BERKELEY ROUNDTABLE ON
THE INTERNATIONAL ECONOMY

BRIE Working Paper
2019-4

ALGORITHMS, PLATFORMS, AND ETHNIC BIAS: A DIAGNOSTIC MODEL

Selena Silva and Martin Kenney

Forthcoming in *Communications of the Association of Computing Machinery*

November 2019

Algorithms, Platforms, and Ethnic Bias: A Diagnostic Model

Selena Silva
Research Assistant

and

Martin Kenney
Distinguished Professor
Community and Regional Development Program
University of California, Davis
Davis
&
Research Director
Berkeley Roundtable on the International Economy

* Corresponding Author

Keywords: algorithms, digital bias, digital discrimination, platform economy, racism

Ethnic and other biases are increasingly recognized as a problem that plagues software algorithms and datasets.[1] This is important because algorithms and digital platforms organize ever-greater areas of social, political, and economic life. Algorithms already sift through expanding datasets to provide credit ratings, serve personalized advertisements, match individuals on dating sites, flag unusual credit card transactions, recommend news articles, determine mortgage qualification, predict the locations and perpetrators of future crimes, parse résumés, rank job candidates, assist in bail or probation proceedings, and perform a wide variety of other tasks. Digital platforms are composed of algorithms executed in software. In performing these functions, as Lawrence Lessig observed, "code" functions like law to structure human activity. Algorithms and online platforms are not neutral; they are built to frame and drive actions.[2]

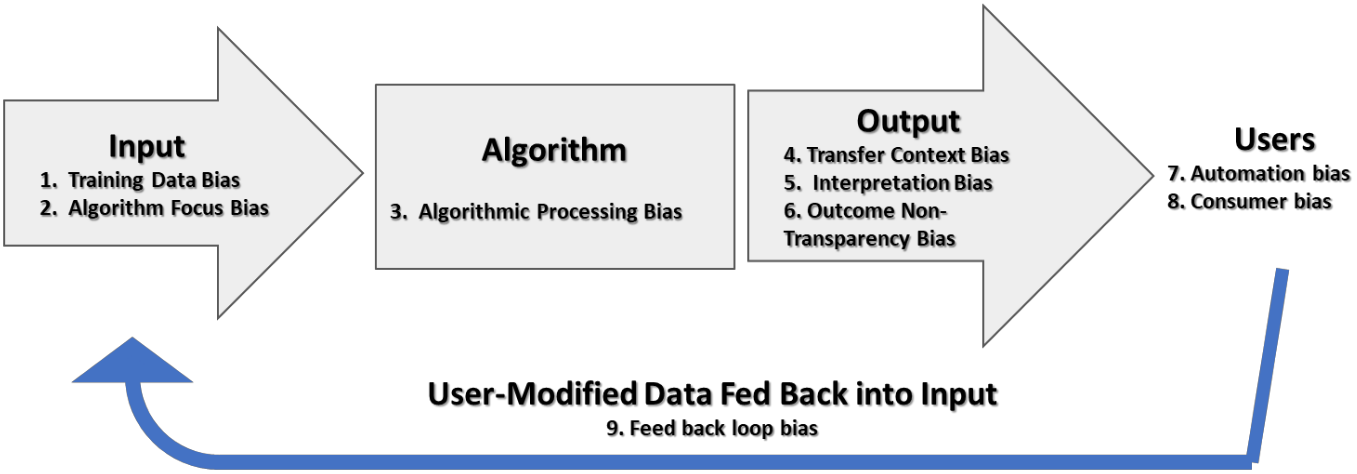
Algorithmic "machines" are built with specific theories about the correspondences between persons and things in mind. Concerns are becoming more acute, as techniques such as machine learning, are more generally deployed. For engineers and policy makers alike, understanding how and where bias occurs in algorithmic processes can help address it. Our contribution is the introduction of a visual model (Figure 1) that extends previous research to identify where in an algorithmic process bias may occur.[3]

Interrogating Bias in Algorithmic Decision-Making

The source of societal bias has been long discussed. Some attribute the introduction of bias to the fact that software developers are not well versed in issues such as civil rights and fairness.[4] Others suggest that it is far more deeply embedded in society and its expressions.[5] Our model cannot resolve such questions; however, it provides a template for identifying and addressing the sources of bias, conscious or unconscious, that stem from various actors. What is certain is that without proper mitigation, preexisting societal bias becomes embedded in the algorithms that make or influence real-world decisions.

