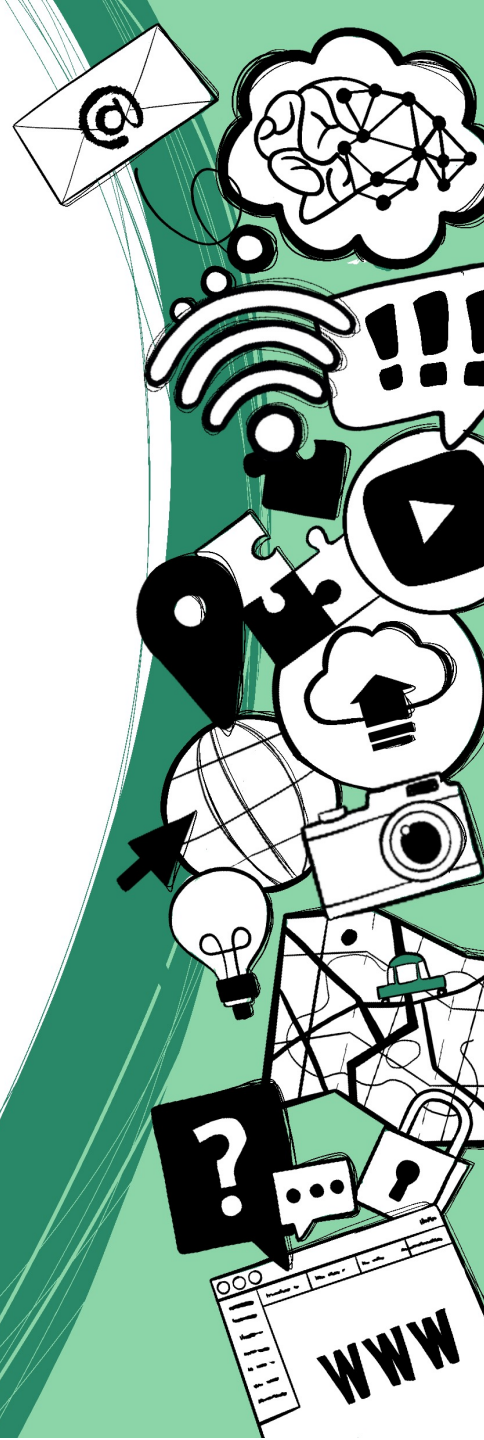
ENARIS

# Bias-Fehler in KI-Systemen



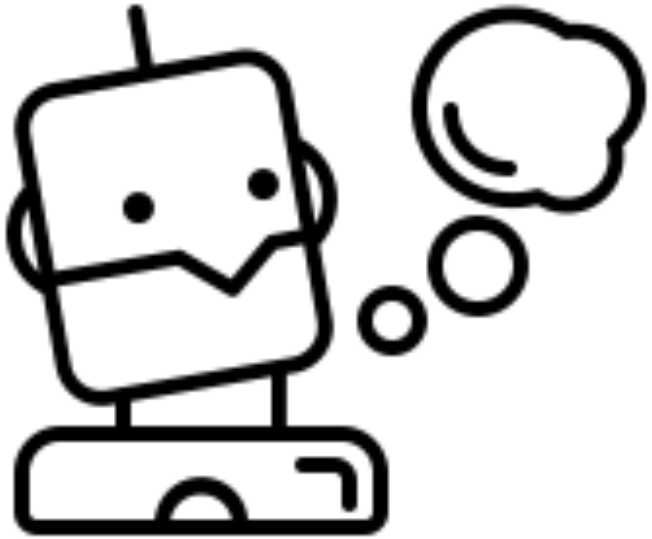

mobilis
interaktív kiállítási központ

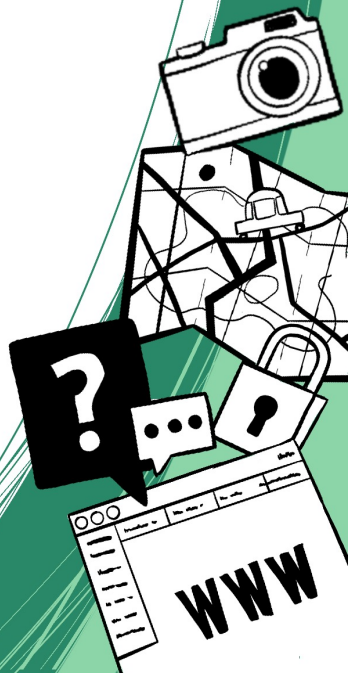
TU Graz

OESTERREICHISCHE COMPUTER GESELLSCHAFT ®
AUSTRIAN COMPUTER SOCIETY

Interreg
Austria-Hungary 2014-2020
European Union – European Regional Development Fund

# Was sind Bias-Fehler überhaupt?



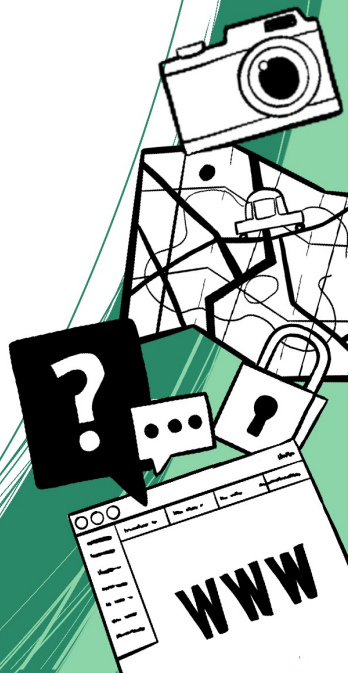


