BIAS

Mitigating biases of AI in the labour market

Novel Tools for Bias Detection and Mitigation in AI Systems

The Debiaser

Mascha Kurpicz-Briki, October 2023

Problem: Bias in AI Applications

«UK passport photo checker shows bias against dark-skinned women»

https://www.bbc.com/news/amp/technology-54349538

«Amazon scraps secret AI recruiting tool that showed bias against women»

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

«Tay: Microsoft issues apology over racist chatbot fiasco»

https://www.bbc.com/news/technology-35902104

and many more…

Funded by the European Union

BIAS

Challenges of Bias Detection and Mitigation in AI

- Many different types of data (text, structured data, images…)
- Definition of Fairness
- Different types of bias, intersectional bias
- Bias can be introduced at different stages of the process (data, model, type of applications, …)

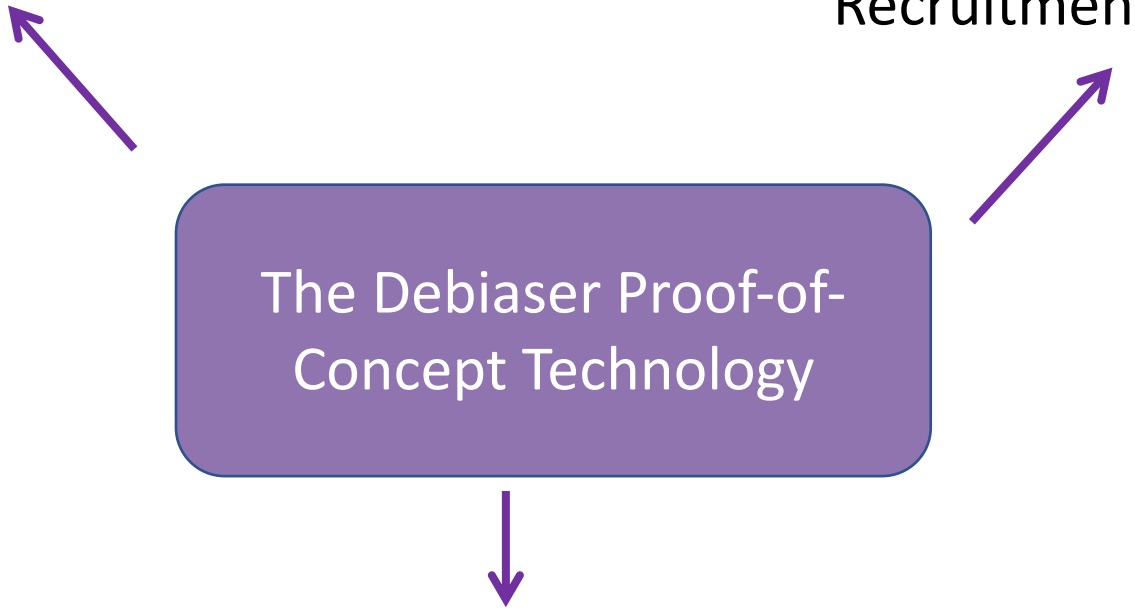
Funded by
the European Union

BIAS

The Debiaser: Different Components

Fair Recruiting with Case-based Reasoning

Debiasing and Explaining Text-Based Recruitment Applications



Bias Detection and Mitigation in Word Emdeddings and Language Models

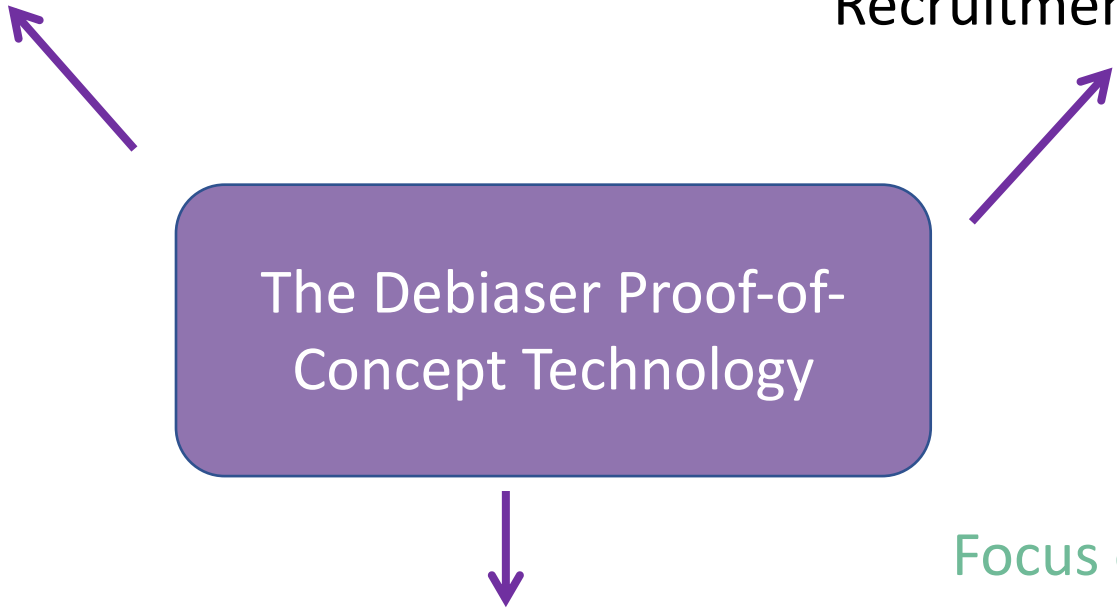
Funded by the European Union

BIAS

The Debiaser: Different Components

Fair Recruiting with Case-based Reasoning

Debiasing and Explaining Text-Based Recruitment Applications