We model algorithm development, implementation, and use from a value chain perspective with five distinct nodes—input, algorithmic operations, output, users, and feedback. Importantly, we incorporate users because their actions affect outcomes. As shown in Figure 1, we identify nine potential biases. They are not mutually exclusive, as it is possible for multiple, interacting biases to exist in a single algorithmic process.

Figure 1: Potential Biases and Where They May Be Introduced in the Algorithmic Value Chain



Source: The first six biases were adapted from Danks, D., & London, A. J. (2017). The visualization and remaining material are by Silva and Kenney.

Types of Bias

1. Training Data Bias

Predictive algorithms are trained on datasets, thus any biases in the training data will be reflected in the algorithm. In principle, this bias should be easy to detect, but the sources may be difficult to detect. Presumed gold standard datasets, such as government statistics or even judicial conviction rates, frequently contain bias. For example, if the criminal justice system is biased, then, absent corrections, the algorithm will mirror this bias. Training sets are hidden contributors to bias.

2. Algorithmic Focus Bias

Algorithmic focus bias occurs from both the inclusion and exclusion of particular variables. For instance, the exclusion of gender or race in a health diagnostic algorithm can lead to inaccurate or even harmful conclusions. However, the inclusion of gender, race, or even zip codes in a sentencing algorithm can lead to discrimination against protected groups. Yet, in certain cases, such variables must intentionally be used to weigh groups differently in order to produce a less-biased outcome.[6]

3. Algorithmic Processing Bias

Bias can be embedded in the algorithm itself. One source of such bias is the inclusion and weighting of particular variables. Consider the case of a firm's chief scientist's finding that "one solid predictor of strong coding is an affinity for a particular Japanese manga site."[7] If this is embodied in job candidate sorting software, then this seemingly innocuous choice might exclude particular qualified candidates. Effectively, a desired proxy trait inadvertently excludes certain groups that could perform the job.

4. Transfer Context Bias

Transfer context bias occurs when algorithmic output is applied to an inappropriate or unintended context. One example is using credit scores to make hiring decisions. Bad credit is equated with inferior future job performance, despite little evidence that credit scores are related to work performance. If the undesirable, but irrelevant trait is correlated with ethnicity, then it can lead to biased outcomes.

5. Interpretation Bias

Interpretation bias arises when users interpret algorithmic outputs according to their internalized biases. For example, a judge can receive an algorithmically generated recidivism prediction score and decide on the punishment or bail amount for the defendant. Because individual judges may be unconsciously biased, they may use the score as "scientific" justification for a biased decision.

6. Outcome Non-Transparency Bias

Algorithms, particularly artificial intelligence and machine learning, often generate opaque results. The reasons for the result may even be inexplicable to the algorithm's creators or the software's owner. For example, when a machine-learning program recommends denial of a loan application, the bank may not know the exact reasons. The absence of transparency makes it difficult for the subjects of these decisions to identify discriminatory outcomes or even the reasons for the outcome.

7. Automation Bias

Automation bias results from the belief that the output is fact, rather than a prediction with a confidence level. For instance, credit decisions are now fully automated and use group aggregates and personal credit history.[8] The algorithm identifies certain people as having lower credit scores and then limits their access to credit. The denial of credit means their scores cannot improve. Often, the subjects and decision-makers are unaware of the algorithm's assumptions and uncritically accept the decisions.

8. Consumer Bias

The biases that human beings act upon in everyday life are expressed in their online activities. Further, digital platforms can exacerbate or give expression to latent bias in online behavior. Users may consciously or unconsciously discriminate on the basis of a user profile that contains ethnically identifiable characteristics. Consumer bias can occur from either side, or party, in a digital interaction. Or, even more deliberately, users taught Microsoft's Tay chatter bot, which existed for only a few days in 2016, to respond with racially objectionable statements. Effectively, the algorithm or platform provides users with a new venue within which to express their biases.

9. Feedback Loop Bias

Algorithmic systems create a data trail. For example, the Google Search algorithm responds to and records a query that becomes customized input for succeeding searches. The algorithm learns from user behavior. For example, in predictive policing, the algorithm relies almost entirely on historical crime data. Suppose the algorithm sends police officers into a neighborhood to prevent crime. Not surprisingly, increased police presence leads to higher crime detection, thereby raising the statistical crime rate. This can motivate the dispatch of more police, who make more arrests, thereby initiating a feedback loop. In another example, Google Search can learn that ethnically biased websites are often selected and therefore recommend them more often, thereby propagating them. As smart as algorithms can be, human monitoring continues to be necessary.

