STANFORD BUSINESS GRADUATE SCHOOL OF

CASE: ETH-23
DATE: 02/19/20

DESIGNING AI FOR ALL:
A PRIMER ON BIAS IN ARTIFICIAL INTELLIGENCE SYSTEMS

Algorithms improve our daily lives by filtering spam emails and curating our social media feeds, but they're also used to recommend prison sentences, identify top job candidates, or guide autonomous vehicles. With each application, the risk of algorithmic bias varies from harmless inconvenience to life threatening and criminal. News reports in recent years are rife with examples of unwanted bias in algorithms used for artificial intelligence (AI):

- Amazon's hiring algorithm discriminated against female job candidates
- Google Photo identified two black people as gorillas
- Broward County's algorithm wrongly flagged black defendants at twice the average rate
- Pokemon Go effectively redlined communities of color for participants
- Apple Card offered higher credit limits to male customers regardless of credit scores
- Facial recognition software vastly underperformed for subjects with darker skin features

This case will highlight sources of algorithm bias and recommendations to address it.

WHAT IS ALGORITHMIC BIAS?

Algorithmic bias describes the presence of systematic errors encoded into models, often leading to disparate outcomes on protected classes. This bias exists in large part because algorithms are artifacts of human society. They encode historical actions, behaviors, and decisions while optimizing for our specific priorities. Algorithms have the potential to improve human decision

Thomas Higginbotham, Zoe Weinberg, Wendy De La Rosa, Professor Jennifer Aaker and Professor Fei Fei Li prepared this case as the basis for class discussion rather than to illustrate either effective or ineffective handling of an administrative situation.

making through objective, data-based insight, but also have the potential to perpetuate human biases and deploy them at scale.[1]

WHERE DOES THE BIAS COME FROM?

Unwanted bias can originate anywhere in the lifecycle of an algorithm. Here are a few common sources:

> **Training Data:** Data used to train the algorithm reflects the social, historical, and political conditions in which it was created—and these data may be biased.[2] For example, the Broward County recidivism algorithm relied on historical data that included bias against African Americans, which led to the algorithm labeling a greater proportion of black defendants as medium or high risk, compared to white defendants. Incomplete or under-representative training data or the existence of historical biases are all major sources of algorithmic bias in this stage.
>
> **Model:** Feature selection—choosing which variables to include or exclude from the model—can obscure bias in algorithms. Models that attempt to counter gender bias through blindness—removing gender from the model—can in some instances *produce* bias.[3]
>
> **Test Data:** Models are trained on data within a specific context. Some algorithms are highly context-dependent and therefore must be trained in environments that are very similar to the intended use. For instance, race and gender variables may mitigate algorithm bias in one context but when placed in the context of the test data, the result could be biased. Misunderstanding the context can be catastrophic.
>
> **Deployment:** Even an unbiased model can be deployed in a way that exacerbates unwanted bias and inequality.[4] Human oversight is needed to prevent misuse or misapplication.

WHAT CAN WE DO ABOUT IT?

Removing unwanted bias does not necessarily equate to fairness. However, the following steps are common recommendations to address bias and fairness in algorithms and artificial intelligence:

[1] This case uses the term "algorithmic bias" in artificial intelligence more specifically to imply "unwanted bias," henceforth used interchangeably with "bias." Machine learning, often considered a subcategory within artificial intelligence, is also used interchangeably for the purposes of this case.

[2] "AI Now," AI Now Institute, New York University, https://ainowinstitute.org/ (January 26, 2020).

[3] Nicole 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/ (February 18, 2020).

[4] Technically, every machine learning model is "biased" because it relies on weighted biases to work, thus it is more accurate to say a model without "unwanted biases." For the same reason, it is important to stay away from the term "de-biasing" (even though it is often used colloquially by engineers), because technically you can't ever fully "de-bias" the model; you can only work to combat / mitigate unwanted biases.

Diversity in Design: Only 12 percent of machine learning researchers are women.[5] The AI Now Institute reported that less than 5 percent of the workforce at leading technology companies are African American. Improving diversity throughout the lifespan of an algorithm is critical and a significant concern today.[6] Operators of algorithms should consider the role of diversity—and the importance of cultural sensitivity—within their work teams, training data, and decision-making processes.[7]

Preprocess and Test Data for Biases: Several methods exist to measure and mitigate/correct for unwanted bias in training data,[8] but there will usually be a tradeoff between fairness and accuracy.[9] This also requires constant monitoring throughout the deployment of the model.[10]

Detect Bias and Auditing: Rigorous testing should be required across the lifecycle of AI systems in sensitive domains. Pre-release trials, adversarial testing techniques, independent auditing to stress test algorithms for protected classes,[11] and ongoing monitoring are necessary to test for bias, discrimination, and other harms.[12]

Transparency and Explainability: Remedying bias in AI systems is almost impossible when these systems are opaque. Transparency is essential, and begins with tracking and publicizing where AI systems are used, and for what purpose.[13] Relatedly, explainability refers to the ability to trace an output to its corresponding inputs along the model, which is especially important for issues of AI governance and liability. Different approaches have been proposed, including the idea of appending Model Cards to every AI system that provide details about the system's construction and limitations.

