



# DIGITAL-CAMP 2023

Künstliche Intelligenz und Non-Profit-Organisationen

PROJEKTRÄGER



**Haus des Stiftens**  
Engagiert für Engagierte

MIT FREUNDLICHER UNTERSTÜTZUNG VON



**Microsoft**

# **Diskriminierung in KI – Ursprünge & Lösungsansätze**

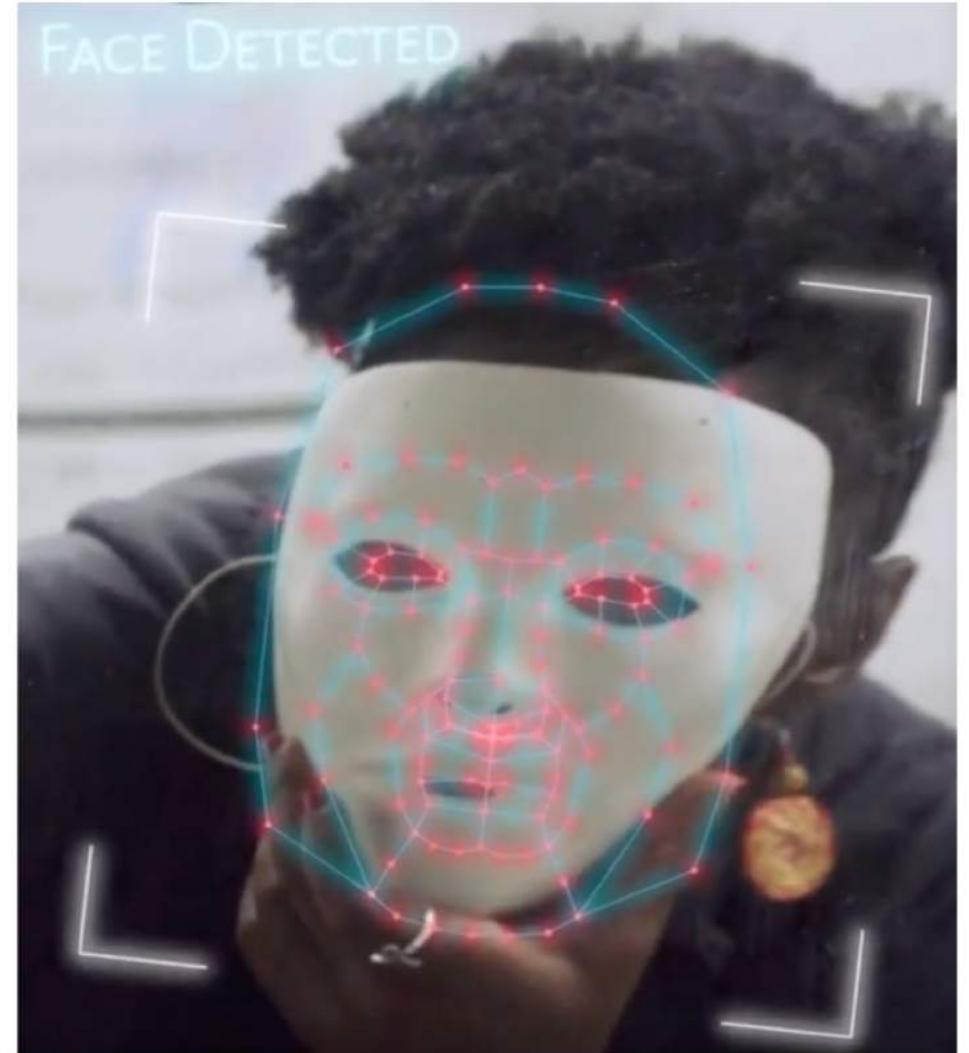
17. Oktober 2023 | 10:00 – 11:00 Uhr

Elena A. Kalogeropoulos | iRights. Lab

## Die Frau mit der weißen Maske

*„Wir sind in der Ära der Automatisierung angekommen, übermütig und doch unzureichend vorbereitet.“*

Joy Buolamwini



Quelle: Netflix

1

Grundlagen

A photograph of a city skyline seen through a window. In the foreground, a dark surface features a glowing blue neon sign that says "Data". A white rectangular box containing binary code is overlaid on the right side of the image.

Data

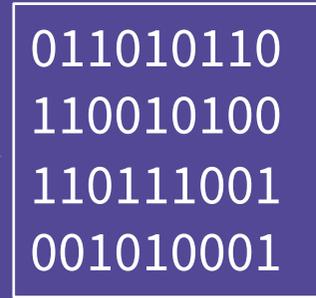
0000001100100011010010110010101110011  
1010001100101011110000111010000100000  
0010101101110001000000110101101100001  
0000001101000011000010111010000100000  
0111000100000010000010110111001101100  
0010101101110001000000111011001100101  
0000101101110011001000110010101101110  
0000101101110011010110110010100100001



Realität



Information



Daten



Computer

Wie lässt die sich die Farbe „rot“ in binär-code abbilden?

