

Technical Perspective

The Impact of Auditing for Algorithmic Bias

By Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor

IF YOU READ news articles on the ethics of AI, you will repeatedly see the phrase “algorithmic bias” popping up. It refers to algorithms producing results that appear racist or sexist or displaying other forms of unfair bias. For example, when Amazon built a machine-learning model to score job applications, trained on historical hiring data at the company, it discovered the system downgraded female applicants—or anyone who mentioned activities associated with women (coach of women’s soccer league, for example).^a In a groundbreaking paper, researchers Joy Buolamwini and Timnit Gebru^b documented substantial discrepancies across skin types and the reliability with which three commercial face classifiers could classify gender in facial images: these classifiers could identify gender with high (99%) accuracy for white men but accuracy for darker-skinned men and all women was lower, with errors as high as 35% on darker-skinned women.


The term “algorithmic bias” is used to refer to these unwarranted or unfair differential results on different groups. Of course, the algorithm itself is not biased, in the sense it is a mathematical object with no views about the world or fairness. The bias is something that we humans attribute to how the algorithm and its model function. Often, the unfair treatment is a consequence of training the model on biased data chosen or generated by humans. The data used to train the Amazon job application classifier was drawn from the history of Amazon hiring decisions, which apparently included a bias against women. The model trained on that data recognized a pattern in human behavior and reproduced—or maybe even

amplified—it.^c What Buolamwini and Gebru’s groundbreaking 2018 “Gender Shades” study drew attention to was the importance of representative training data for facial recognition classifiers. They demonstrated this by conducting an algorithmic audit, testing the performance of classifiers on a benchmark dataset (the Pilot Parliaments Benchmark they constructed) with explicit attention to representation across classes of skin color and gender.

The audit that Buolamwini and Gebru published brought the risk of bias in commercial facial classification software, which was already sold in the marketplace (by Microsoft, IBM, and Face++), into public view. It showed a technique for identifying and reducing these errors, giving developers, companies, and regulators tools to assess at least one type of possible bias in AI models.

The goal of the research described in the following paper by Inioluwa Deborah Raji and Joy Buolamwini was to see whether and how companies producing commercial facial classification software responded to the publication of the Gender Shades results. The broader question is whether the practice of algorithmic auditing can help reduce algorithmic bias in the world. Raji and Buolamwini conducted a similar audit, using the Pilot Parliaments Benchmark, of the classifiers available in 2018 from the three companies that were originally audited in 2017, and of two other companies (Amazon and Kairos) that had not been part of the original study. What they found is the originally audited companies had released new classifiers with significantly reduced error rates on non-white-male faces. The companies that had not been originally audited, however, still had much higher error rates for non-white-male faces, as high as 31% on

darker-skinned females. This study does not prove that the original audits caused the audited companies to respond, but it clearly showed it was possible to reduce bias and that routine AI development processes would benefit from increased attention to the risk of bias. In a blog post published after the original audit was released, Microsoft discussed how they had modified training data and testing practices to ensure their facial recognition software performed comparably on different skin tones and genders.^d

What this paper demonstrates is twofold: First is the potential for a new mode of computing research that serves an independent auditing function for algorithmic systems, perhaps akin to work in cybersecurity or safety engineering research that looks to identify, probe, measure, and spread knowledge of common system vulnerabilities to motivate remedial efforts. The second is the potential of this kind of computing research to inform and suggest future tools and modes of algorithmic governance, both internal (corporate audits or product development protocols, such as requirements for intersectional bias testing) and external (independent algorithmic audits by regulators, especially for high-risk systems). The impact of this paper is its demonstration of computing research potential to do more than propose novel techniques or results; it can probe and expose the limitations of systems already in use and impacting people’s lives, with an eye to raising the technical and professional standards for computing excellence. 

d <http://bit.ly/3fL7129>

Vincent Conitzer, Carnegie Mellon University’s Foundations of Cooperative AI Lab and the University of Oxford’s Institute for Ethics in AI.

Gillian K. Hadfield, University of Toronto’s Schwartz Reisman Institute for Technology and Society and the Vector Institute for Artificial Intelligence.

Shannon Vallor, The University of Edinburgh’s Centre for Technomoral Futures.

Copyright held by authors.

a <https://reut.rs/3NRwQRe>

b J. Buolamwini and T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 2018 Conf. Fairness, Accountability and Transparency*.

c The potential for an ML system to show bias that is greater than what occurs in the training data has been shown in Leino et al.’s Feature-wise Bias Amplification. In *Proceedings of ICLR 2019*; <https://arxiv.org/pdf/1812.08999.pdf>