Human Oversight: To develop equitable and trustworthy technology, we must understand how AI performs in practice, and guide and shape the way AI interacts with

[5] Tom Simonite, "AI is the Future—But Where are the Women," August 17, 2018, https://www.wired.com/story/artificial-intelligence-researchers-gender-imbalance/ (January 26, 2020).

[6] Gwen Moran, "A.I. Has a Bias Problem, and Only Humans Can Make It a Thing of the Past," October 26, 2019, https://fortune.com/2019/10/26/fixing-artificial-intelligence-bias-problem/ (January 26, 2020).

[7] Nicole Turner Lee, Paul Resnick, and Genie Barton, op. cit.

[8] Note: Mitigating bias in artificial intelligence can be done during pre-processing (reweighing, disparate impact remover, collecting additional data), in-processing (adversarial debiasing), or post-processing (equalized odds post-processing, reject option classification). See, for instance, Rachel. E.K. Bellamy, et al., "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," October 3, 2018, https://arxiv.org/pdf/1810.01943.pdf, https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies

[9] Borealis AI Blog, "Tutorial #1: bias and fairness in AI," August 19, 2019, https://www.borealisai.com/en/blog/tutorial1-bias-and-fairness-ai/ (January 26, 2020).

[10] Note: If historical data is riddled with discrimination (e.g., loan data, hiring data), a team may have to generate a new synthetic dataset from scratch, often called a "Golden Set," which is vetted by experts who are trained in unconscious bias for testing and evaluation purposes. In addition, it's important to note that "stripping out" protected categories like sex and race is not enough, given how many proxies there are for any given trait.

[11] Cade Metz, "We Teach A.I. Systems Everything, Including Our Biases," *The New York Times*, November 11, 2019, https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html (January 26, 2020).

[12] Sarah Myers West, Meredith Whittaker, and Kate Crawford, "Discriminating Systems" AI Now, April 2019, https://ainowinstitute.org/discriminatingsystems.pdf (January 26, 2020).

[13] Ibid.

humans, their vital social structures and institutions, and the international order.[14] Human judgement is still needed to ensure AI-supported decision making is fair. Where is this needed? In what form should it be provided? In what situations should a fully autonomous solution be permissible?

Consumer Literacy: Given the increased use of algorithms in many aspects of daily life, all potential subjects of automated decisions would benefit from knowledge of how these systems function. Just as computer literacy is now considered a vital skill in the modern economy, understanding how algorithms use an individual's data may soon become a necessity.[15]

Ethical Frameworks: Failing to imbue ethics into AI systems, we may be placing ourselves in a dangerous situation of allowing algorithms to decide what's best for us.[16] Designers and engineers can avoid unwanted bias by incorporating rigorous ethical reviews and fairness checks into the design and implementation process.

ADDITIONAL MATERIALS AND RESOURCES:

Read:

- Cathy O'Neill, Weapons of Math Destruction (2016)
- Brookings, Algorithmic Bias Detection and Mitigation: Best Practices and Politics to Reduce Consumer Harms (2019)
- Toward Data Science, Algorithmic Solutions to Algorithmic Bias: A Technical Guide (2019)
- Google, Machine Learning Fairness for Developers (2019)
- Sam Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness (2017)
- Martin Wattenberg et al., Attacking Discrimination with Smarter Machine Learning (2016)
- Sorelle A. Friedler, On the (Im)possibility of Fairness (2016)
- Solon Barocas, Big Data's Disparate Impact (2016)
- Blaise Agüera y Arcas, et al., Physiognomy's New Clothes (2017)
- Kate Crawford, Artificial Intelligence's White Guy Problem (2016)
- Andrew D. Selbst, et al., Fairness and Abstraction in Sociotechnical Systems (2018)

Watch:

- AI Ain't I A Woman
- Hardt and Barocas, Fairness in Machine Learning tutorial from NIPS 2017
- 21 Definitions of Fairness
- D4BL 2019 Welcome and Keynote

[14] "Human Impact," Human-Centered Artificial Intelligence, Stanford University, https://hai.stanford.edu/research/human-impact (January 26, 2020).

[15] Nicole Turner Lee, Paul Resnick, and Genie Barton, op. cit.

[16] Vyacheslav Polonski, "The Hard Problem of AI Ethics - Three Guidelines for Building Morality Into Machines," February 28, 2018, https://www.oecd-forum.org/users/80891-dr-vyacheslav-polonski/posts/30743-the-hard-problem-of-ai-ethics-three-guidelines-for-building-morality-into-machines (January 26, 2020).