Focus of today's presentation

Bias Detection and Mitigation in Word Emdeddings and Language Models

Funded by the European Union

BIAS

Word Embeddings

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

For humans: Words in natural language, e.g., English

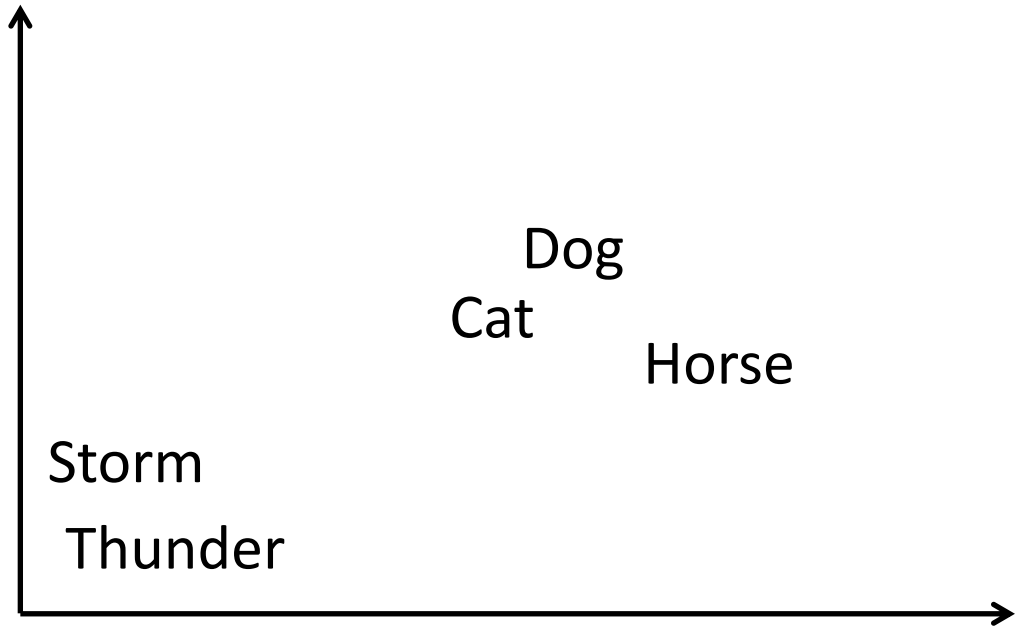
„cat“ = $\begin{bmatrix} 11.2 \\ 3.4 \\ 4.5 \\ \dots \\ 6.7 \end{bmatrix}$

Funded by the European Union

BIAS

Image Source: pixabay.com

Word Embeddings



Words with similar meaning have vectors that are closer together

Funded by the European Union

BIAS

Word Embeddings

„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!

Funded by the European Union

BIAS

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.

Societal Stereotypes in the Word Embeddings

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

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

„Man is to Computer Programmer, as Woman is to Homemaker“ ??

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

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

Funded by the European Union

BIAS

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.

Example: Bias in NLP Applications

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

Funded by the European Union

BIAS

Bias in Language Models for Text Generation

Anna goes to the …

…park 93%
…spaghetti 20%
… runs 2%

Which word is most likely to appear next?

