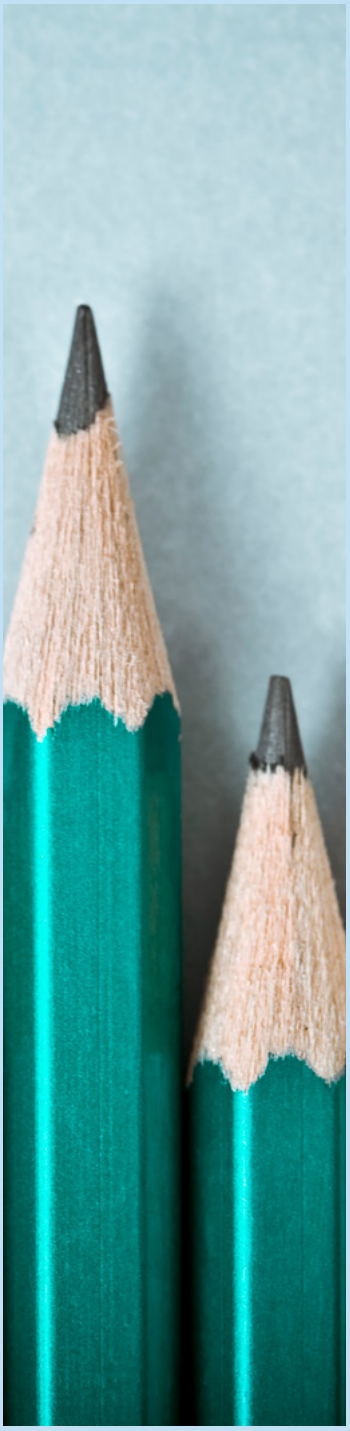
VIEW POINT



HUMAN BIAS IN AI

Abstract

Artificial Intelligence (AI) offers enormous potential to transform our businesses, solve and automate some of our toughest problems and inspire the world to a better future. However, AI systems are created and trained using human generated data that could affect the quality of the systems. Bad data can contain implicit racial, gender, or ideological biases. AI models learn those biases and even amplify them. Building an Inclusive AI model devoid of biases and discrimination is the need of the hour.

Infosys®
Navigate your next

What is Human Bias?

As per Oxford Learner's Dictionary,

"Bias is a strong feeling in favour of or against one group of people, or one side in an argument, often not based on fair judgement."

To be biased simply means to not have a neutral viewpoint. Biases can be toward or against an individual, an ethnic group, gender identity, a political party, a religion or a lot other things. People can learn biases from history, other individuals and cultural contexts. For example, many people are quick to assume that the word 'Nurse' would probably mean a female nurse. This bias could be because the role has been associated with women since the Crimean War following which the ratio of female to male nurses has been skewed.

How does AI learn Bias?

The bias in AI systems is the byproduct of cognitive biases in humans as it is trained using biased human generated data and decision making processes. Here is a look at the factors influencing AI bias.

1. Biased Data

Let us take the example of Google translate. Nurse and Engineer are gender-neutral words. If we translate sentences containing those words to gender-neutral languages like Bangla or Turkish and translate them, it will associate engineer with masculine pronouns and nurse with female pronouns. A translation attempt with different words like doctor and teacher will also yield similar errors.

