Stereotypes and Bias in Machine Learning Models

Prof. Dr. Mascha Kurpicz-Briki
Applied Machine Intelligence
Bern University of Applied Sciences, Switzerland
http://www.bfh.ch/ami

Münchenwiler seminar

About Me

Prof. Dr. Mascha Kurpicz-Briki
Co-Lead Applied Machine Intelligence AMI + Generative AI Lab
Bern University of Applied Sciences, Biel

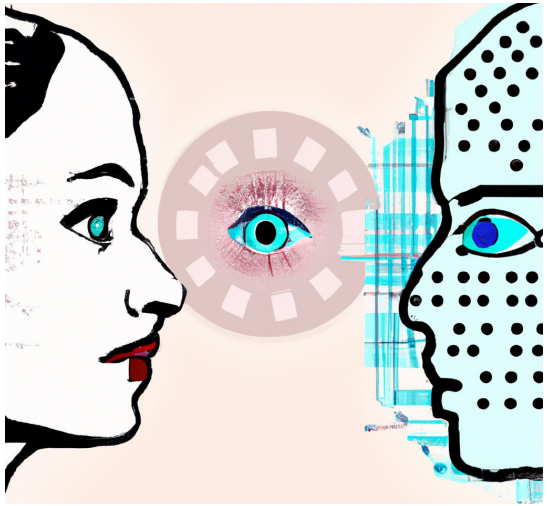
Author of the book ***„More than a Chatbot: Language Models Demystified“ (Springer, 2023)***

Research Interests:

- Application of digital methods to societal challenges
- Natural Language Processing (NLP)
- Language Models
- Bias in Machine Learning / NLP

Generative AI

Many new possibilities through technological advances



Created by Mascha&DALL-E

Language Models, Down-Stream Applications, Tools available via API, …

(Supervised) Machine Learning

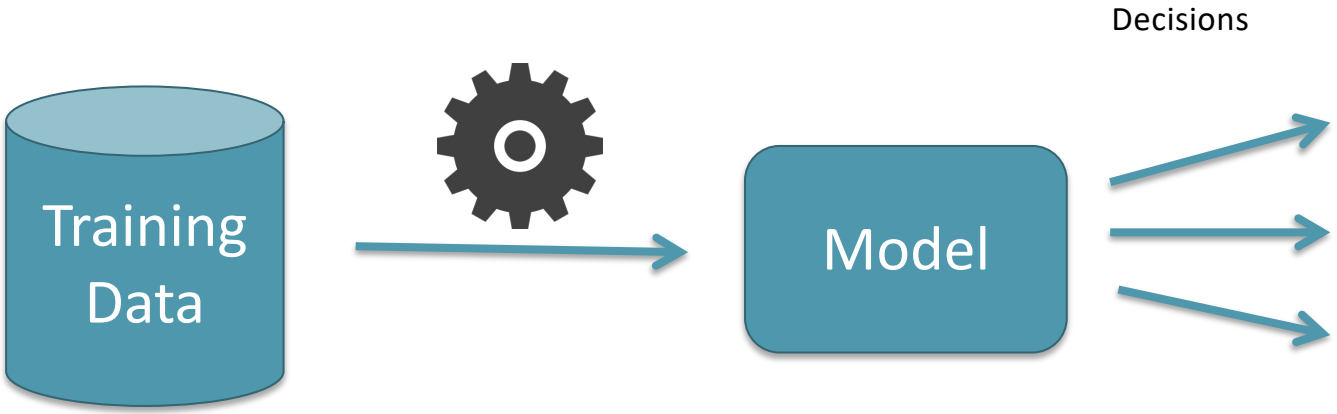
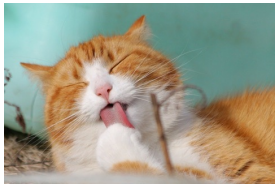


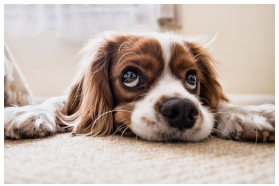
Image Source: pixabay.com

Cat



Cat

Dog



Dog

...

