

Bias in Artificial Intelligence

Seminar-Human in the age of Artificial Intelligence

August 16, 2022

Presented by Abdullah Amawi. University of Göttingen

Table of contents

- Introduction
 - What is Bias?
 - What is AI Bias?
 - Types of AI Bias
- Sources of Bias
- Examples
 - Bias In Healthcare
 - Bias In NLP
 - Bias In Recommender Systems
- Debiasing
- Conclusion & Opinion
- Questions & Discussion

Introduction

What Is Bias

- **Bias definition:** The action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment[2].
- **Unconscious bias:** A type of bias that the person is not aware of but can influence the decisions of the person[2].

Underlying Bias Sources In AI

- The National Institute of Standards and Technology

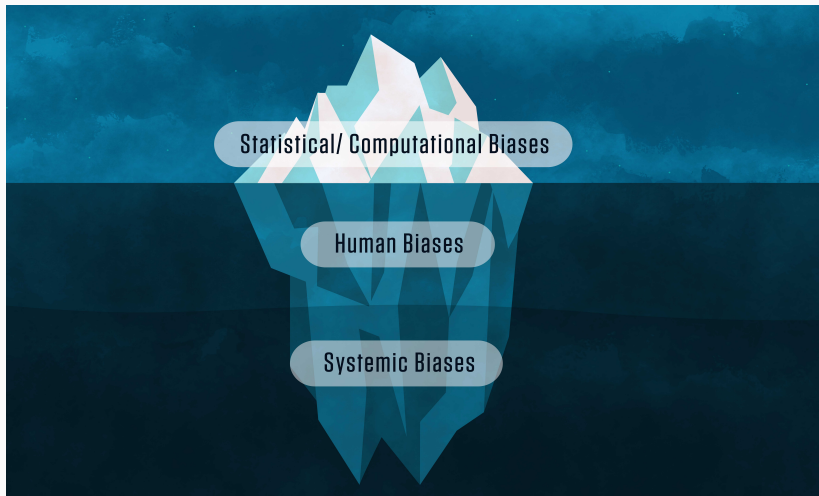


Figure 1: Bias sources according to NIST[12]

What Is AI Bias

- **AI bias** is an anomaly in the output of machine learning algorithms, due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data[6].
- Also known as **Algorithmic bias**, it is the tendency of algorithms to reflect human biases[4]

AI Bias Reflects Society's Biases

A Mckinsey study found that[11]:

- **Models may be trained** on data from human choices or data from social or historical disparities
- **Data may be biased** by the way they are gathered or chosen for use
- **User-generated** data may lead to a bias feedback loop
- **A machine learning** system may potentially detect statistical connections that are considered socially inappropriate or unlawful

Types of AI Bias

- **Algorithmic AI bias(Data bias):** Where algorithms are trained using biased data
- **Societal AI bias:** It is the type of bias that we have in our society, that the AI developers, or designers may carry into their own solutions[10].
- **Cognitive biases:** Unconscious errors in thinking that affects individuals' judgements and decisions[6].
- **Lack of complete data:** When the data that is used to train or develop the algorithm is incomplete.

Bias In AI/ML Pipeline

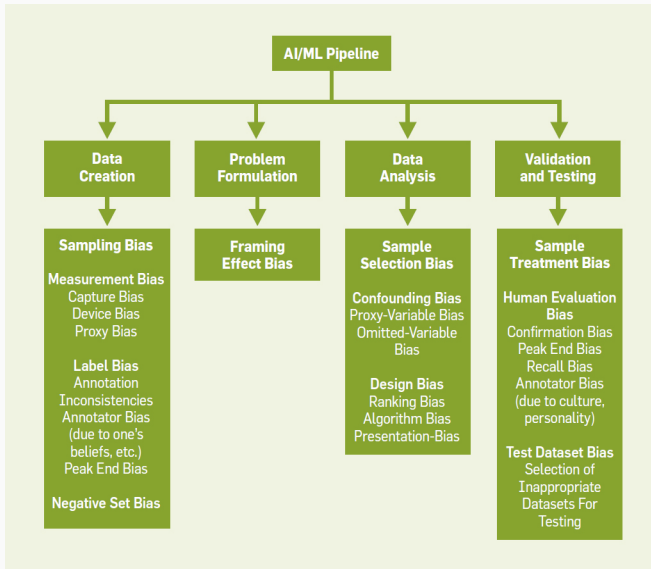


Figure 2: Bias in AI/ML Pipeline according to ACM[1]

