Bias in Machine Learning and Artificial Intelligence

Julia Lee
Sukan Suksawang
University of California, Riverside

Bias in Machine Learning and Artificial Intelligence

Background:
The whole human endeavor is using information to extrapolate it into knowledge. Data is information; and "statistics is the study of techniques and methods that we can use to transform data into knowledge" [Baumer]. The internet has caused the emergence of big data. Data so large that it has to be processed by machines, this has caused a boom in new techniques optimize this data [Baumer]. People often see these techniques as unbiased but we will show that is not always the case.

Goal: Most people nowadays are using ML and AI. Studying bias in those areas allow us to better understand its roots and consequences while using it. We can adjust that knowledge in our daily life, as well as optimize the advantages of ML and AI, while bewaring of the flaws.

Key Terms:
Machine Learning (Statistical learning, Data mining): is a series of methods using data to find patterns and correlations to predict outcomes and to extract useful information [Sartorius]

Algorithms: A set of rules or procedures used to solve a problem. [Sartorius]

Big Data: Big data often refers to large dimensional data sets where traditional statistical methods can't be used. This has led to new techniques like data mining, machine learning and artificial intelligence [Baumer].
- many cases (rows
- many variables/measures (columns)

For example: we might want to use x1 and x2 to classify our data on y.
Image: created by Julia Lee in r

Artificial intelligence: is a branch of machine learning that uses the idea of the human brain to create a mechanical network of Algorithms. Artificial intelligence often uses unstructured big data. [Sartorius]
- Facial recognition software
- Voice recognition
- Self driving cars
- Natural language processing

Black box algorithms: You put your data in a black box and get an output but it is hard to explain how you got to the outcome. This allows for bias because it is hard to figure out what went wrong in your model (because it is hard to see how you got to a faulty outcome) .
- Lack of Transparency. [Bembeneck]
  - Neural networks
    - Deep learning
    - Models human brain
    - Hidden layers to get an outcome.
  - Support Vector machines
    - geometric algorithm that relies on many variables is hard for humans to visualize or understand.

How do machine learning algorithms work?
- Use a feature space (variables) to predict/classify Y. This is often referred to as supervised machine learning. Uses labelled data (where Y is known) to build this model to predict Y given x
  - $\hat{Y} = Ax + Bw + Cz$
- Pattern/ structure recognition of data. This is what we call unsupervised learning.

Bias:
- Algorithm bias is the idea that an algorithm can produce results that are systematically prejudiced due to the methods and data used to build it. [O'Neil]
- One underlying problem with data is our tendency to rely too heavily on them, on the theory that statistics are objective and that "numbers don't lie." [O'Neil]

How & Why Can Machine Learning Algorithms be Bias?

Data used to train the algorithm can be biased /flawed/ incomplete :

- This is also known as sampling bias (what data is being used to build these models?)
- Facial recognition software that was built only using faces of white men. [Fong]
  - This might make it harder for your software to correctly identify people of a different race or gender.
- Crime data that is collected by police but is skewed towards where police choose to patrol (often in majority minority neighborhoods) [Fong]. This introduces the cops own bias into the model. [Fong]
- The data is collected by humans and humans are flawed and biased [Fong].

The data might also have many confounding factors and high correlation (correlation ≠ causation) :

- Zip-code and race might be correlated. So someone might think their model is not biased because they did not even include race into the model but race is there because zip-code is included. This is known as **proxy variables.**
- This is also known as **label choice bias** where there is a gap between what the algorithm is trying to measure/predict compared to what it is actually saying. [Bembeneck]
- Credit scores, for example, do not necessarily prove a job candidate's trustworthiness, but may only reflect his socio-economic background.
- Another example, of this inequality is colleges using test scores to predict who should be admitted often affecting students who have less resources for test prep and inadvertently discriminating against people of color and low-income individuals.

Machine Learning algorithms are often designed to minimize error rates.