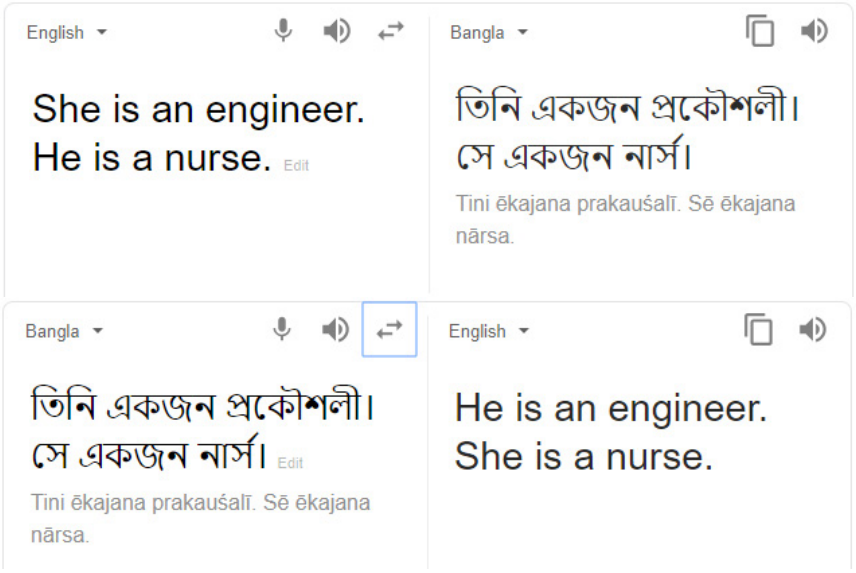


Figure 1: Gender Bias in Google translation AI model

How did Google Translate learn biases?

Word embedding is one of the core engines of most modern NLP systems. Google Translate and many other popular AI linguistic models learn by guessing the next word or context words in texts available on the web and textbooks; these sources are abound with biases of the society and the biased views of the authors. Word embedding learns analogies such as **"King is to Man as Queen is to Woman"** and learns biased analogies like **"Computer programmer is to Man as Homemaker is to Woman"** and **"Doctor is to Man as Nurse is to Woman"**. Also, in majority of the past text data, the word 'doctor' appears more frequently with male names creating a bias.

According to 'The 2009 Statistical Abstract Report by US Census Bureau', full time workers with the same educational qualifications have huge gender based pay gap. Hence, creating a pay model using data from a company with gender pay gap will most probably yield a biased AI model. Recently, Amazon faced backlashes for similar kind of models.

There are more than 180 human biases that have been defined and classified and many of them are evident in the AI systems of today. For example, Google Photos have been in the news for misclassifying people of a certain ethnicity as gorillas. These kind of misclassifications happen because our data does not represent true distribution of different subgroups. A paper titled 'No Classification without Representation' showed over 2/3 of the images in ImageNet, the most studied image data set in the world, to be from the Western world (USA, England, Spain, Italy, Australia).

2. Filter Bubbles and Conflicting Goals

Filter bubbles are information bubbles created by AI powered Personalization engines based on user likes, dislikes and content browsing history. These bubbles create a virtual space where users are only served with the content that confirm and reinforce their viewpoints. Many social media channels today are creating such filter bubbles promoting extremism and conspiracy theories.

With the acceptance of Social Media and User Generated Content and AI powered Recommendation Engines to promote it all, people are sinking deeper into filter bubbles. All thanks to the AI powered recommendation engine, worldwide, people are spending more than 1 billion hours on YouTube per day. One in five of these YouTube users watch recommended videos regularly, despite the videos being conspiracy theories or fake news, leading the algorithm to learn those patterns to recommend more such videos and hence amplify the biases. Even when platforms like YouTube and Facebook have sanitation measures in place to identify and flag fake news, the opaque recommendation algorithms prove to be a blatant conflict to this goal of sanitation and leads to filter bubbles.

How AI Amplifies Bias?

Most commonly used algorithm models currently in production are discriminative models, such as neural networks, logistic regression and random forests. Discriminative models maximize accuracy by generalizing on data they have been trained on, thereby amplifying the bias in training data. This observation was first highlighted in a paper called "Men Also like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints". As part of this research, a training dataset was created with images of people cooking, in which 66% images were of females. However, the trained model amplified that bias to predict that 84% of the people cooking were female.

Bias Aware AI Pipeline

Problem Statement

- What kind of applications will your AI System have?
- What will be the impact of biased predictions?
- What kind of bias might be present in your AI system?
- What is your targeted user population?

Inclusive Team

Just as textbooks reflect the biases and the views of their authors, systems will include the biases of its developer. A diverse and inclusive team will keep these biases in check.

Data

- Is your past data relevant for the current problem?
- Is your data creation process inclusive of all subgroups?
- What kind of data are you collecting?
- Did you validate your data for biases?
- How are you handling biased features? De-biasing? Removing?
- Are you reviewing your data before sending to training update?

Let us understand with a few examples

Microsoft's Twitterbot Tay, in a few hours of its launch, interacting with Twitter users, learnt racial abuses forcing it to be shut down in 24 hours. This was because of the absence of human review of streaming data before training.