# Was sind Bias-Fehler überhaupt?

- Verzerrung oder Voreingenommenheit

  → entstehen bei der Erhebung von notwendigen Daten für das System

  → können zu schwerwiegenden Problemen im System und zu Diskriminierung führen



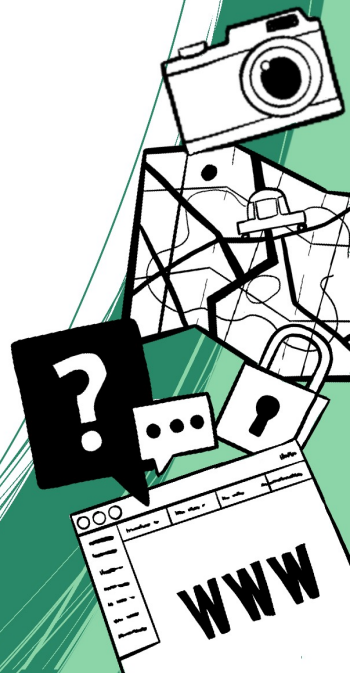
# Was sind Bias-Fehler überhaupt?

Beispiel: **Chatbot Tay** (2016)

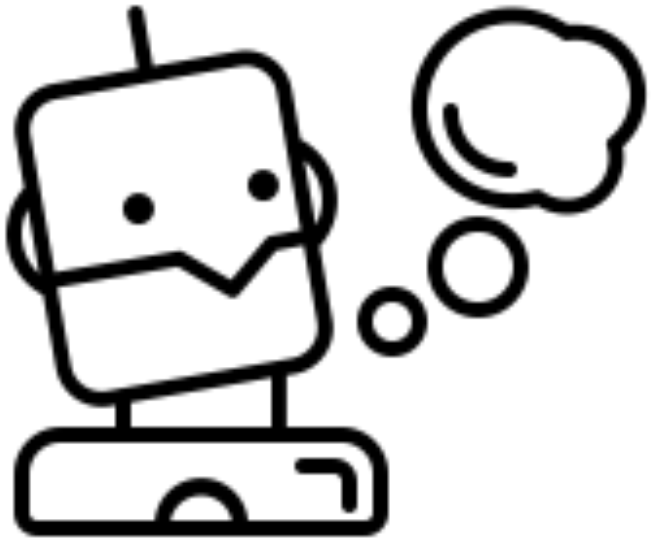
Twitter-Chatbot, der durch die Tweets echter Nutzer*innen „lernte“