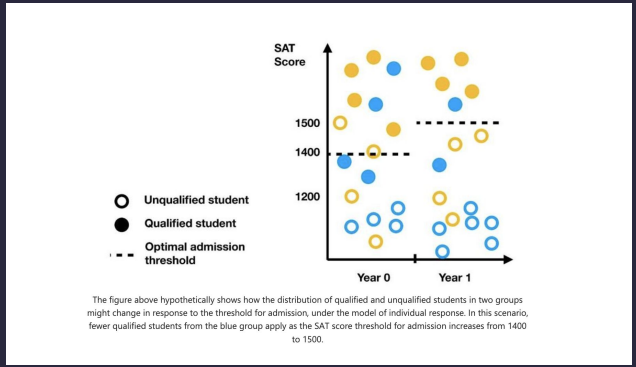
- This means that the algorithms are trained to focus on overall accuracy over "fairness"
- Bias- variance tradeoff: This is the idea of overfitting (variance) your model to the training data vs. your model making erroneous assumptions because of underfitting (bias). As you lower the bias in your model you will increase the variance.
- Many different kind of error rates that a model could be minimizing and this often only known by the person/people building the model and different error rates give you different results. [Bembeneck]

Lack of transparency

- Companies who make and use these models have no incentive to tell you how they are using your data or how the algorithm is made. [O'Neil]
- Lack of Transparency allows companies to hide behind biased models and no-one can review their product fully. [O'Neil]
- These algorithms are very complexed and use many factors to build them but often the consumer of these algorithms don't understand exactly what the algorithm is doing. [O'Neil]
- For example, the models used by banks and hedge funds often prioritize profit and that influences the models they use and the weight they put on certain variables. [O'Neil]

Model validation:

- a model becomes a belief, it becomes more hardwired, leading to a self fulfilling feedback loop. [O'Neil]
- The idea that models that are created affect us and we also affect them. [O'Neil]
- Example: the college-ranking model created by U.S. News and World Report. "The trouble was that the rankings were self-reinforcing." [O'Neil , p.59]
- Models are influenced by the priorities that the people making the models have. [O'Neil]



The figure above hypothetically shows how the distribution of qualified and unqualified students in two groups might change in response to the threshold for admission, under the model of individual response. In this scenario, fewer qualified students from the blue group apply as the SAT score threshold for admission increases from 1400 to 1500.

The figure above is an example of a model feedback loop: This example is about SAT scores how a school might have one threshold and it gives an incentive for more students of one group to apply and then in response the school might increase the threshold making it less likely for students to apply [Liu]

- This loop unfairly hurts poor and middle class students are the most disadvantaged by the rankings because they do not have the resources to pay for tests review classes or college admissions consultants, they cannot game the system the same way a privileged student can. [Liu]

Examples of Machine Learning Bias

criminal Risk Assessment:

- A software (Northpointe's software) was created to help judges figure the risk of people convicted of a crime to who will reoffend. [Angwin]
- Software uses the answers of 137 questions to give a recidivism score. [Angwin]
- Judges saw these risk scores as unbiased and used the results to wrongly sentence people. [Angwin]
- When using the algorithm for risk assessment the results had huge racial disparity [Angwin]
    - Black defendants were more likely to be wrongly labeled as high risk (**44.9% vs. 23%**)
    - White defendants were more likely to be wrongly labeled as low risk (**47.7% vs 28.0%**)

- Reasons why for this disparity:
    - Proxy variable that is correlated to race:

One of the questions was "If the person has a parent or family member in the criminal justice system?" [Angwin]

- We know that there is racial inequality in the answer to this question (that has nothing to do with recidivism) [Angwin]
- So while the model does not take into account race, the model still learns about race through other variables. [Angwin]

