## Gemensam vinjett TGTU82 och TATA24

- Fokus bör läggas på att undersöka sambandet mellan linjär algebra, språkmodeller och bias t ex hur metoder från linjär algebra kan användas för att upptäcka bias i språkmodell + diskutera resultaten.
- Tyngdpunkt kan läggas på antingen "bias" eller på språkmodeller och deras relation till linjär algebra

## Model bias and linear algebra

Earlier news headlines on bias in algorithms (source [1]):

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of racially biased algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in Al: building fairer algorithms

Bias in Al: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with Al voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

The Best Algorithms Struggle to Recognize Black Faces Equally

Artificial Intelligence has a gender bias problem – just ask Siri

Today one of the most popular kind of machine learning model is the Large Language Model (LLM), such as ChatGPT. What about bias in LLMs? Does it exist? What effects can it have? How can it be uncovered, studied and maybe even addressed?

What do you need to know about LLMs to understand and study the potential biases? (For an introduction to Natural Language Processing, see [2], and other sources on the web.) How do LLMs relate to things you are learning in linear algebra – how do they represent words and their meaning, e.g. using embeddings? In your reading for the TGTU82 seminar you find [3]. How do word embeddings work? Can your knowledge about vectors be used to study LLMs then? How is bias represented in the vector space?



Where does bias in LLMs come from? What are you correcting when you try to counter bias? And what remains uncorrected? Can you think of different ways to counter unwanted biases in LLMs? How would you address the problem?

If you want to see how things work in practice, try out a simple language model using the Jupyter notebook at [4] and try to find some biases, or try some of the other examples there. How is a word represented? What does "similar" mean, and how can you see if something is 'close' in this model? What happens if we do a calculation like "king" – "man" + "woman" or "king" – "he" + "she"? What happens in the vector space? What results do you get? What about "doctor" – "he" + "she"? Are the results expected? Unexpected? What about other gender-identities? Can you generate results that reflect a non-binary gender perspective? Why or why not?

## References:

- [1] https://towardsdatascience.com/algorithm-bias-in-artificial-intelligence-needs-to-be-discussed-and-addressed-8d369d675a70
- [2] https://foundations-of-ai-and-ml.ida.liu.se/content/nlp/intro
- [3] Gonen, H. & Y. Goldberg (2019) Lipstick on a Pig: Debiasing Methods Cover up Systemic Gender Biases in Word Embeddings But do not Remove Them. arXiv:1903.03862v2 (på Lisam)
- [4] https://gitlab.liu.se/jansn19/jts-tata24/-/blob/main/Kod/wordembeddings-new.ipynb