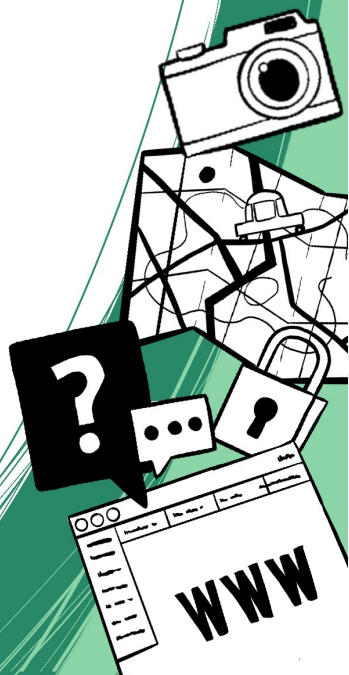
→ begann sehr schnell diskriminierende Nachrichten zu veröffentlichen


→ nach nur **16 Stunden** wurde der Chatbot wieder offline genommen



# Welche Arten von Bias-Fehler gibt es?

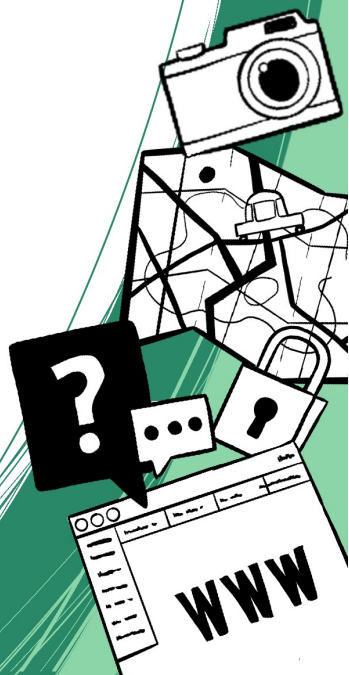




# Welche Arten von Bias-Fehler gibt es?

- **algorithmic AI bias** or "**data bias"**: durch eingespeiste Daten ein Bias-Fehler begangen (statistische Verzerrung der Daten)

- **societal AI bias:** von der Gesellschaft indoktrinierte Normen; aber auch Stereotype erschaffen blinde Flecken oder Voreingenommenheit (gesellschaftliche Vorurteile)

