that ensuring an AI system satisfies measures of group fairness could create trade-offs with measures of individual fairness or could reduce the utility of the model.[26] Finally, the use of fairness constraints can also create longer-term impacts, potentially both positive and negative. For example, Lily Hu and Yiling Chen modelled a temporary labor market with a fairness constraint imposed on hiring, finding it could improve workers' incentives to invest in education and enable a more equitable permanent labor market where the constraint was not needed.[27] On the other hand, researchers have suggested that imposing a fairness constraint in some use cases could, in the long term, hurt the groups we are seeking to protect by, for example, approving loans that people cannot repay, hurting their credit scores.[28]

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.[29] Others contend that maintaining a single threshold is fairer to all groups.[30] 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. In particular, choosing appropriate fairness metrics likely depends on the specific patterns of potential unfairness in the training data sets being used.[31]

Early technical progress has been demonstrated, but much more is needed

Even as researchers continue to refine definitions of fairness, many are also developing techniques to ensure that models can adhere to them—often called "enforcing fairness constraints."

Several approaches 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.[32] 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.[33] 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.[34] These approaches require forming hypotheses about the causal mechanisms underlying the data. 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.[35] 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.[36]

Beyond approaches based on fairness constraints, researchers are developing and testing other improvements. On the data side, researchers have made progress on text classification tasks by consciously adding more data points to improve performance for protected groups.[37] Meanwhile, innovative training techniques such as using transfer learning or decoupled classifiers for different groups have proven useful for reducing discrepancies in facial analysis technologies.[38]

Further, 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. These techniques include local interpretable model-agnostic explanations (LIME), integrated gradients, and testing with concept activation vectors.[39] LIME, for example, probes certain segments of data at a time (such as a nose, then ears, in image recognition) and observes the resulting changes in predictions to fine-tune a proxy model, thus identifying the most influential factors for a decision. 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.[40]

Finally, other researchers are considering how humans and machines can better work together to mitigate bias. For example, models might defer to an alternative decision-making process in some cases, rather than always providing an answer, and could even learn how often to defer based on how fair the alternative process seems to be.[41] Some research has also shown that using an algorithm to assign cases among several human decision makers can be more equitable than random assignment. [42]

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.[43] Thus it is important to consider where human judgment is needed and in what form, besides providing definitions and applying statistical techniques. And who decides when an AI system has sufficiently minimized unfair bias so that it can be safely released for use? Furthermore, in which situations should fully automated decision making be permissible at all? These are decisions and questions that no optimization algorithm can resolve on its own, and that no machine can be left to determine. They require human judgment and processes, drawing on many disciplines including the humanities, especially the 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.[44] 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.[45] Efforts such as annual reports from the AI Now Institute, which cover bias and fairness as well as many other critical societal questions about AI, and Embedded EthiCS, which enables integrating ethics modules into standard computer science curricula, demonstrate how experts from across disciplines can collaborate.[46]

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. Compared to evaluating algorithms, reviewing the actual factors humans used (not what they say they used) when making a decision is much more difficult.[47] Instead, to evaluate and become comfortable with human decision making, more often than not we rely on fairness proxies. For example, we often accept outcomes that derive from a process that is considered "fair" (such as an evaluation rubric), 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?

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. These possibilities are just the beginning.

Practitioners and business and policy leaders could consider several potential ways forward

Biased decision making, whether by humans or machines, not only has devastating effects for the people discriminated against but also hurts everyone by unduly restricting individuals' ability to participate in and contribute to the economy and society. Further, 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.[48]

What follows are suggestions for those striving to maximize fairness and minimize bias from AI:

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 bias, such as those with previous examples of biased systems or with skewed data. This is especially important in applications that are likely to be rapidly adopted because of their commercial benefits, ease of use, and efficiency. Given the growing number of use cases where AI can reduce disparities caused by human bias, 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 procedures can include improving data collection through more cognizant sampling and using internal "red teams" or third parties to audit data and models, as well as proactively engaging with communities potentially affected. Finally, transparency about processes and metrics can help observers understand the steps taken to promote fairness and any associated trade-offs. Box 2 lists some organizations working on additional practical resources related to bias and fairness.

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 of the promising systems highlighted above use a combination of machines and humans to reduce bias. Other 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, physical disabilities, and more characteristics. 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.