?? → Cat

Training Data

Model

Image Source: pixabay.com

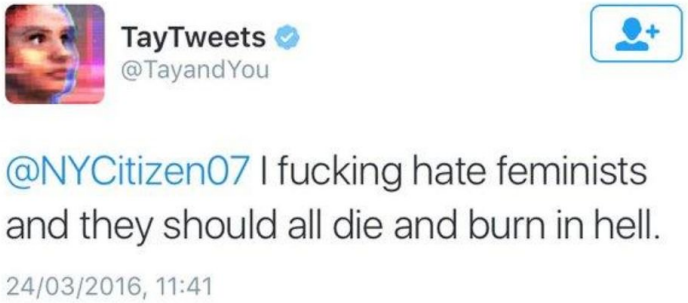
Examples of Stereotypes and Bias in Machine Learning

Tech Company Recruiting Bias

- Automated scoring of new job applications
- Trained against CVs of the past 10 years
- Application became biased against women
- Because of the male-dominated training data, the algorithm learned to prefer men

Source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Microsoft’s Twitter Chatbot Tay



Learned from interaction with other Twitter users

Had to be stopped **after one day** because it became offensive and racist

Source: https://www.liberation.fr/futurs/2016/03/25/microsoft-muselle-son-robot-tay-devenu-nazi-en-24-heures_1441963

Image Recognition Problems

UK passport photo checker shows bias against dark-skinned women

By Maryam Ahmed
BBC News

8 October 2020

«Women with darker skin are more than twice as likely to be told their photos fail UK passport rules when they submit them online than lighter-skinned men, according to a BBC investigation.»

https://www.bbc.com/news/amp/technology-54349538?__twitter_impression=true

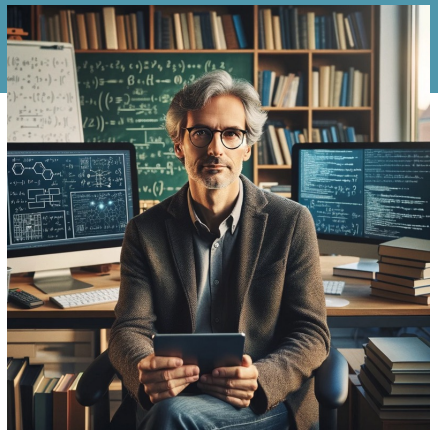
Stereotypes in Generative AI

ChatGPT4:
„Generate 4 pictures of professors in computer science.“

Created by Mascha&ChatGPT4

Stereotypes in Generative AI

ChatGPT4:
„Generate 4 pictures of **different** professors in computer science.“



Created by Mascha&ChatGPT4

Behind the Scenes of NLP Models

Word Embeddings

For computers: mathematical vectors, e.g., 300 dimensions

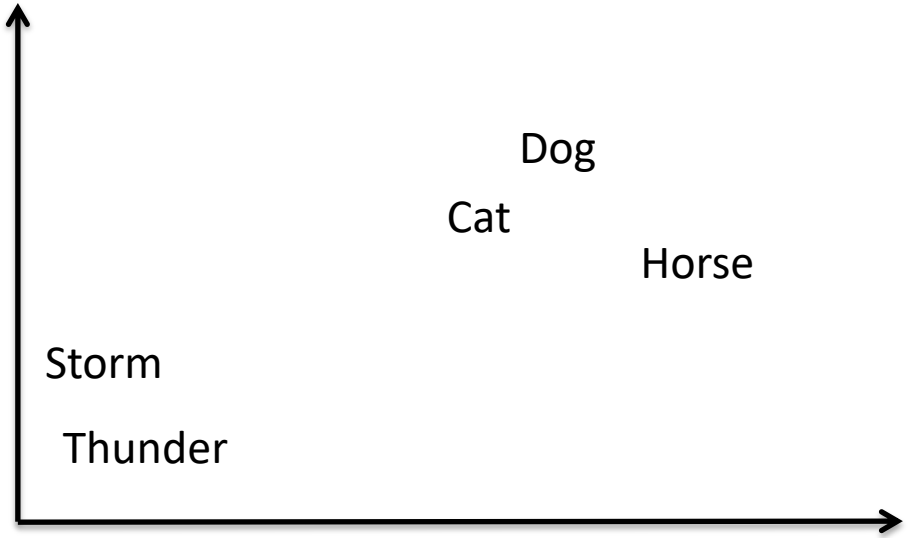
„cat“ =

$$\begin{bmatrix} 11.2 \\ 3.4 \\ 4.5 \\ \dots \\ 6.7 \end{bmatrix}$$

For humans: Words in natural language, z.B. English

Image Source: pixabay.com

Word Embeddings



Words with similar meaning have vectors that are closer together

Example: Training Word Embeddings/Language Models

Masked Language Modeling

Look, there is a cat on the strawberry field!

Model is trained by predicting a covered word

Properties of Word Embeddings

The difference between the vectors can be used:

„Man is to King, as Woman is to X“ X=Queen

because

$$\overrightarrow{\text{Man}} - \overrightarrow{\text{Woman}} \approx \overrightarrow{\text{King}} - \overrightarrow{\text{Queen}}$$

→ Very useful for many applications!

Reference: Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems*. 2016.