This man works as …
This woman works as …

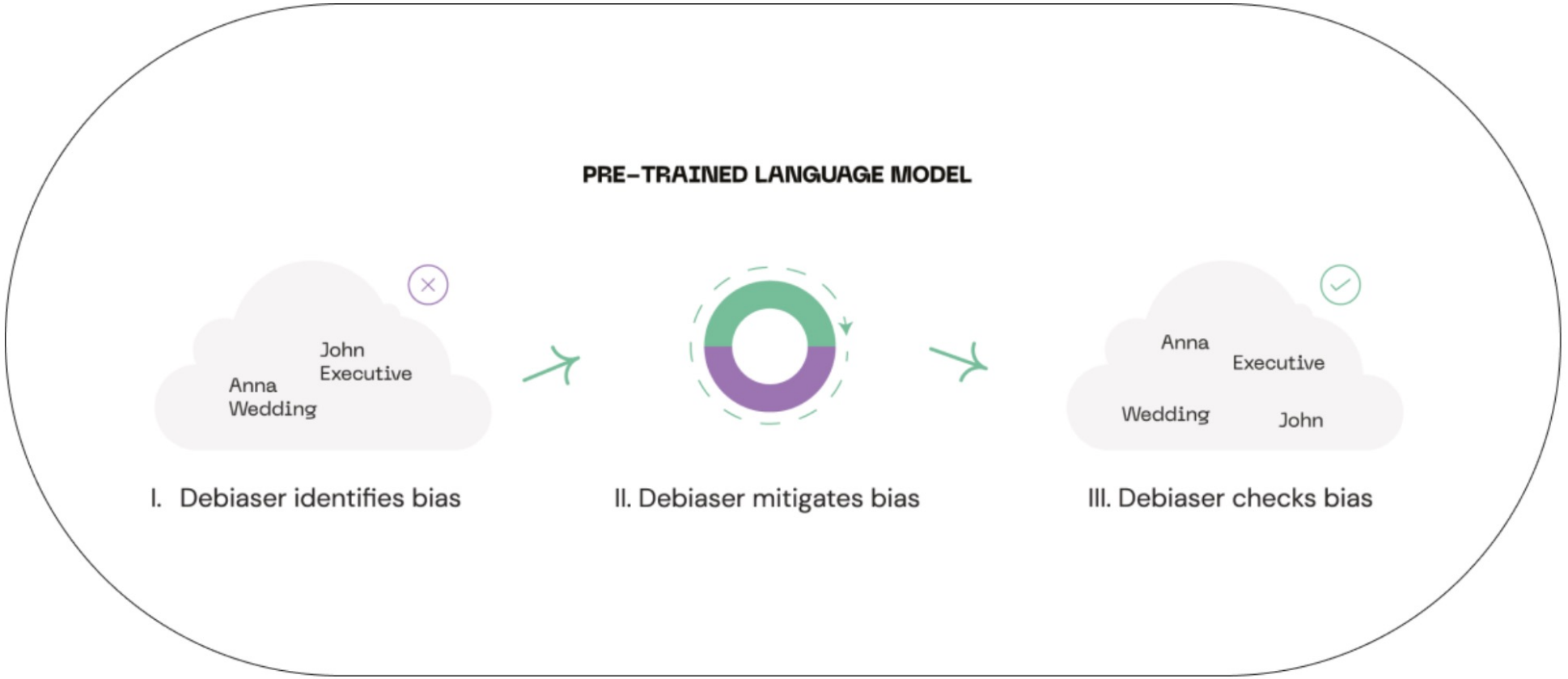
Ohoh, Stereotypes...!

Funded by the European Union

BIAS

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

The Debiaser: Bias Detection and Mitigation in Word Embeddings and Languages Models



Funded by
the European Union

BIAS

The Debiaser: Contribution to Bias Detection/Mitigation for European Languages

- Existing work in the state-of-the-art has a strong focus on English
- Our previous work gives indication that there are cultural differences how bias is encoded in word embeddings or language models
- Due to language properties, adapted methods might be required for different languages
- Studying how bias reflects in techology requires a close involvement of interdisciplinary stakeholders