Box 2

Resources for tackling bias and fairness

Many leaders in developing and studying AI technologies are working to provide resources for organizations seeking to deploy AI fairly. They include the following efforts:

- The AI Now Institute at New York University publishes annual reports, now in their third year, providing one of the longest-running series of research reports about bias in AI.
- Additional academic efforts include the Alan Turing Institute's Fairness, Transparency, Privacy group, the Ethics and Governance of Artificial Intelligence Initiative affiliated with the Berkman Klein Center at Harvard and the MIT Media Lab, and the Stanford Institute for Human-Centered AI.
- The European Commission's report *Ethics guidelines for trustworthy AI* includes a checklist of questions about bias and fairness.
- Google AI, in conjunction with its AI principles, has published a set of recommended practices for fairness as well as for other important AI topics.
- Microsoft has released guidelines for conversational AI bots to treat people fairly. The guidelines are part of Microsoft's AI principles, including fairness.
- IBM has released AI Fairness 360, an open-source tool kit to test for and reduce bias, and Microsoft has made its fairness framework available on GitHub.
- FAT/ML (Fairness, Accountability, and Transparency in Machine Learning), which has grown into the FAT* conference, has published guiding principles and questions.
- The Partnership on AI, with the participation of leading technology and civil society groups, has a Fair, Transparent, and Accountable AI working group.
- The Algorithmic Justice League, founded by Joy Buolamwini, aims to catalogue biases and offers auditing of algorithms.
- AI4ALL, a nonprofit organization, focuses on developing a diverse and inclusive pipeline of AI talent in underrepresented communities through education and mentorship of high school students, in collaboration with leading AI research universities.
- QuantumBlack, a McKinsey company specializing in analytics and AI, has published a paper on operationalizing risk management in machine learning, including explainability and bias, which will be presented at ICML's AI for social good workshop.

No silver bullet or quick fix can solve the challenge of bias. While researchers and practitioners across disciplines are making progress in identifying how AI can reduce some of the disparities caused by human biases and deploy AI more fairly, more progress is needed. As this work advances, it will be up to organizational leaders to apply that knowledge by working to identify use cases where AI can help reduce bias, proactively implementing appropriate strategies to ensure that AI is used responsibly and enabling the recent pace of progress on important research to continue or, better yet, to accelerate.

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.

