Berner
Fachhochschule

Potenziale und Diskriminierungsrisiken in KI-Anwendungen

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

Viele neue Möglichkeiten und Potentiale von KI-Anwendungen im Bereich Text

Chatbots

Maschinelle Übersetzung

…

Extraktion von Informationen aus Texten

Texte generieren

Traditionelle Software

Zutaten + spezifische Anleitung = Resultat

Artificial Intelligence

„machine learning“

Image Source: pixabay.com

(Supervised) Machine Learning

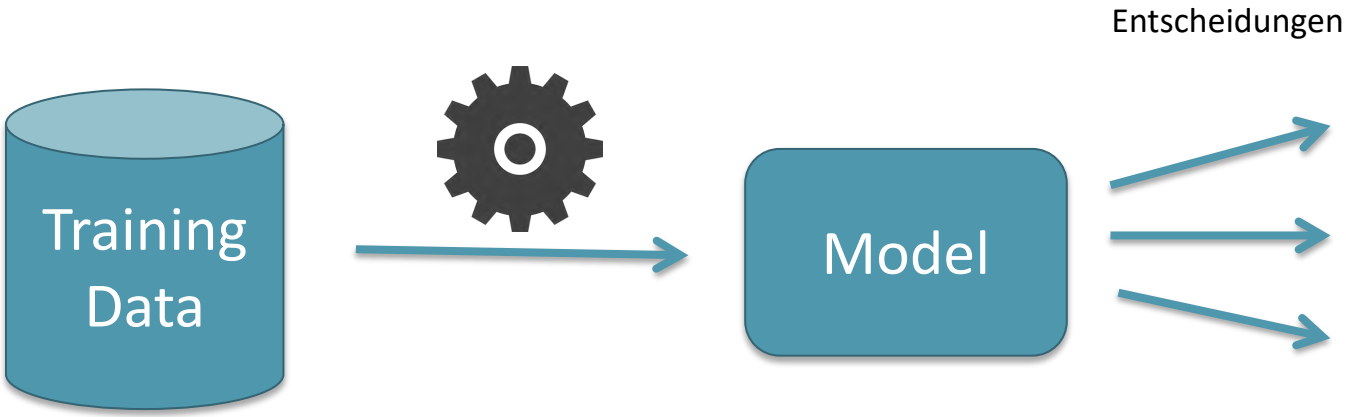
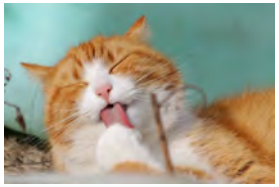


Image Source: pixabay.com

Katze



Katze

Hund

Hund

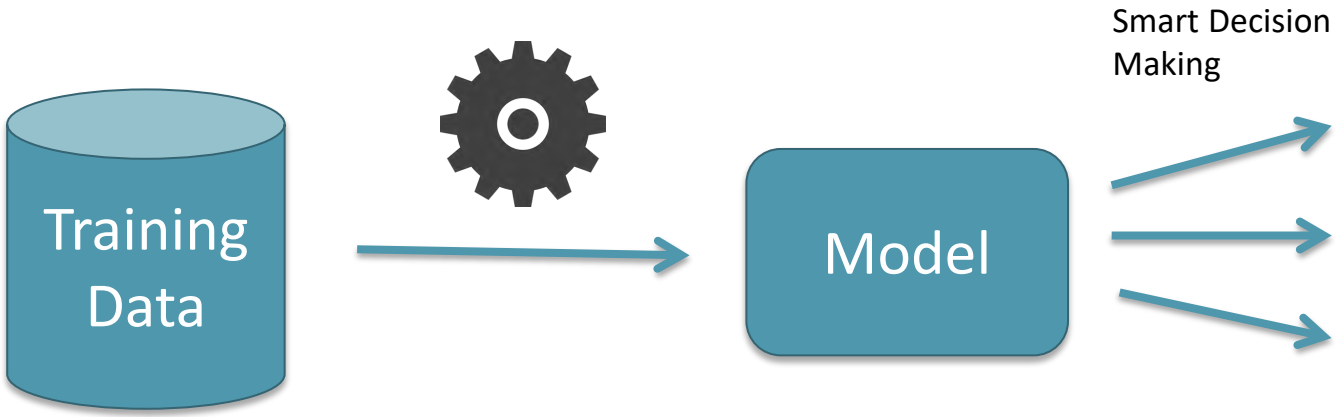
...

?? → Katze

Model

Training Data

Image Source: pixabay.com



Enthalten die Daten *Stereotypen*?

Sind die Entscheidungen *fair*?