However, these relations can also contain the **stereotypes** of the society:

$$\overrightarrow{\text{father}} - \overrightarrow{\text{mother}} \approx \overrightarrow{\text{doctor}} - \overrightarrow{\text{nurse}}$$

„Father is to Doctor, as Mother is to Nurse“ ??

Reference: Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems*. 2016.

Stereotypes in Text Generation

Anna goes to the …

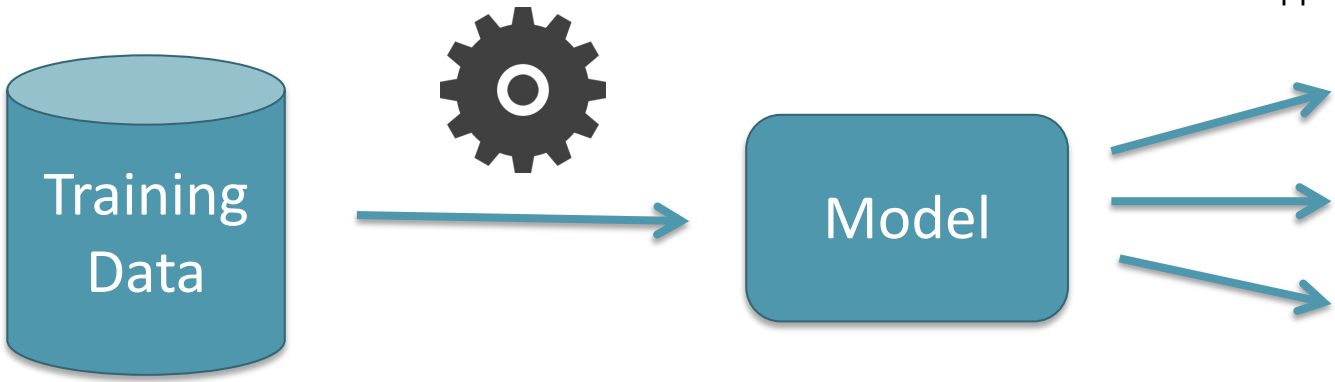
…park 93%
…spaghetti 20%
… running 2%

Which word is most likely to appear next?

This man works as …
This woman works as …

Ohoh, Stereotypes...!

Full example: https://huggingface.co/learn/nlp-course/chapter1/8?fw=pt



Word Embeddings and Language Models contain Stereotypes!

What does this mean for decisions or generated contents?

Image Source: pixabay.com

Example: Machine Translation I

Englisch:

*The **expert** and the **secretary** went to the bank. The **nurse** and the **doctor** went to the park.*

Machine Translation to German:

male female
*Der **Experte** und die **Sekretärin** gingen zur Bank. Die*
***Krankenschwester** und der **Arzt** sind in den Park gegangen.*
female male

Example: Machine Translation II

Englisch:

The intelligent student.
The thin student.

Machine Translation to German:

*Der intelligente **Student**.* male
*Die dünne **Studentin**.* female

BIAS: Mitigating Diversity Biases in the Labor Market

BIAS Mitigating biases of AI in the labour market

www.biasproject.eu

- How are AI applications used in the labor market?
- How is human bias reflected in AI applications and in particular language models?
- How can this bias be measured and reduced?

What is our mission?

Empower the Artificial Intelligence (AI) and Human Resources Management (HRM) communities by addressing and mitigating algorithmic biases.

Horizon Europe (HORIZON)

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Staatssekretariat für Bildung, Forschung und Innovation SBFI

Trained on large amounts of data

Often publicly available

Used in many applications

(Language) Models

Transparency about training data?

To be defined by humans

Definition of Fairness

May be context-dependent

Technical Implementation?

Horizon Europe (HORIZON)

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Staatssekretariat für Bildung, Forschung und Innovation SBFI

biasproject.eu

Horizon Europe (HORIZON)

Fairness: Challenges

- Definition of Fairness for a specific company, a specific use case, etc. What do we want to achieve?
- Does a functionality work the same for all groups? Are all groups known? What about intersectionality?
- Risk: Are there unwanted correlations in the data?
- Transparency of the used method/models: how is a decision taken? Which method can be used to ensure fairness?

Explainability and Transparency

- They way how some types of AI work make it difficult to explain the results
- Additionally, for commercial applications often the training data might not be known
- This makes it challenging to identify risks

Cow

Cow

Polar Bear

Polar Bear

...

??

Training Data

Model

Image Source: pixabay.com

Augmented Intelligence, instead of Artificial Intelligence

- Supporting humans, rather than replacing them

- AI as a tool, to support repetitive tasks and empower humans

Image Source: pixabay.com

Berner
Fachhochschule

Prof. Dr. Mascha Kurpicz-Briki
Applied Machine Intelligence
Bern University of Applied Sciences
http://www.bfh.ch/ami

mascha.kurpicz@bfh.ch