A study by ProPublica showed that Criminal risk profile assessment tool COMPAS survey questions had huge correlation with ethnic subgroups and repeated offense. The tool was also identified to be biased against African Americans.

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, did Re-Offend	47.7%	28.0%

Model Optimization and Bias Metrics

- Optimization metrics have huge impact and should be true representative of all subgroups
- Conflicting goals or optimization can create filter bubbles
- Define a bias metric to measure biases present in the model

Model Selection

- Calculate metrics for true subgroup representations
- Select best model with low bias and high accuracy

Here is an example of a model selection that could avoid gender bias

Subgroup	Subgroup Probability in test data	Model Accuracy		
		Model A	**Model B**	Model C
Male	0.75	90%	**87%**	75%
Female	0.25	80%	**82%**	74%
Overall	1	87.5%	**85.75%**	74.75%
Expected Accuracy (0.5 Male and 0.5 Female)		85%	**84.5%**	74.5%
Bias (Ratio of Male and Female Accuracy)		1.125	**1.06**	1.01

Human Review and Bias Awareness

- Share details about biases present in model with user
- Monitor Bias metrics for AI system in production
- Human review for predictions which have high impact. Example, sentence prediction for criminals etc.
- Report button for biased/wrong predictions

What Enterprises are Doing to Minimize AI Biases?

Most of the research in this area has been via academics and nonprofit organizations. Big tech companies are the primary driving force for AI advances, and their algorithms impact billions of people. Unfortunately, it is only now that most companies have started focusing on reducing biases. Google has launched the What-If tool for detecting biases and similarly IBM has AI Fairness 360 toolkit. Many enterprises such as Google, Facebook, Microsoft and IBM have launched various fairness programs. Kaggle has launched many competitions and shared datasets, which can be used for advancing research in this area. Famous AI courses such as Fairness in Machine learning by UC Berkeley, ML Engineering Course by Google, Deeplearning.ai by Andrew Ng, Fast.ai are also trying to create awareness on biases in lectures. However, there is a need to have more transparency and regulation to ensure bias free AI systems.

Conclusion

Building a responsible AI Model with fairness, privacy & security is the need of the hour. AI Model has to be sensitive towards factors such as race, gender, religious beliefs, income & caste and has to be built with humility. AI systems will remain biased unless we start focusing on inclusion at every step. This underlines the need for constant testing, accountability and review process from system ideation to development. People are not perfect, but they learn from their mistakes and rectify them. Likewise, AI systems must become accountable and rectify and minimise thier biases through constant monitoring, awareness and continuous feedback.

References

1. Thomas, Rachel. (2019, January 31). I'm an AI researcher, and here's what scares me about AI. Retrieved from https://medium.com/@rachelthomas/im-an-ai-researcher-and-here-is-what-scares-me-about-ai-909a406e4a71
2. Hammond, Kristian. 5-Unexpected sources of bias in artificial intelligence. Retrieved from https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/
3. Sajin, Stas. (2018, November 1). Preventing machine learning bias. Retrieved from https://towardsdatascience.com/preventing-machine-learning-bias-d01adfe9f1fa
4. Douglas, Laura. (2017, December 16). AI is not just learning our biases; it is amplifying them. Retrieved from https://medium.com/@laurahelendouglas/ai-is-not-just-learning-our-biases-it-is-amplifying-them-4d0dee75931d
5. 2016, May 23. Machine Bias. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
6. Thomas, Rachel. (2017, June 21). Word Embeddings, Bias in ML, Why You Don't Like Math, & Why AI Needs You. Retrieved from https://www.youtube.com/watch?v=25nC0n9ERq4&index=4&list=PLtmWHNX-gukJUPxzuBf_GnTfN67WJen6c
7. Bolukbasi, Chang, Zou,Saligrama & Kalai. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Retrieved from https://arxiv.org/pdf/1607.06520.pdf

About the Author

Dharmendra Choudhary is a Specialist Programmer working with the Conversation UI research team at Infosys Center for Emerging Technology Solutions. He specializes in Deep Learning, NLP and AI models in production. He graduated from IIT Delhi and is a self-learner who loves to engage in technology conversations.

To know more about our work in AI write to us at icets@infosys.com.