Sources of Bias

Sources of Bias In AI Models

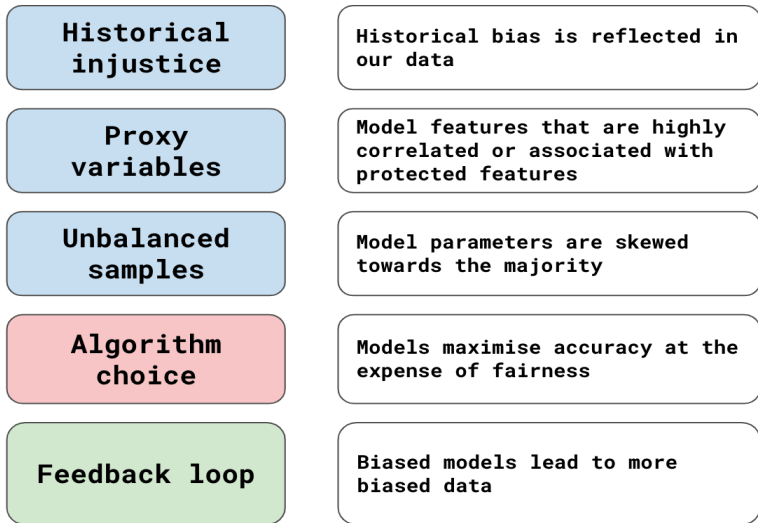


Figure 3: Common sources of bias[7]

Sources 1 - Historical injustice

- **Historically**, discrimination and bias existed, and still exist[7].
- This historical bias **can be reflected in our data**.
- A recent example is from a model developed by Amazon to help automate recruitment[7].

Sources 2 - Proxy variables

- **Proxy variables** are variables that are associated to **protected variables**, which are sensitive variables such as race or gender[7].

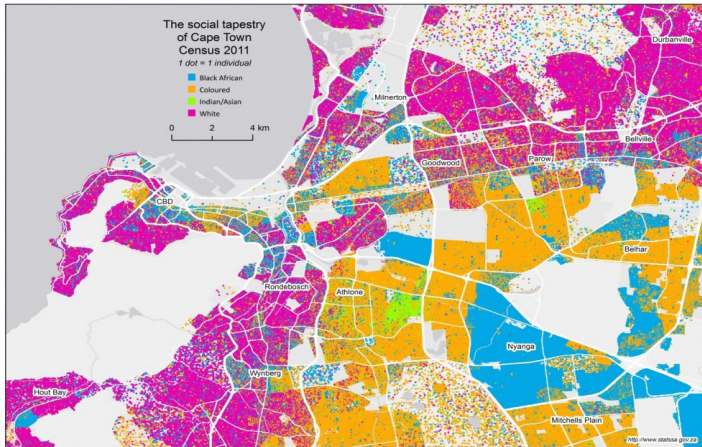


Figure 4: Racial groups in Cape Town[7]

Sources 3 - Unbalanced samples

- Datasets can be **imbalanced**, which results into **poor performance** or **biased results**.

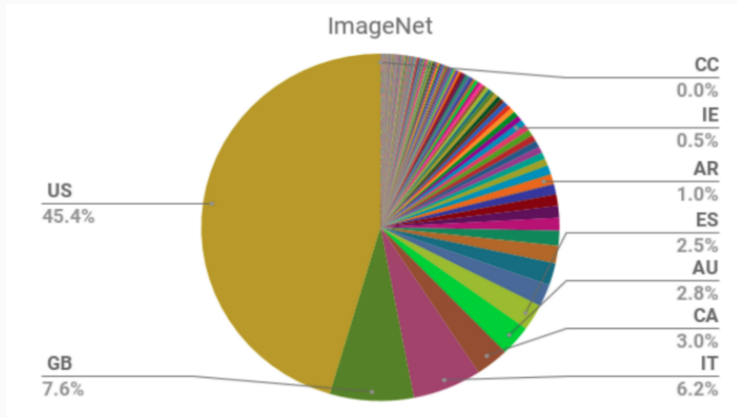


Figure 5: Geo-diversity of ImageNet dataset[7]

Sources 4 - Algorithm choice

- Some algorithms are **less interpretable** than others.
- This makes it **harder to identify** the source of bias and correct it.

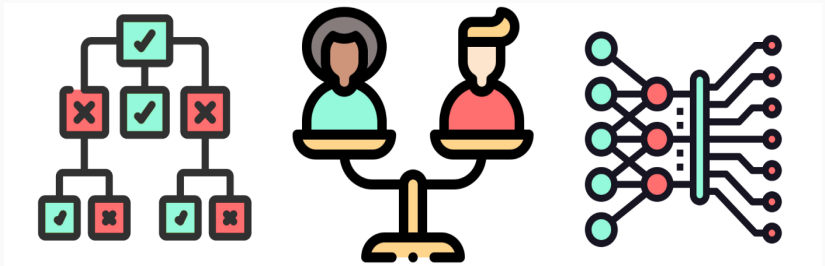


Figure 6: Algorithm bias illustration [7]

Sources 5 - User Feedback loop

- Biased models may become more biased.
- Since users that the model don't perform good on will not use the solution, the model will become even worse for them[7].