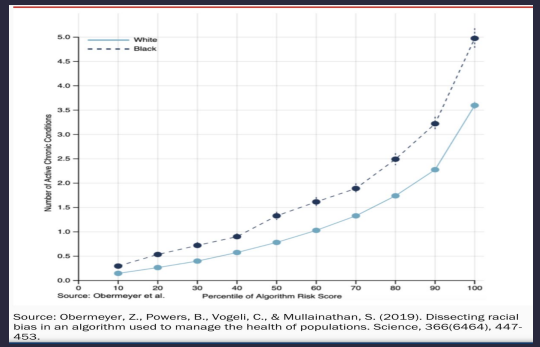
Medical Bias:

- Medical algorithm created to create risk scores for patients. [Bembeneck]
    - White and black patients were being unequally given risk scores despite having the same health status. [Bembeneck]
- The model reported to be predicting medical need but turned out to be measuring the means people had to pay for care. [Bembeneck]
- Medical costs were being used as a proxy for medical care.
- Label choice bias. [Bembeneck]

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)



Source: Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Google sentiment analysis:

- Google developed text analysis software [Lander]
- Text analysis to assign sentiment to words and phrases.
- The algorithm was trained on internet conversations randomly and indiscriminately [Lander]
- Google's algorithm assigned negative sentiment to phrases like "Jew", "Gay","Black" [Lander]
- AI algorithms for text analysis are built by training on the data that it is given.
    - If the outside text fed to the algorithm uses these phrases in a negative way it will affect the outcome.
    - All the data is being weighed equally.
- The algorithms ended up just reflecting the material it was feeding on even if it is not true. [Lander]

Facial recognition:

- Facial recognition software does a better job at recoginizing faces of white cis-gender males.[Breland]
- Police use this software to identify suspects. [Najibi]
- Recognition software is trained on data of only the faces of the engineers building the software, mainly white men. [Breland]
- Studies have shown that for faces of women of color the algorithm was the least accurate. [Najibi]

These are both examples of sampling bias where the data used to build these models are flawed and causes biased to be baked into the model. [Lander]

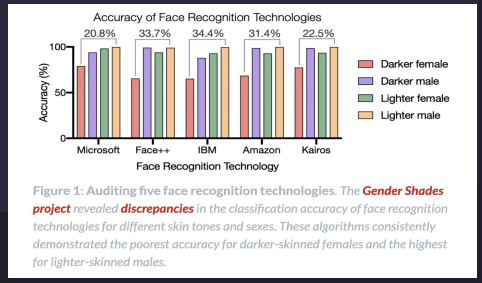


Figure 1: Auditing five face recognition technologies. *The **Gender Shades project** revealed **discrepancies** in the classification accuracy of face recognition technologies for different skin tones and sexes. These algorithms consistently demonstrated the poorest accuracy for darker-skinned females and the highest for lighter-skinned males.*

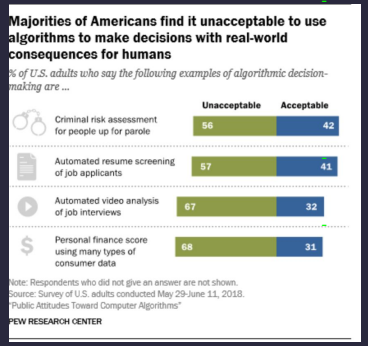
Response Towards Algorithms & Bias

Public perceptions toward algorithms:

- **A pew research 2018 poll found that majority of Americans thought it was unacceptable for algorithms to make decisions in that impact their lives across many topics: criminal justice, job applications,interviews, and finances**
- **Respondents gave these reasons for their distrust:**
  - "They violate privacy "[Smith] : This was the top concern for those worried about personal finance scores and consumer data, with 26% of respondents citing this as their reason. [Smith]
  - "They are unfair" [smith]: This was more of a concern for those worried about personal finances but also cited by those concerned about an automated employment system. [Smith]
  - "Humans are complex, and these systems are incapable of capturing nuance."[Smith]: This was the top concern for those who worried about resume screening and interviews. [Smith]
  - "They remove the human element from important decisions."[Smith]: this reason was a concern for those who found criminal risk scores unacceptable and over 50% of respondents for all situations said this was a concern. [Smith]