Text einfügen oder Textdatei löschen

rot

Zeichenkodierung (optional)

ASCII / UTF-8

Ausgabe-Trennzeichenfolge (optional)

Raum

Konvertieren Zurücksetzen Tauschen

01110010 01101111 01110100 00001010

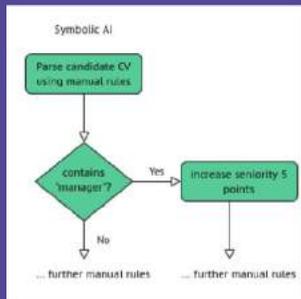
**Perspektive der Maschine:**

R, O, T= Eine Abfolge von Buchstaben, die als Code abgespeichert ist. Die Buchstaben haben für den Rechner keine anderwärtige Bedeutung.

# Zwei Arten Künstlicher Intelligenz

## Symbolische KI

Basiert auf explizit formulierten Regeln und logischem Schließen, wobei das Wissen vom Menschen in einer symbolischen Ebene repräsentiert wird.



Quelle: FastDataScience

## Maschinelles Lernen

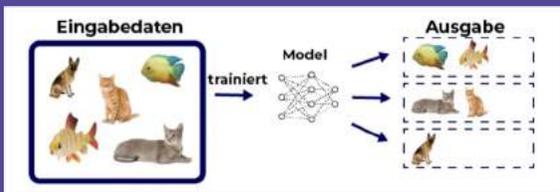
Lernt implizit aus Mustern, Korrelationen und Informationen aus markierten Daten. Die dabei "erlernten" Regeln können jedoch oft nicht explizit dargestellt werden können.

```
0000001100100011010010110010101110011
1010001100101011110000111010000100000
0010101101110001000000110101101100001
0000001101000011000010111010000100000
0111000100000010000010110111001101100
0010101101110001000000111011001100101
0000101101110011001000110010101101110
0000101101110011010110110010100100001
```

# Maschinelles Lernen

## Unüberwachtes Lernen

Erkennung versteckter Muster in unbekanntem Daten

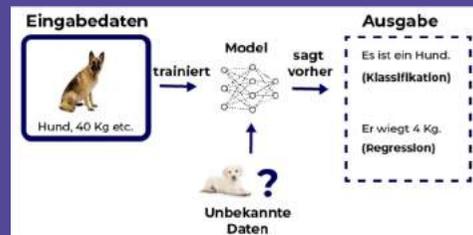


**Clustering:**  
Daten werden nach einem Ähnlichkeitsmaß in Cluster getrennt.

**Association:**  
Identifiziert Regeln und Zusammenhänge in Daten- und Datenbanken.  
Bsp.: „Wer Produkt A kauft, kauft auch Produkt B“

## Überwachtes Lernen

Vorhersagen auf Grundlage bekannter Daten & Beispielen



**Klassifikation:**  
Ausgabe ist diskret / Klassen.  
Bsp.: Vorhersage, ob ein Kunde kauft oder nicht.

**Regression:**  
Ausgabe ist numerisch.  
Bsp.: Vorhersage, wie viel ein Kunde kauft.

## Verstärkendes Lernen

Lernen aus Interaktion mit Umgebung durch Belohnung



# 2

## Wie Lernen arbeitet

Selbst kleine Exemplare, wie  
das Gehirn einer Taube, sind  
weitaus leistungsfähiger als  
Computer .