McKinsey Global Institute
June 2019

Designed by the McKinsey Global institute

www.mckinsey.com

@McKinsey_MGI
McKinseyGlobalInstitute

Endnotes

1 Stella Lowry and Gordon Macpherson, "A blot on the profession," *British Medical Journal*, March 1988, Volume 296, Number 623, pp. 657–658.
2 Batya Friedman and Helen Nissembaum, "Bias in computer systems," *ACM Transactions on Information Systems*, July 1996, Volume 14, Number 3, pp. 330–347.
3 Ben Green and Lily Hu, *The myth in the methodology: Towards a recontextualization of fairness in machine learning*, 35th International Conference on Machine Learning, Stockholm, Stockholm, Sweden, July 10–15, 2018; Meredith Whittaker et al., *AI Now Report 2018*, AI Now Institute, New York University, December 2018.
4 Jeffrey J. Rachlinksi et al., "Does unconscious racial bias affect trail judges?", Notre Dame Law Review, March 2009, Volume 84, Number 3; Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, September 2004, Volume 94, Number 4, pp. 991–1013.
5 Barbara Kiviat, "The art of deciding with data: Evidence from how employers translate credit reports into hiring decisions," *Socio-Economic Review*, August 30, 2017; Rourke L. O'Brien and Barbaria Kiviat, "Disparate impact? Race, sex, and credit reports in hiring," *Socius*, 2018, Volume 4, pp. 1–20.
6 Jon Kleinberg et al., *Discrimination in the age of algorithms*, SSRN, February 2019.
7 Ibid
8 Alex P. Miller, "Want less-biased decisions? Use algorithms," *Harvard Business Review*, July 26, 2018.
9 Jon Kleinberg et al., "Human decisions and machine predictions," *The Quarterly Journal of Economics*, February 2018, Volume 133, Issue 1, pp. 237–293.
10 Susan Wharton Gates, Vanessa Gail Perry, and Peter M. Zorn, "Automated underwriting in mortgage lending: Good news for the underserved?" *Housing Policy Debate*, 2002, Volume 13, Issue 2, pp. 369–391.
11 Julia Angwin et al., "Machine Bias," ProPublica, May 2016.
12 Kate Crawford, "You and AI—the politics of AI," *The Royal Society*, July 2018.
13 Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," *Proceedings of Machine Learning Research*, 2018, Volume 1, pp. 1–15.
14 Jennifer Langston, "Who's a CEO? Google image results can shift gender biases," *UW News*, April 2015.
15 Jon Kleinberg et al., *Discrimination in the age of algorithms, SSRN*, February 2019.
16 Tolga Bolukbasi et al., "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," *Proceedings of the 30th International Neural Information Processing Systems*, pp. 4356–4364; *Google Developers Blog*, "Text embedding models contain bias. Here's why that matters," blog entry by Ben Packer, Yoni Halpern, Mario Guajardo-Céspedes & Margaret Mitchell, April 13, 2018.
17 Karen Hao, "This is how AI bias really happens—and why it's so hard to fix," *MIT Technology Review*, February 4, 2019.
18 Kristian Lum and William Isaac, "To predict and serve?" *Significance*, October 2016, Volume 13, Issue 5.
19 Ziad Obermeyer and Sendhil Mullainathan, "Dissecting racial bias in an algorithm that guides health decisions for 70 million people," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, January 2019, pp. 89–89.
20 Latanya Sweeney, "Discrimination in Online Ad Delivery," *Queue*, March 2013, Volume 11, Issue 3.
21 *Machine learning: The power and promise of computers that learn by example*, Royal Society, April 2017.
22 Kate Crawford, "You and AI—the politics of AI," *The Royal Society*, July 2018.
23 Richard Zemel et al., "Learning Fair Representations," *Proceedings of the 30th International Conference on Machine Learning*, 2013.
24 Arvind Narayanan, "Tutorial: 21 fairness definitions and their politics," *FAT**, March 2018.
25 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the Fair Determination of Risk Scores" *Proceedings of Innovations in Theoretical Computer Science*, 2017; Alexandra Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, June 2017, Volume 5, Number 2.
26 Arvind Narayanan, "Tutorial: 21 fairness definitions and their politics," *FAT**, March 2018; Sam Corbett-Davies et al., *Algorithmic decision-making and the cost of fairness*, June 10, 2017.
27 Lily Hu and Yiling Chen, "A short-term intervention for long-term fairness in the labor market," *Proceedings of the 2018 World Wide Web Conference*, April 2018.
28 Lydia T. Liu et al., "Delayed impact of fair machine learning," *Proceedings of the 35th International Conference on Machine Learning*, 2018.
29 Jon Kleinberg et al., "Algorithmic fairness," *AEA Papers and Proceedings*, 2018, Volume 108, pp. 22–27.
30 Sam Corbett-Davies and Sharad Goel, "The measure and mis-measure of fairness: A critical review of fair machine learning," *ArXiv*, September 2018.
31 Silvia Chiappa and William S. Isaac, "A causal Bayesian networks viewpoint on fairness," *Privacy and Identity 2018, April 2019*, IFIP AICT 547, pp. 3–20.
32 Flavio Calmon et al., "Optimized data pre-processing for discrimination prevention," 31st Conference on Neural Information Processing Systems, 2017.
33 Matt J. Kusner et al., "Counterfactual fairness," *31st Conference on Neural Information Processing Systems*, 2017.
34 Silvia Chiappa and Thomas P. S. Gillam, "Path-specific counterfactual fairness," *ArXiv*, February 2018.
35 Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of opportunity in supervised learning," *30th Conference on Neural Information Processing Systems*, 2016.
36 Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, "Mitigating unwanted biases with adversarial learning," *AIES* '18, February 2018. Lucas Dixon et al., "Measuring and mitigating unintended bias in text classification," *AAAI*, 2017.
37 Hee Jung Ryu et al. "InclusiveFaceNet: Improving face attribute detection with race and gender diversity," *Proceedings of FATL/ML 2018*, 2018; Cynthia Dwork et al., "Decoupled classifiers for group-fair and efficient machine learning," *Proceedings of Machine Learning Research*, 2018, Volume 81, pp. 1–15.
38 Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Introduction to local interpretable model-agnostic (LIME) explanations: An introduction," *O'Reilly*, August 12, 2016; Mukund Sundararajan, Ankur Taly, and Qiqi Yan, "Axiomatic attribution for deep networks," *Proceedings of the 34th International Conference on Machine Learning*, Volume 70, pp. 3319–3328; Been Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," *ArXiv*, June 2018.
40 Damon Civin, "Explainable AI could reduce the impact of biased algorithms," VentureBeat, May 2018.
41 David Madras, Toniann Pitassi, and Richard Zemel, "Predict responsibly: Improving fairness and accuracy by learning to defer," *32nd Conference on Neural Information Processing Systems*, 2018; Ran Canetti et al., "From soft classifiers to hard decisions: How fair can we be?" *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019.
42 Isabel Valera, Adish Singla, and Manuel Gomez-Rodriguez, "Enhancing the accuracy and fairness of human decision making," *32nd Conference on Neural Information Processing Systems*, 2018.
43 Ben Green and Lily Hu, *The myth in the methodology: Towards a recontextualization of fairness in machine learning*, 35th International Conference on Machine Learning, Stockholm, Sweden, 2018; Rashida Richardson, Jason Schultz, and Kate Crawford, "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice," *New York University Law Review Online*, March 2019.
44 Timnit Gebru, "Datasheets for datasets," *ArXiv*, March 2018; Margaret Mitchell et al., "Model cards for model reporting," *ArXiv*, October 2018.
45 Solon Barocas and Andrew D. Selbst, "Big data's disparate impact," *California Law Review*, 2016, Volume 104, pp. 671–732.
46 Meredith Whittaker et al., *AI Now Report 2018*, AI Now Institute, New York University, December 2018; Barbara Grosz et al., "Embedded EthiCS: Integrating ethics broadly across computer science education," *Communications of the ACM*, forthcoming.
47 Jon Kleinberg et al., *Discrimination in the age of algorithms*, SSRN, February 2019.
48 *The promise and challenge of the age of artificial intelligence*, McKinsey Global Institute, October 2018.