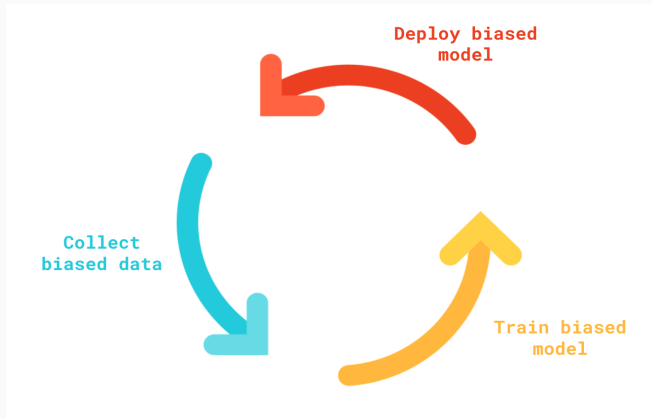


Figure 7: Algorithm feedback loop bias increment [7]

Examples.

Bias in Healthcare - 1

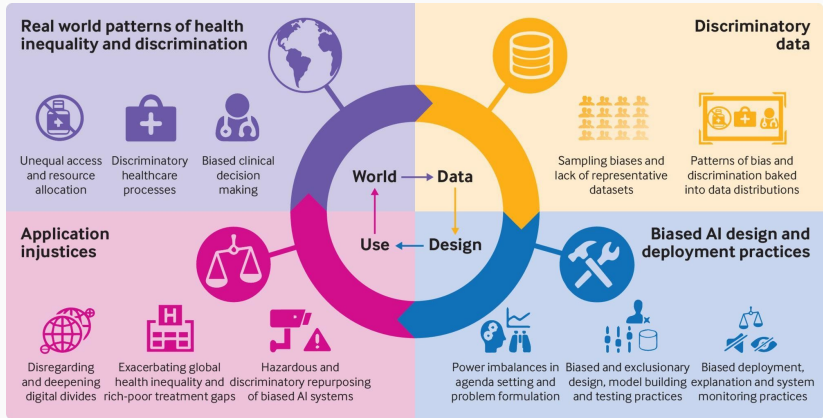


Figure 8: How bias occur in healthcare [11]

Racism in the American healthcare system[13]:

- In October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals.
- Goal was to predict which patients would likely need extra medical care.
- Researchers found that the system heavily favoured white patients over black patients.
- Even though race was not used, another variable, highly correlated was used.

AI bias may **have worsen COVID-19** health disparities for people of color[8].

- The **Journal of the American Medical Informatics Association** argued that biased models may further impacted people of color during the COVID-19 pandemic.
- Researchers noted, COVID-19 **prediction models** can present serious shortcomings, **especially regarding potential bias**.
- One of the most frequent problems were **unrepresentative data samples**.
- It is **worsened by existing disparities** in healthcare and **systemic racism**.

Bias in Recommender Systems - 1

This Recommender systems example is based on Chen et al. work[14]

- **Recommender systems** are a heavily researched area.
- Those systems demonstrate that they **have AI bias**.
- **Multiple types of Bias** exist in recommender systems.
- It is important to study the **impacts and the solutions**.

Bias in Recommender Systems - 2

Table 1: Bias in recommender systems

Types	Stages in Loop	Cause
Selection Bias	User >Data	Users' self-selection
Exposure Bias	User >Data	Item popularity, system intervention
Conformity Bias	User >Data	Conformity
Position Bias	User >Data	Exposed to top of lists
Inductive Bias	Data >Model	Added by researchers
Popularity Bias	Model >User	Algorithm, unbalanced data
Unfairness	Model >User	Algorithm, unbalanced data
Bias amplification in Loop	All	Feedback loop

Bias in Recommender Systems - 3

- Illustration of bias in recommender systems and where it occurs

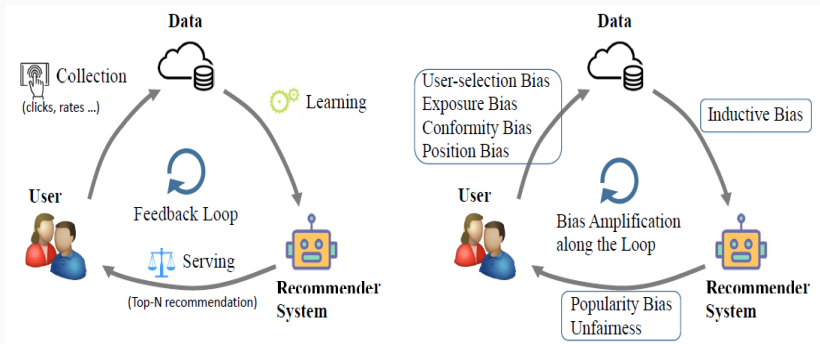


Figure 9: Bias in recommender systems [7]

This NLP example is based on Sun et al. work[15]

- In our society we have gender bias.
- This gender bias may be transferred into our NLP models.
- NLP models maybe propagate and even amplify gender bias.
- An example is how resume filtering systems may give preference to male applicants.