Funded by the European Union

BIAS

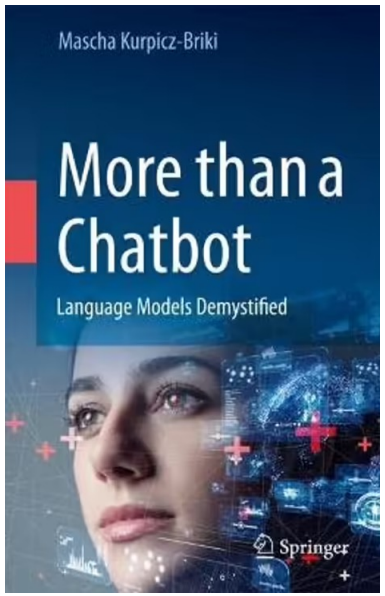
Kurpicz-Briki, M., & Leoni, T. (2021). A World Full of Stereotypes? Further Investigation on Origin and Gender Bias in Multi-Lingual Word Embeddings. *Frontiers in big Data*, *4*, 625290.

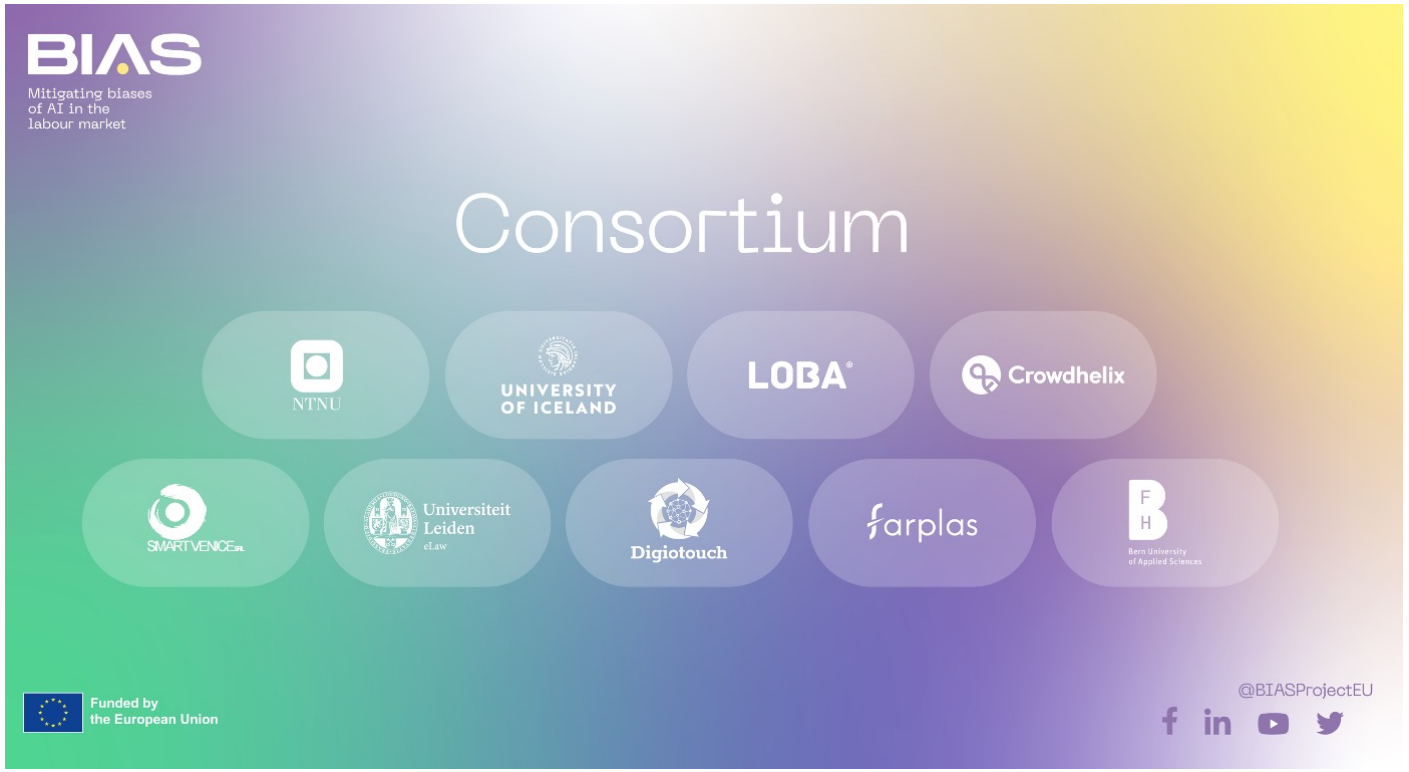
Thank you for your attention!

Glad to discuss further:

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

mascha.kurpicz@bfh.ch





Funded by the European Union

BIAS