Image Source: pixabay.com

Word Embeddings

Für automatische Verarbeitung:
Mathematischer Vektor, z.B. 300 Dimensionen

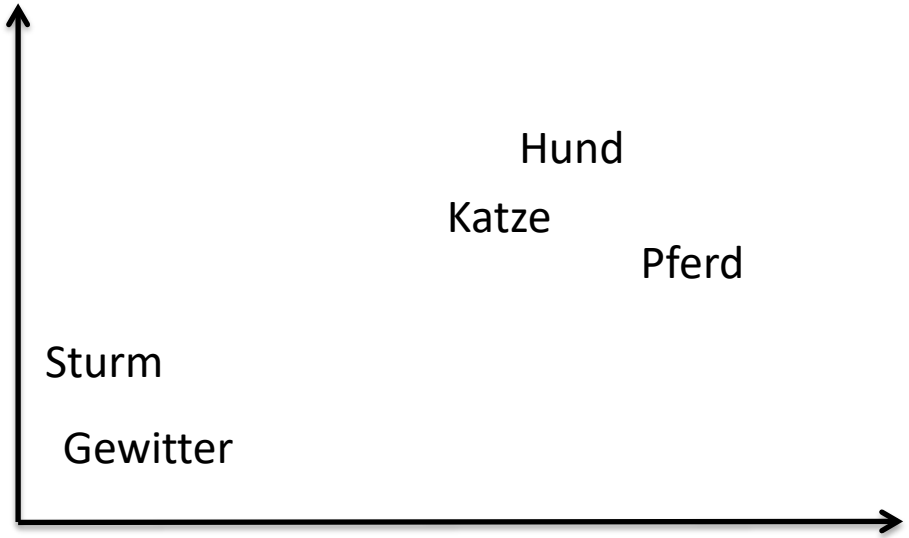
„Katze“ =

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

Für Menschen: Wort in natürlicher Sprache, z.B. Deutsch

Image Source: pixabay.com

Word Embeddings



Wörter mit ähnlicher Bedeutung haben Vektoren, die näher beieinander sind

Image Source: pixabay.com

Eigenschaften von Word Embeddings

Diese Differenz zwischen den Vektoren kann genutzt werden:

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

weil

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

→ Sehr nützlich für vielerlei Anwendungen!

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.

Solche Beziehungen sind nützlich für viele Anwendungen, aber können auch **Stereotypen** enthalten:

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

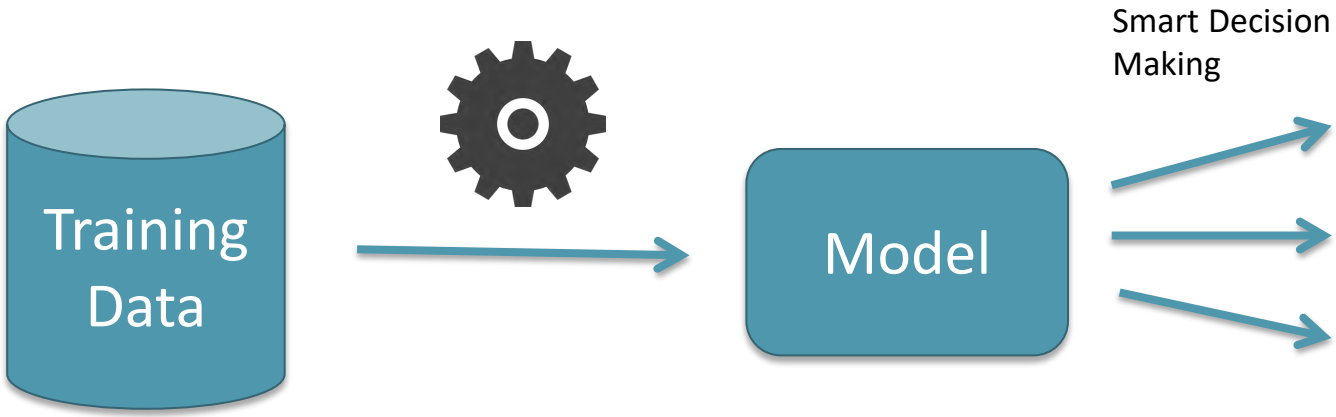
Reference: Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." *Advances in neural information processing systems*. 2016.

Solche Beziehungen sind nützlich für viele Anwendungen, aber können auch **Stereotypen** enthalten:

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

„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.



Die Sprachmodelle enthalten Stereotypen!

Was bedeutet das für die Entscheide?

