Algorithmic Bias

Artificial intelligence plays an increasingly pervasive role in society: it determines credit scores, whether an individual receives a loan, how many police officers patrol a neighborhood, and much more.[i] A central component of an artificial intelligence system is the algorithm, a step-by-step procedure that enables a computer to perform calculation, data processing, and automated reasoning tasks.[ii] Today, most applications of artificial intelligence are based on deep learning algorithms,[iii] with which computers can find and amplify patterns from vast amounts of data and produce predictions that inform future algorithmic decisions.[iv]

Flawed Data Produce Algorithmic Bias

In spite of significant advances in artificial intelligence, a growing body of research indicates that algorithms are capable of amplifying real-world biases. Algorithmic bias can be caused by multiple factors and during various stages of the deep learning process, but one particularly concerning factor is the use of training data that are unrepresentative of the United States population and training data that reflect historical inequalities.[v]

The harm produced by the use of unrepresentative data is illustrated in the case of facial recognition technology. In an MIT Media Lab study, computer systems using facial images to recognize skin color and gender could correctly classify light-skinned men 99% of the time, but could only correctly classify dark-skinned women as little as 65% of the time.[vi] According to the researchers, the root of this problem is existing benchmark data sets, which tend to overrepresent light-skinned men and underrepresent darker-skinned people in general.[vii] For example, the Labeled Faces in the Wild (LFW) data set, which is composed of celebrity faces and has served as a gold standard benchmark for facial recognition technology, was estimated to be 77.5% male and 83.5% white.[viii] If facial recognition systems are trained on data sets that fail to reflect the entire population, such as the LFW data set, then algorithmic learning will be skewed toward those specific characteristics.

In addition, the use of data reflecting historical inequalities can lead to the perpetuation of prejudice against marginalized groups. For instance, Amazon's now-discontinued recruiting algorithm extracted data from resumes submitted to the company over a ten-year period.[ix] The algorithm was programmed to recognize word patterns in the resumes, which were submitted predominantly by white men and compared to the company's majority-male engineering department. The focus on word patterns, rather than relevant skill sets, contributed to the penalization of resumes that contained the word "women's" and the downgrading of resumes submitted by women who had attended women's colleges.[x] By hurting female applicants' chances of being hired, the recruiting algorithm solidified gender bias at a company whose

global gender balance, as of December 2018, is 58.3% male and where men hold 73.2% of managerial roles.[xi]

Impact on Communities of Color

As algorithmic decision-making becomes increasingly ubiquitous, there is more and more evidence that algorithms can perpetuate negative stereotypes about communities of color. For example, typing "Asian girls" or "Latina girls" into a Google search bar leads to countless search results and images that fetishize and sexualize women from these racial backgrounds.[xii] In addition, a software engineer discovered in 2015 that Google Photos' image recognition algorithms were classifying his African American friends as "gorillas."[xiii]

Other forms of algorithmic bias are more overtly malicious, such as the case of facial recognition technology. Building on the MIT Media Lab study's findings, a study released by the National Institute of Standards and Technology (NIST) in December 2019 revealed higher rates of false positives for people of color when NIST researchers tested 189 different facial recognition algorithms' abilities to match a photo of a person to another photo of the same person in a database ("one-to-one" matching) and to determine whether the person in a photo has any matches in a database ("one-to-many" matching).[xiv] Higher error rates could put individuals at a higher risk of experiencing data security breaches and/or being falsely accused of a crime. The researchers ultimately suggested that there is a link between an algorithm's performance and the data on which it was trained, identifying "more diverse training data" as one of several factors that "may prove effective at mitigating demographic differentials with respect to false positives."[xv]

Algorithmic bias also impacts health care. Publicly available medical data sets tend to overrepresent white men, especially in the United States.[xvi] In the study of human genomics, which examines the structure, function, evolution, and mapping of human genomes, a 2016 meta-analysis of 2,511 studies from across the world found that 81% of participants in genome-mapping studies were of European descent – meaning that researchers who download publicly-available data to study disease are far less likely to use the genomic data of people of African, Asian, Hispanic, or Middle Eastern descent.[xvii] In this application, algorithmic bias "can lead to the recapitulation of longstanding health disparities" for people of color.[xviii]

In light of these examples of biased algorithmic decision-making, it is crucial that internet-based companies disclose how their algorithms process personal data, invest in implicit bias and diversity trainings for their employees, and implement other internal policies that promote transparency and equity. Without such measures, applications of artificial intelligence that are based on flawed data will continue to disproportionately harm communities of color.

---

[i] "Life in a Quantified Society," Open Society Foundations (May 2019), https://www.opensocietyfoundations.org/explainers/life-quantified-society.

[ii] Stephen F. DeAngelis, "Artificial Intelligence: How Algorithms Make Systems Smart," WIRED, https://www.wired.com/insights/2014/09/artificial-intelligence-algorithms-2/.

[iii] Karen Hao, "This is how AI bias really happens – and why it's so hard to fix," MIT Technology Review (Feb. 4, 2019), https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

[iv] Karen Hao, "What is machine learning?" MIT Technology Review (Nov. 17, 2018), https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/#deep-learning.

[v] Nicol Turner Lee, Paul Resnick, and Genie Barton, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” Brookings Institution (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.
[vi] Joy Buolamwini and Timnit Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification,” http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.
[vii] *Ibid.*
[viii] *Ibid.*
[ix] Nicol Turner Lee, Paul Resnick, and Genie Barton, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” Brookings Institution (May 22, 2019), https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/.
[x] *Ibid.*
[xi] “Our workforce data,” Amazon, https://www.aboutamazon.com/working-at-amazon/diversity-and-inclusion/our-workforce-data.
[xii] Jonathan Cohn, “Google’s algorithms discriminate against women and people of color,” The Conversation (Apr. 24, 2019), http://theconversation.com/googles-algorithms-discriminate-against-women-and-people-of-colour-112516.
[xiii] James Vincent, “Google ‘fixed’ its racist algorithm by removing gorillas from its image-labeling tech,” The Verge (Jan. 12, 2018), https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai.
[xiv] “NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software,” National Institute of Standards and Technology (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software.
[xv] Patrick Grother, Mei Ngan, and Kayee Hanaoka, “Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects,” (Dec. 2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf.
[xvi] Dave Gershgorn, “If AI is going to be the world’s doctor, it needs better textbooks,” Quartz (Sep. 6, 2018), https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.
[xvii] Dave Gershgorn, “If AI is going to be the world’s doctor, it needs better textbooks,” Quartz (Sep. 6, 2018), https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.
[xviii] Kadija Ferryman and Mikaela Pitcan, “Fairness in Precision Medicine,” Data & Society (Feb. 2018), https://datasociety.net/wp-content/uploads/2018/02/Data.Society.Fairness.In_.Precision.Medicine.Feb2018.FINAL-2.26.18.pdf.