ITIF | INFORMATION TECHNOLOGY & INNOVATION FOUNDATION

Defending Digital Series

AI Bias Is Correctable. Human Bias? Not So Much

DAVID MOSCHELLA | APRIL 2022

Defending Digital Series, No. 5: Individual decisions are shaped by our values, beliefs, experiences, inclinations, prejudices, and blind spots. These "biases" can easily leak into information system design. But overall and over time, modern technology will prove a force for more fair and objective societal actions.

You've probably heard the joke. Two campers see a bear approaching. One starts putting on her sneakers. The other asks: "What are you doing? You can't outrun a bear." The response is obvious: "I don't have to outrun the bear; I just have to outrun you."

So it is with the "objectivity" of information systems. The question is not whether digital services are—or ever can be—completely unbiased. It's whether they can outrun the available alternatives. Much more often than not, they can and will.

PERVASIVE HUMAN BIASES

The word *bias* is now mostly used as a pejorative, but biases should be seen along a spectrum of human attitudes spanning our values, beliefs, allegiances, opinions, preferences, interests, stereotypes, prejudices, and misunderstandings. This broader definition shows how deeply biases are built into the human psyche, and why they can never be fully eliminated. Some biases are seen as "good" and to be built upon; others are "bad" and in need of remedies. Sometimes there is a clear consensus as to what is good or bad, but often this a matter of debate.

Additionally, it's all too easy to believe that bad biases are the result of a lack of understanding, and if only people were more informed, society would be much more objective and fair. While lack of knowledge is certainly a major source of bias, professional expertise doesn't fare much better. Whether we are looking at judges, lawyers, professors, scientists, doctors, engineers, architects, writers, journalists, politicians, investors, economists, managers, coaches, consultants, or computer programmers, sharp differences and entrenched opinions are the norm. Deep experience and expertise do not necessarily lead to objective consensus. As behavioral scientists have long noted, subject matter experts tend to:

1. Rely too much on societal and professional stereotypes
2. Overvalue their personal experiences, especially recent ones
3. Overvalue their personal *gut feel*
4. Prefer anecdotes that confirm their existing views

5. Have limited knowledge of statistics and probability
6. Resist admitting mistakes
7. Struggle to keep up with the skills and literature in their fields
8. Burn out and/or make mistakes in demanding work environments
9. Avoid criticizing, evaluating, or disciplining their peers
10. Become less open-minded over time

For decades, we have seen the unfortunate results of these traits in criminal sentencing, student grading, medical diagnoses and treatments, hiring and salary negotiations, financial services, editorial coverage, athletic evaluations, political processes, and many other areas. Fortunately, data analytics, expert systems, algorithmic operations, and—most recently—machine learning can establish a software and data foundation upon which better and more consistent decision-making can be built. It's hard to think of any other approach that can potentially address these long-standing societal challenges so directly.

AI BIASES IN PERSPECTIVE

While AI offers many important benefits, the potential risks get most of the media attention.[1] Will AI usher in a surveillance society? Will it eliminate jobs and diminish human worth? Will it increase the divide between the haves and have nots? Will any one nation gain a decisive edge? Will humans still be in control? Compared to these existential questions, the challenge of machine learning bias is both more manageable and a lesser danger. It is less dangerous because AI can mitigate human shortcomings, and it is more manageable because AI bias is correctable and businesses and governments have strong incentives to do so.

That this bias exists is undeniable. A machine-learning system is only as objective as its underlying dataset. If this data reflects historical biases, then so will the system. If certain populations are not sufficiently represented statistically, then the system will be weak in those areas. These situations are analogous to the way surveys work. If a survey sample is not representative of the real world, or if the wording of a survey's questions is slanted in some way, these deficiencies will affect the survey's results. Similarly, if a facial recognition system doesn't have enough faces from people of color, it will be less effective in identifying those faces, just as some Asian systems perform worse on non-Asian faces.

The good news is that just as survey professionals strive to design random samples and ask neutral questions, so can AI developers take steps to assure the quality of their underlying data. Systems can, for example, be tested for bias by isolating criteria such as race, gender, location and other factors. Facial recognition weaknesses can and have been corrected through better data. These and similar techniques will only improve over time, as business practices mature and as the volume of relevant data steadily increases. While eliminating all bias is impossible (and often not desirable), building systems that outrun the average human is a very achievable task.

RELATED CRITIQUES

Unfortunately, building systems that demonstrably outperform or complement human decision-making is not enough to satisfy many of today's AI critics, as machine-learning systems are now

disparaged on numerous related fronts. Most of these critiques are not about bias per se. Rather, they reflect an underlying antitechnology, anticorporate bias. But if citizens are to accept the widespread use of AI, then the machine-learning community must do a better job of answering the five questions below:

1. **What sort of AI systems do we want?** Different societies will need to find their own balance between, for example, the security and convenience benefits of an effective facial recognition system versus the potential risks of oppressive visual surveillance.
2. **How well do AI systems work in the real world?** As with any innovation, governments and businesses need to evaluate when a system is ready for widespread use. It's a familiar cost/benefit/risk calculation. But some systems have been branded as biased when the real problem is that they just aren't good enough yet.
3. **Should AI decisions be explainable?** While the logic behind data analytics and rule-based algorithms is often transparent, the inner workings of most machine-learning systems are not, and won't be for the foreseeable future. Important AI applications, such as medical image diagnoses, are essentially *black boxes* that need to be evaluated on the quality of their output, not their lack of *explainability*.[2]
4. **What recourse should citizens have?** Given the lack of explainability, citizen recourse may well be required in sensitive areas such as loan approvals or school acceptances. Sometimes, business and government AI operators will have to be able to do more than just say, "that's what the system says," and at least be able to match the recourse available in the pre-AI era.
5. **Are the results of these systems socially acceptable?** An AI application might be based on a representative dataset and provide accurate output in the areas it was designed for. But if that output has an adverse effect upon a particular race, class, gender, or other group, the system might still be criticized as biased. We often see this with auto insurance, where premiums are based on a variety of risk factors that sometimes result in different rates for different demographic groups.

Fears of AI bias also reflect the familiar hi-tech double-standard. A single potent example of an AI failure—such as the racist rantings of Microsoft's chatbot, Tay, or a deadly accident involving a self-driving Tesla—will get vastly more media coverage than everyday car crashes or human misbehavior. Likewise, organizations such as the Algorithmic Justice League will continue to effectively publicize AI's risks, failures, and injustices, sometimes in ways that are misleading. It's important that the AI community be able to respond to concerns such as those in AJL's movie, *Coded Bias*.[3]

CONSCIOUS VS. UNCONSCIOUS BIAS

Thus far, the bias we have discussed has been mostly unconscious in nature. Organizations almost never set out to build a system that discriminates by age, gender, race, disability, or any other factor, and most AI developers want to use the best data available. However, as these developers are usually technologists at heart, it's fair to say that they sometimes fail to prioritize this issue—although this is now changing. Similarly, many human biases are also unconscious in nature. We may think that we are being impartial and fair, but our minds are full of stereotypes, preconceptions, self-interests, confirmation biases, and other discriminatory forces.

Conscious bias is another thing altogether. Conservatives might have a conscious bias towards systems that treat everyone the same, while liberals are more likely to factor in historical inequities. Should college admissions be based mostly on empirical test scores, or should the backgrounds of applicants be given much more weight? Should criminal sentencing be uniform, or is adjusting for individual circumstances essential to justice? Similar debates exist in virtually every professional field. As there are no consistently right or wrong answers, changes in policy are often a matter of which conscious biases prevail at any one time or in any one situation.

Conscious bias will be of increasing importance in the design of machine-learning systems going forward. Developers will almost certainly improve the quality of the data they are relying on. But when this data still produces results that society deems "problematic," what should be done? It's not hard to imagine a future where AI developers are expected to train their systems to produce more socially desirable outcomes. In this scenario, AI applications that were originally designed to mitigate "bad" unconscious human biases may soon be expected to entrench "good" conscious ones. Such systems will no longer just have to outrun the bear; they will be the bear, as human and machine decision-making coevolve, and increasingly become one and the same.

About This Series

ITIF's "Defending Digital" series examines popular criticisms, complaints, and policy indictments against the tech industry to assess their validity, correct factual errors, and debunk outright myths. Our goal in this series is not to defend tech reflexively or categorically, but to scrutinize widely echoed claims that are driving the most consequential debates in tech policy. Before enacting new laws and regulations, it's important to ask: Do these claims hold water?

About the Author

David Moschella is a non-resident senior fellow at ITIF. Previously, he was head of research at the Leading Edge Forum, where he explored the global impact of digital technologies, with a particular focus on disruptive business models, industry restructuring and machine intelligence. Before that, David was the worldwide research director for IDC, the largest market analysis firm in the information technology industry. His books include *Seeing Digital – A Visual Guide to the Industries, Organizations, and Careers of the 2020s* (DXC, 2018), *Customer-Driven IT* (Harvard Business School Press, 2003), and *Waves of Power* (Amacom, 1997).

About ITIF

The Information Technology and Innovation Foundation (ITIF) is an independent, nonprofit, nonpartisan research and educational institute focusing on the intersection of technological innovation and public policy. Recognized by its peers in the think tank community as the global center of excellence for science and technology policy, ITIF's mission is to formulate and promote policy solutions that accelerate innovation and boost productivity to spur growth, opportunity, and progress. For more information, visit itif.org.

ENDNOTES

1. Robert D. Atkinson, "'It's Going to Kill Us!' and Other Myths About the Future of Artificial Intelligence" (ITIF, June 2016), https://itif.org/publications/2016/06/06/its-going-kill-us-and-other-myths-about-future-artificial-intelligence.
2. Joshua New and Daniel Castro, "How Policymakers Can Foster Algorithmic Accountability" (Center for Data Innovation, May 2018), https://datainnovation.org/2018/05/how-policymakers-can-foster-algorithmic-accountability/.
3. Michael McLaughlin and Daniel Castro, "The Critics Were Wrong: NIST Data Shows the Best Facial Recognition Algorithms Are Neither Racist Nor Sexist" (ITIF, January 2020), https://itif.org/publications/2020/01/27/critics-were-wrong-nist-data-shows-best-facial-recognition-algorithms.