Table 2: Representation bias in NLP[15]

Task	Example of Representation Bias in the Context of Gender
Machine Translation	Translating "He is a nurse. She is a doctor" to Hungarian and back to English becomes "She is a nurse. He is a doctor"(Douglas,2017)
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).

Bias in NLP - 3

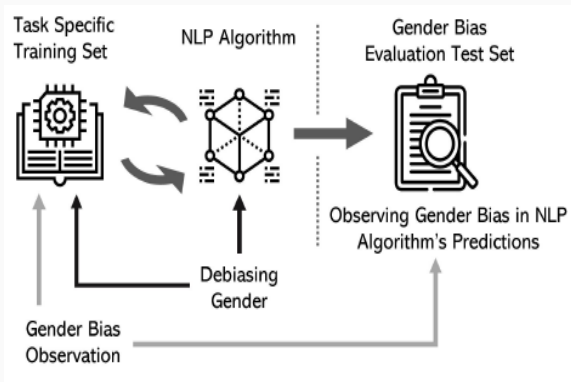


Figure 10: Algorithm feedback loop bias increment [7]

Debiasing

Minimizing Bias

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider



Figure 11: Minimizing Bias in AI according to Mckinsey[11]

Fixing Bias in AI Systems

There are general steps to fix bias in AI[11]:

- **Understand the algorithm and data**, to access where did the bias or unfairness occur; For example:
 - **Examine the training dataset**, is it a good representation? is it large enough?
 - **Perform subpopulation analysis** that would calculate the model metrics for the subgroups.
 - **Monitor the model over time**, since AI models can change over time, bias in them can also change over time.
- **Establish a debiasing strategy**
 - **Technical strategy** involving tools that can help you to identify bias sources.
 - **Operational strategies % Organizational strategy** such as how you collect your data, how it is verified through metrics.

Fixing Bias in AI Systems - 2

- **Improve human-driven processes** as you identify biases in training data.
- Decide on where to have an **automated decision Vs human involvement**.
- **Follow a multidisciplinary approach** by including experts that understand each application area.
- **Diversify your organisation**, having a diverse Ai community in your organization makes it easier to identify biases.

Tools to Reduce Bias

- **AI Fairness 360**[3]
 - Open source library from IBM.
 - Can test bias in models and datasets.
- **IBM Watson OpenScale**[9]
 - Another solution from IBM.
 - Performs bias checking in real time when the AI is making decisions.
- **Google's What-If Tool**[5]
 - A Tool offered by Google.
 - Can improve fairness and visualize model behavior.

Conclusion & Opinion

Conclusion.

- Bias is an existing phenomena in humans that we tend to carry out into AI systems.
- There are many types of Bias, both in humans and AI.
- Learning more about bias in ourselves to try to tackle bias in AI
- There are existing tools to tackle AI bias and its sources.
- Bias in AI is a very complicated issue that needs further studying and improvement in our solutions.









Thank you!

Questions?







Additional resources

References

References i

-  <https://cacm.acm.org/magazines/2021/8/254310-biases-in-ai-systems/fulltext>.
-  <https://dictionary.cambridge.org/dictionary/english/bias>.
-  <https://github.com/trusted-ai/aif360>.
-  <https://levity.ai/blog/ai-bias-how-to-avoid>.
-  <https://pair-code.github.io/what-if-tool/>.
-  <https://research.aimultiple.com/ai-bias/>.
-  <https://towardsdatascience.com/algorithm-fairness-sources-of-bias-7082e5b78a2c>.
-  <https://www.healthcareitnews.com/news/ai-bias-may-worsen-covid-19-health-disparities-people-color>.

References ii

-  <https://www.ibm.com/cloud/watson-studio>.
-  <https://www.lexalytics.com/>.
-  <https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans>.
-  <https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights>.
-  <https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/>.
-  J. CHEN, H. DONG, X. WANG, F. FENG, M. WANG, and X. HE.
Bias and debias in recommender system: A survey and future directions.
2021.



T. Sun, A. Gaut, S. Tang, Y. H. and Mai ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang.

**Mitigating gender bias in natural language processing:
Literature review.**

2019.