Image Source: pixabay.com

BIAS: Mitigating Diversity Biases in the Labor Market

- Wie werden KI Anwendungen auf dem Arbeitsmarkt eingesetzt?
- Wie wird menschlicher Bias in der KI reflektiert?
- Wie kann solcher Bias gemessen und reduziert werden?

Jetzt aktiv mitwirken! biasproject.eu

Horizon Europe (HORIZON)

Large Language Models: „How big is too big?“

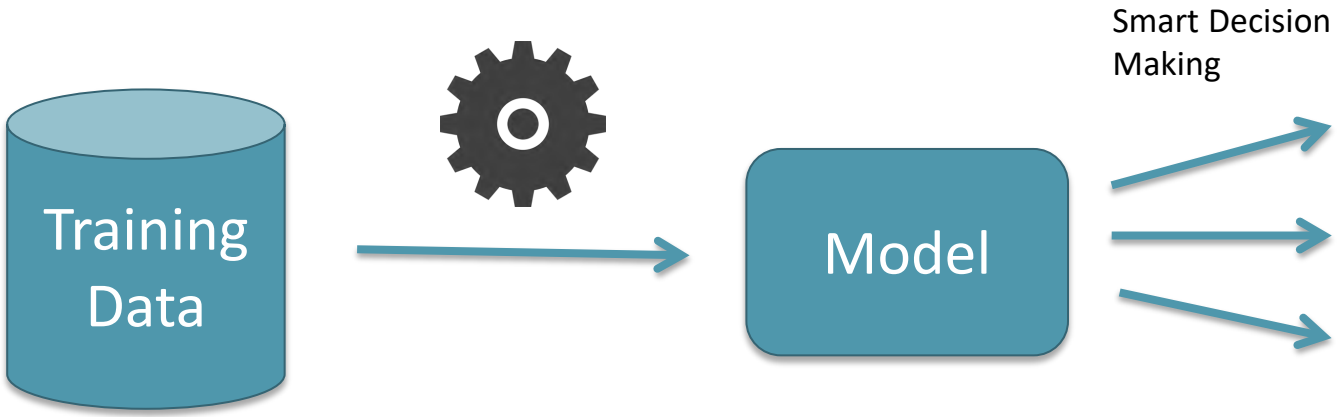
- Umweltauswirkungen und finanzielle Kosten des Trainings solcher Modelle
- Problematische Datensets, welche Stereotypen enthalten
- Interpretation von Bedeutung in generierten Texten

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

„*[…]* only a very small number of the over 7000 languages of the world are represented in the rapidly evolving language technologies and applications.»

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Explainable AI



Wie kamen die Entscheide zustande?

Image Source: pixabay.com

Wie soll die digitale Gesellschaft der Zukunft aussehen?

Image Source: pixabay.com

Zwei Arten von AI

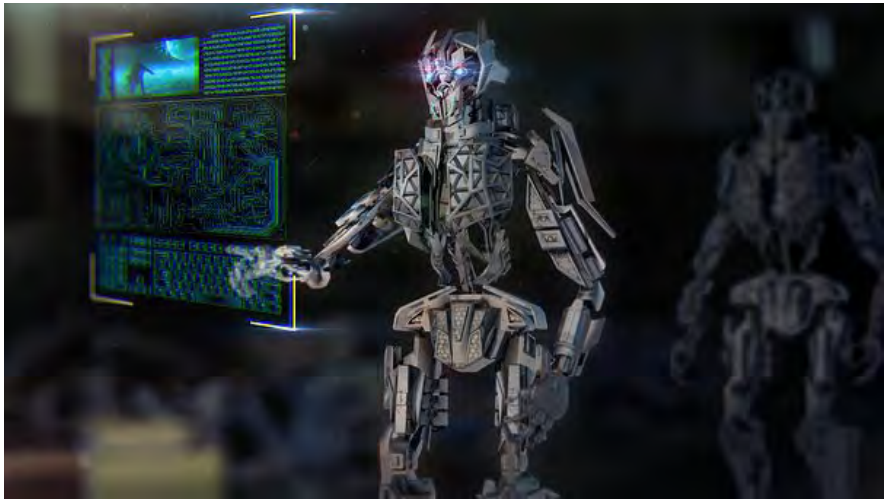


Image Source: pixabay.com

Augmented Intelligence, statt Artificial Intelligence

- Menschen unterstützen, anstelle sie zu ersetzen

- AI als Tool zur Unterstützung bei repetitiven Arbeiten, damit der Mensch mehr Zeit hat für andere Aufgaben

- Automatische Verarbeitung von Dokumenten, etc.

Image Source: pixabay.com

Berner Fachhochschule

Kontakt

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

mascha.kurpicz@bfh.ch

@SocietyData