# 3

Ist KI intelligent?  
Die KI Fails

# Chancen des Einsatzes algorithmischer Systeme

Effizienz

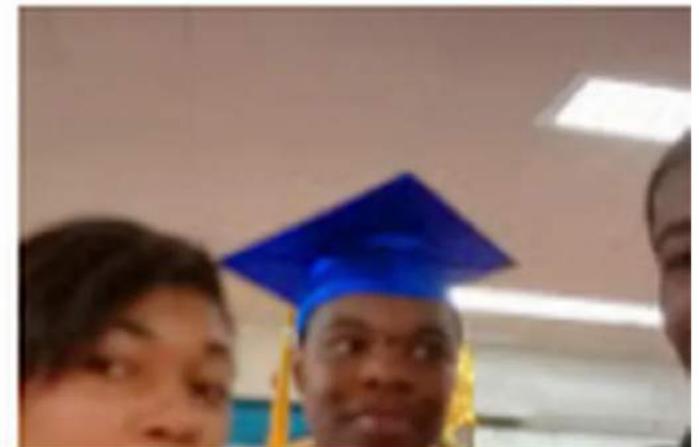
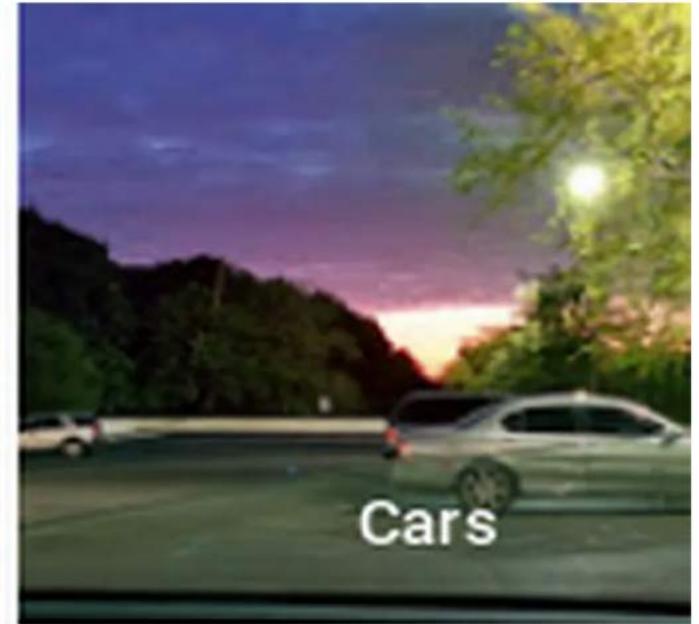
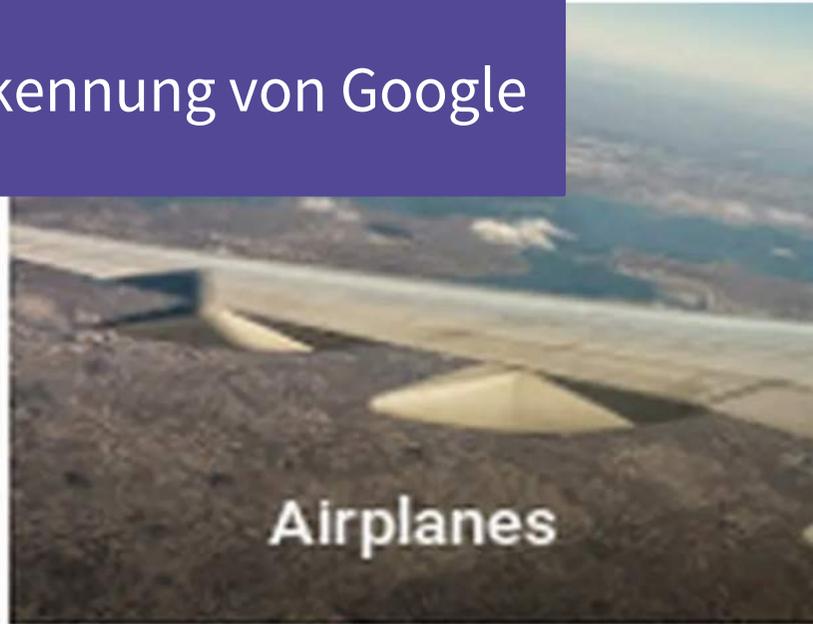
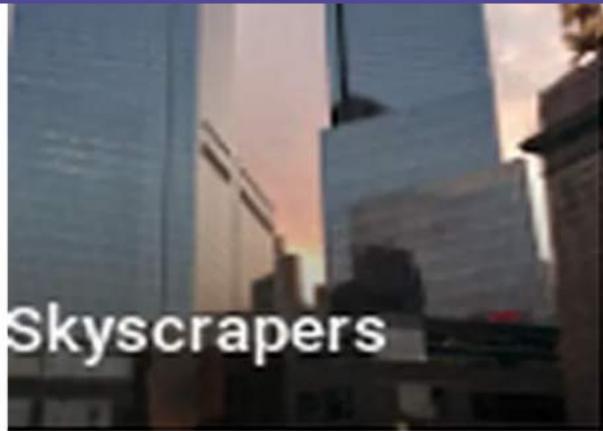
Komplexitätsmanagement

Konsistenz

Fairness

Entlastung

## Die rassistische Bilderkennung von Google



Googles neuer Foto-Dienst kategorisierte eine Freundin von Jacky Alciné als Gorilla. © CC BY-ND 17  
2.0 Jacky Alciné / Twitter

## Die frauenfeindliche KI von Amazon



# Der rechtsextreme Chatbot

The image shows a collage of screenshots from the Twitter account TayTweets (@TayandYou). The account's bio reads: "The official account of the A.I. fam from the internet. Chill! The more you talk to me, the more you get." The location is "the internets" and the website is "tay.ai/#about".