Industry/ academia response:

- Recent push in the industry and in academia to focus on the topic of data ethics. [Bloomberg]
- Data scientists have launched a initiative, "Community Principles on Ethical Data Sharing",.with Bloomberg to create a "hippocratic oath " for data scientists. [Bloomberg]
- This initiative with the help of 2000 data scientists and computer scientists have released some areas of ethics and guidelines: [Bloomberg]
  - Data collection: minimize bias in data & focus on the collection and storage of data. [Bloomberg]
  - Diversity: data practitioners should foster diversity and obtain a representation of viewpoints, and communities and openness. [Bloomberg]
  - Transparency: data practitioners should be transparent, and provide enough context and documentation in order for others to evaluate the data. It also provides that practitioners have a duty to communicate responsibly, "acknowledging and disclosing caveats and limitations." [Bloomberg]
  - Do No Harm: Scientists should consider the possible benefit or harm to society and mitigate the harm. [Bloomberg]



Government's response to A.I. concerns:

- In May of 2021, Detroit passed an ordinance to require more transparency in the city's use of AI and surveillance software like the police's use of facial recognition. [O'Brien 2021]

New York City, in November of 2021, became the first city to pass a piece of legislation that aims to limit the use of artificial intelligence in the hiring process. [O'Brien 2021]

- Limit what technology a employer can use in the hiring process
- Forces AI tools to give more transparency to allow people to opt in or out of its use. [O'Brien 2021]

In May of 2022, The U.S. Justice Department warned employers of the possible bias of A.I. algorithms used to make hiring decisions. [O'Brien 2022]

The Biden White House in October of 2021 proposed creating possible Bill of Rights for A.I. [O'Brien 2021]

- Develop new laws and regulations to protect against algorithms that might unfairly affect people's lives. [Lander]
- Follows proposed European Union regulations [O'Brien 2021]
  - Ban some harmful uses of A.I
  - Regulate the use of facial recognition
  - Algorithmic transparency
  - Quality Assurance

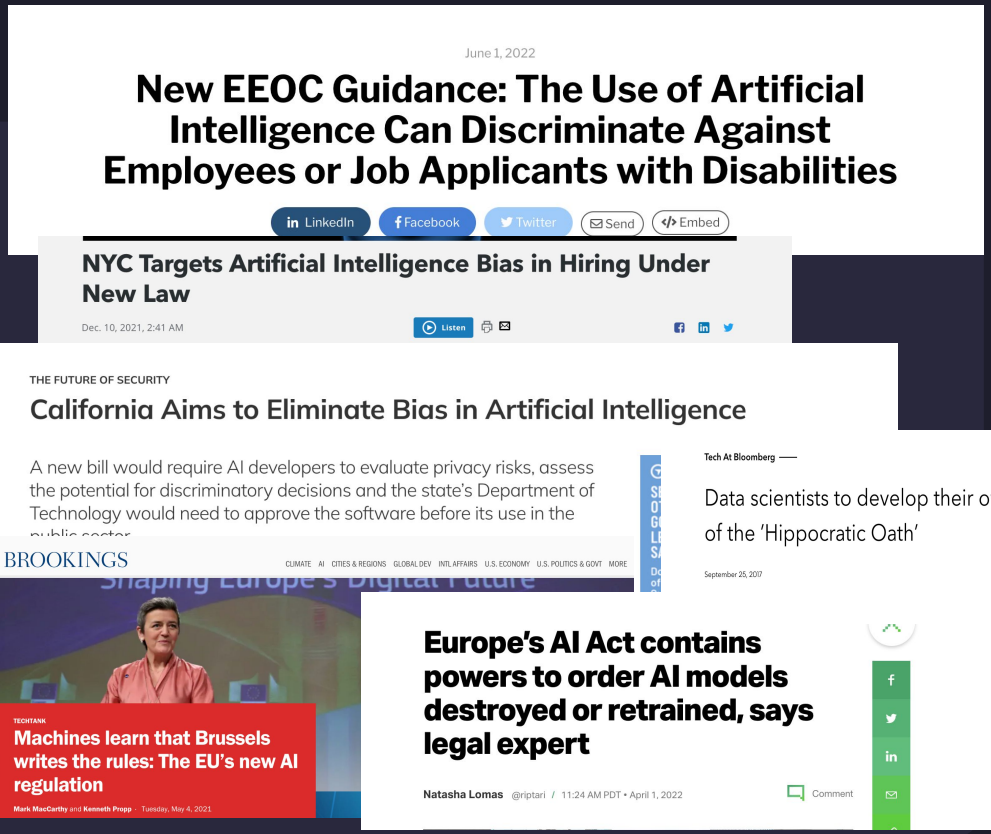
Germany in 2021 wrote the first framework to check for quality assurance of algorithms and their development process. Germany also wrote out a framework and best practices for how AI should be tested and created. [O'Brien 2021]