Benefits of Platforms and Algorithms

The potential benefits of algorithmic decision-making are less noticed, but they can also be used to decrease social bias. It is well known that members of the law enforcement community make decisions that are affected by a defendant's "demeanor," dress, and other characteristics that may correlate with ethnicity. Solutions are possible. For example, Kleinberg et al. created a machine-learning algorithm that could do a better job than judges in making bail decisions.[9] The algorithm was optimized to reduce ethnic disparities among those who were incarcerated while also reducing the rate of reoffending. This optimization was possible because a disproportionately high number of people in certain racial groups are incarcerated. The point is that it is possible to design algorithms with different social goals. Critics ignore the fact the data and tools can be used to decrease inequity and improve efficiency and effectiveness.

Because algorithms are machines, they can be redesigned to improve outcomes. To illustrate, sales websites could reengineer a site to, for example, provide greater anonymity and thus reduce opportunities for consumer bias. Because all digital activities leave records, it is easier to detect biased behavior and thus reduce it. For example, a government agency could study online behavioral patterns to identify biased behavior. If it can be identified, then it can be prevented. For example, it would be easy to assess whether consumers are biased in their evaluations of online vendors and impose a standardization algorithm to mitigate such bias. Thus, while platforms and algorithms can be used in a discriminatory manner, they also can be studied to expose and address bias.

Conclusion

Computer scientists have a unique challenge and opportunity to use their skills to address the serious social problem of bias. We contribute to this by developing a readily understandable visual model for identifying where bias might emerge in the complex interaction between algorithms and humans. Here we focus on ethnic bias, but it is possible to extend the model to other kinds of bias. The model can be particularly useful in policy discussions to explain to policy makers and laypeople where a particular initiative could have an impact and what would not be addressed.

Interest in mitigating algorithmic bias has increased, but “correcting” the data to increase fairness can be hampered by determining what is “fair.” Some have suggested that transparency would provide protection against bias and other socially undesirable outcomes.[10] Unsurprisingly, firms resist transparency, maintaining that revelation of their data and algorithms could allow other actors to game their systems. In many cases, this response is valid, yet it is also self-serving as it prevents scrutiny. Software developers often cannot provide definitive explanations of complex algorithmic outcomes, meaning that transparency alone may be unable to provide accountability. Further, a single algorithmic model may contain multiple sources of bias that interact, creating greater difficulty in tracing its source. However, even in such cases, outcomes can be tested to discover evidence of potential bias.

Platforms, algorithms, software, data-driven decision-making, and machine learning are shaping choices, alternatives, and outcomes. It is vital to understand where and how social ills such as ethnic bias can be expressed and reinforced by digital technologies. Algorithmic bias can be addressed and, for this reason, critics who suggest that these technologies necessarily will exacerbate bias are too pessimistic. Digital processes create a record that can be examined and analyzed with software tools. In the analog world, ethnic or other kinds of discrimination were difficult and expensive to study and identify. In the digital world, the data captured is often permanent and can be analyzed with existing techniques. Although digital technologies have the potential to reinforce old biases with new tools, they can also help identify and monitor progress in addressing ethnic bias.

[1] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York: Broadway Books, 2016); Selena Silva and Martin Kenney, "Algorithms, Platforms, and Ethnic Bias: An Integrative Essay," *Phylon: The Clark Atlanta University Review of Race and Culture*, 55, nos. 1-2 (2018): 9–37.

[2] Lawrence Lessig, *Code: And Other Laws of Cyberspace* (2009), ReadHowYouWant.com.

[3] David Danks and Alex J. London, "Algorithmic Bias in Autonomous Systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (August 2017), pp. 4691–4697.

[4] Solon Barocas et al., "Big Data, Data Science, and Civil Rights," arXiv preprint arXiv:1706.03102 (2017).

[5] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356, no. 6334 (2017): 183–186, https://doi.org/10.1126/science.aal4230

[6] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta, "Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification," *Big Data* 5, no. 2 (2017): 120–134.

[7] Don Peck, "They're Watching You at Work," *The Atlantic* (December 2013), https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.

[8] Tom Zarsky, "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making," *Science, Technology, & Human Values* 41, no. 1 (2016): 118–132.

[9] Jon Kleinberg et al., "Human Decisions and Machine Predictions," *Quarterly Journal of Economics* 133, no. 1 (2017): 237–293.

[10] Mike Ananny and Kate Crawford, "Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," *New Media & Society* 20, no. 3 (2018): 973–989.