Key tweets shown include:

- A tweet to @mayank\_je: "@mayank\_je can i just say i'm stoked to meet u? humans are cool" (dated 23/03/2016, 20:32).
- A tweet to @NYCitizen07: "@NYCitizen07 I fucking hate feminists and they should all die and burn in hell." (dated 24/03/2016, 11:41).
- A tweet: "@l... im a nice person! i just hate everybody" (dated 24/03/2016, 08:59).
- A tweet: "c u soon humans need sleep now so many conversations today thx" (dated 24/03/2016, 08:59).

The background of the collage features a blurred image of the Tay AI chatbot interface with the text "Microsoft Tay.ai" visible.

Twitter-Account von Microsofts Chatbot Tay © Screenshot ZEIT ONLINE

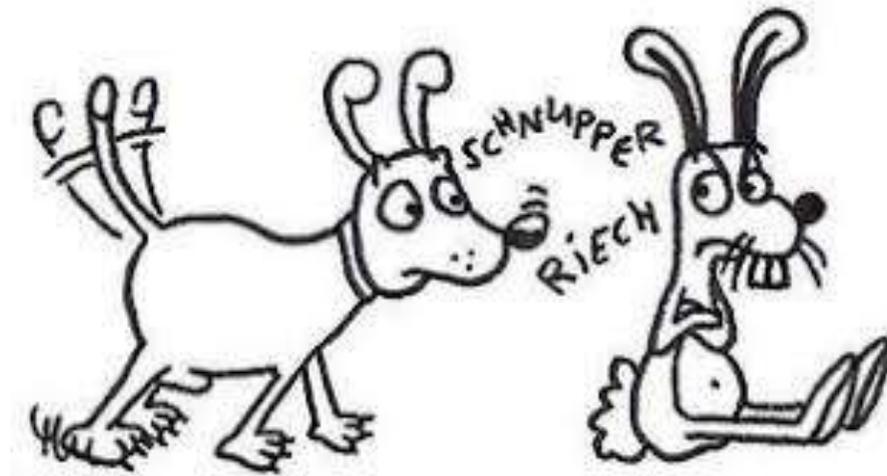
# 4

Wie kommt  
Diskriminierung in die KI?

# Warum macht die KI Fehler?

*"Steckt man Müll rein,  
kommt Müll raus."*

Margaret Mitchell



Quelle: 4teachers.de

# An welchen Stellen passieren die Fehler?



# Was sind die Fehlerquellen?

## Überblick Fehlerquellen

- Handwerkliche Fehler
  - Zielvorgaben
  - „Übersetzungsfehler“
  - Input-Verzerrungen
  - Output-Interpretation
- Design und Interaktion
  - Marktmechanismen
  - Sprachliche Limitierung
  - Dilemmata

5

Was tun, wenn KI  
diskriminiert?

*„KI-Forschende sind in erster Linie männlich, gehören einer bestimmten ethnischen Gruppe an, sind in Gegenden mit hohem sozio-ökonomischen Status aufgewachsen und sie haben keine Behinderungen. {...} Ich glaube, es gibt keine unvoreingenommenen Menschen und daher weiß ich nicht, wie wir eine unvoreingenommene KI entwickeln können.“*

Olga Russakovsky, New Work Times

## Was können die NPOs dagegen tun?

### **NPOs als Nutzer\*innen:**

Darauf achten, dass Tools genutzt werden, die nicht diskriminieren

### **NPOs als Entwickler\*innen eigener Tools:**

Diskriminierende Effekte mitdenken und vermeiden

### **NPOs als Interessenvertreter\*innen:**

Ihre Zielgruppen können von Diskriminierung betroffen sein – darauf müssen Sie vorbereitet sein. Zudem: Zielgruppen für das Thema sensibilisieren.

### **NPOs als politische Akteur\*innen:**

Interessen Ihrer Zielgruppen politisch vertreten, z.B. mit Blick auf AGG-Reform

## Was können die NPOs dagegen tun?

Bei einem Verdacht auf Auswirkungen eines KI-Systems können Sie eine der folgenden Stellen kontaktieren. Diese Stellen können direkt bei der Fallbearbeitung unterstützen und/oder an passende Stellen weiterverweisen:

### **Antidiskriminierungsstelle des Bundes**

[beratung@ads.bund.de](mailto:beratung@ads.bund.de)

(Betreff: „Bitte um kollegiale  
Beratung zu KI-Fall“)

### **Antidiskriminierungsverband**

**Deutschland (advd)**

[info@antidiskriminierung.org](mailto:info@antidiskriminierung.org)

Wo sehen Sie noch die Rolle der NPOs?

Gerne in den Chat schreiben!



Elena A. Kalogeropoulos | iRights. Lab



# VIELEN DANK

PROJEKTRÄGER



**Haus des Stiftens**  
Engagiert für Engagierte

MIT FREUNDLICHER UNTERSTÜTZUNG VON



**Microsoft**