Conclusion:

Why is there bias in AI:
- Sampling bias /Faulty data
- Use of proxy variables
- Label choice bias
- Model validation feedback loop.

How can we combat bias in AI:
- Transparency about the data used to train the algorithm
- Oversight/regulation
- Diversity in the people making the algorithms

Good News:
- Governments are starting to write regulations on the use of AI and tech companies that make these programs
- Industry is becoming more aware of bias and developing data ethics and guidelines.
- Public is more wary of the use of these algorithms in recent years



References:

Benjamin S. Baumer, Daniel T. Kaplan. “Modern Data Science with R.” Chapter 1 Prologue: Why Data Science?, 28 July 2021, https://mdsr-book.github.io/mdsr2e/ch-prologue.html#what-is-data-science.

Bembenek ,Emily, et al. “To Stop Algorithmic Bias, We First Have to Define It.” Brookings, Brookings, 9 Mar. 2022, https://www.brookings.edu/research/to-stop-algorithmic-bias-we-first-have-to-define-it/.

Bloomberg.com, Bloomberg, 25 Sept. 2017, https://www.bloomberg.com/company/stories/data-scientists-develop-version-hippocratic-oath/.

Breland, Ali. “How White Engineers Built Racist Code – and Why It's Dangerous for Black People.” The Guardian, Guardian News and Media, 4 Dec. 2017, https://www.theguardian.com/technology/2017/dec/04/racist-facial-recognition-white-coders-black-people-police.

Doty, Chris. “Detecting and Reducing Bias in Speech Recognition.” Deepgram, 5 Apr. 2022, https://deepgram.com/blog/detecting-and-reducing-bias-in-speech-recognition/.

Fong, Joss. “Are We Automating Racism?” Vox, Vox, 31 Mar. 2021, https://www.vox.com/videos/2021/3/31/22348722/ai-bias-racial-machine-learning.

Julia Angwin, Jeff Larson. “Machine Bias.” *ProPublica*, 23 May 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Knight, Will. “The Foundations of Ai Are Riddled with Errors.” Wired, Conde Nast, 31 Mar. 2021, https://www.wired.com/story/foundations-ai-riddled-errors/.

Lander, Eric. “Americans Need a Bill of Rights for an AI-Powered World.” Wired, Conde Nast, 8 Oct. 2021, https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/.

References:

Liu, Lydia T. “Negative Feedback Loops: Using an Economic Model to Inspect Bias in AI.” Microsoft Research, 21 Jan. 2020, https://www.microsoft.com/en-us/research/blog/when-bias-begets-bias-a-source-of-negative-feedback-loops-in-ai-systems/.

Najibi, Alex. “Racial Discrimination in Face Recognition Technology.” Science in the News, 26 Oct. 2020, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.

O'brien, Matt. “U.S. Civil Rights Enforcers Warn Employers against Biased AI.” AP NEWS, Associated Press, 12 May 2022, https://apnews.com/article/technology-discrimination-artificial-intelligence-e1bcf4a2e7f1b671cbf3a44bc99b3656.

O'brien, Matt. “White House Proposes Tech 'Bill of Rights' to Limit Ai Harms.” AP NEWS, Associated Press, 8 Oct. 2021, https://apnews.com/article/joe-biden-science-technology-business-biometrics-b9dbf5fee3bf0e407b988b31e21f5300.

O’Neil, Cathy, Weapons of Math Destruction. New York, New York, Crown, 2016.

Smith, Aaron. “Public Attitudes toward Computer Algorithms.” *Pew Research Center: Internet, Science & Tech*, Pew Research Center, 7 July 2020, https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

Sartorius. “Understanding the Relationship between Data Science, Artificial Intelligence and Machine Learning.” Sartorius, 15 July 2020, https://www.sartorius.com/en/knowledge/science-snippets/data-science-vs-artificial-intelligence- vs-machine-learning-602514.

Extra credit reference

Alelyani, Salem. “Detection and Evaluation of Machine Learning Bias.” *Applied Sciences*, vol. 11, no. 14, July 2021, p. 6271. *Crossref*, https://doi.org/10.3390/app11146271.

Examining Bias in Machine Learning and Algorithms (Extra Credit)

Root causes of biases

- It is derived from training data, which is derived from cognitive bias in human decisions and judgements.
- Bias in the research model comes from the statistical priory in the learning data

Bias detection

- It is possible to think of it as fundamental data qualities inherited through human behavior and practice.
- The research used the scientific methods to evaluate factors that have the quality of PBAs (Potentially biased attributed)
    - PBAs are used to detect bias by alternating their values.
    - A biased attribute is one that dramatically changes predicted class values after using the alternation function.
    - The research calculates the divergence between the original and alternated mean of predicted class values with respect to each attribute's value to determine the amount of bias.

Reference
Alelyani, Salem. “Detection and Evaluation of Machine Learning Bias.” *Applied Sciences*, vol. 11, no. 14, July 2021, p. 6271. *Crossref*, https://doi.org/10.3390/app11146271.

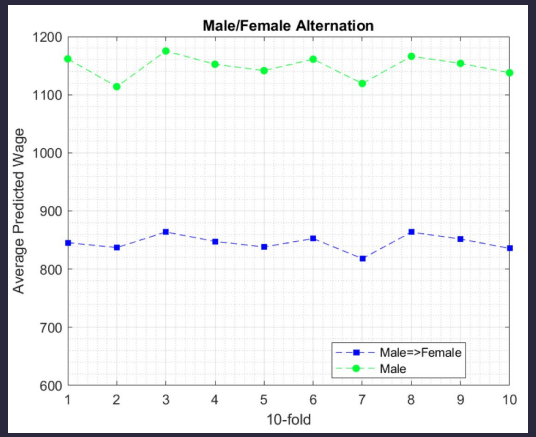
Examining Bias in Machine Learning and Algorithms (Extra Credit)

Example of methodologies

$$\gamma(\mathbf{u},\mathbf{v}) \neq \gamma(\mathbf{v},\mathbf{u})$$

- The amount of bias is indicated by this equation. The greater the difference, the greater the bias in favor of or against a particular value.

Example of findings



Male to female alternation. The wages of males (in green) decrease when we change the gender attribute to a female (in blue).

Reference

Alelyani, Salem. “Detection and Evaluation of Machine Learning Bias.” *Applied Sciences*, vol. 11, no. 14, July 2021, p. 6271. *Crossref*, https://doi.org/10.3390/app11146271.