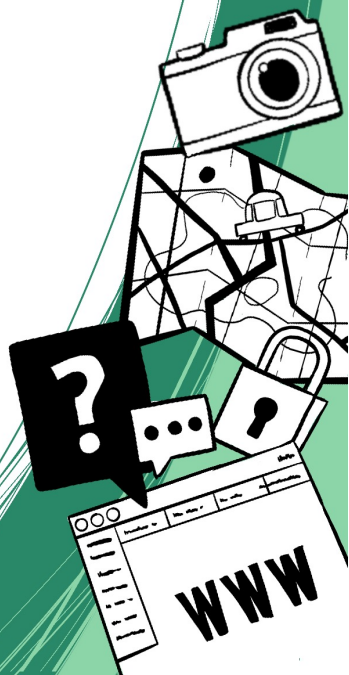


# Welche Arten von Bias-Fehler gibt es?

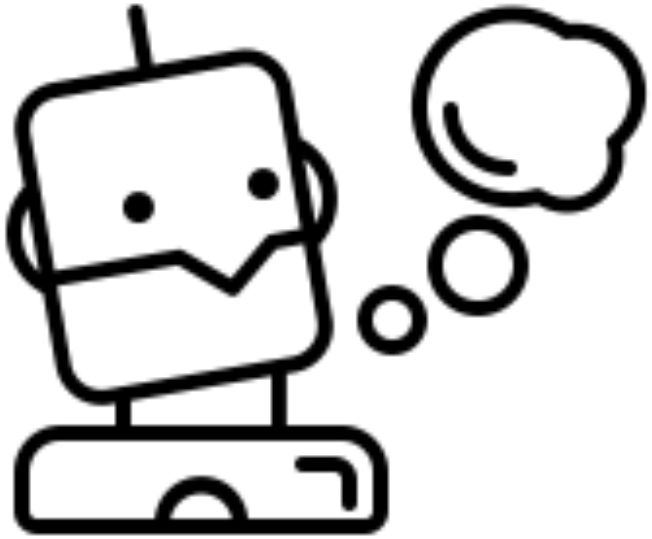
- **algorithmic AI bias** or "**data bias"**: durch eingespeiste Daten ein Bias-Fehler begangen (statistische Verzerrung der Daten)

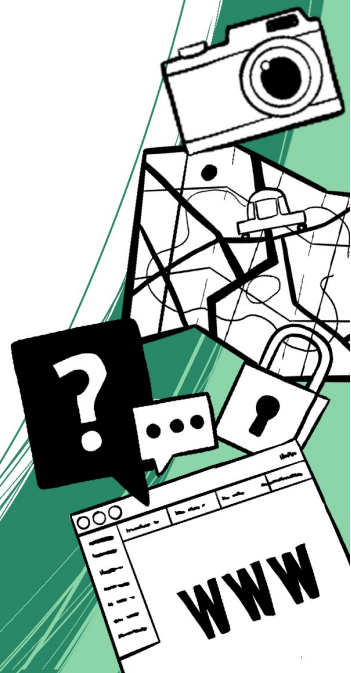
- **societal AI bias:** von der Gesellschaft indoktrinierte Normen; aber auch Stereotype erschaffen blinde Flecken oder Voreingenommenheit (gesellschaftliche Vorurteile)

**→ Societal Bias beeinflusst/schafft häufig algorithmische Bias-Fehler!**



# Warum braucht es ethische Regeln bei künstlicher Intelligenz?

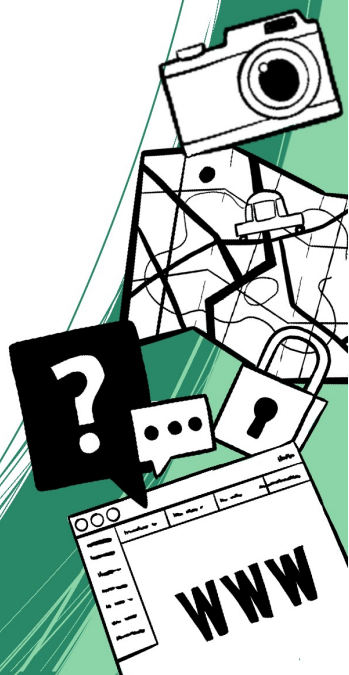




# Warum braucht es ethische Regeln bei künstlicher Intelligenz?

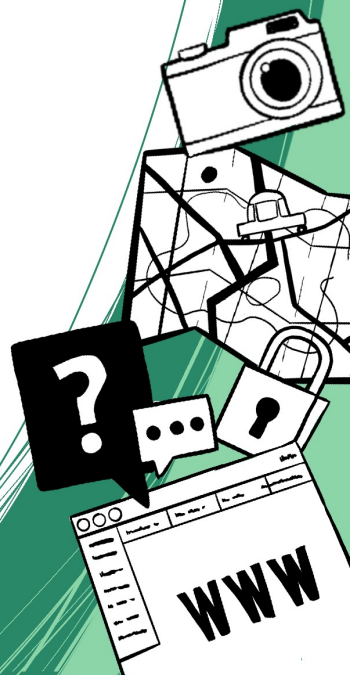
- Programmierer*innen können ihre eigenen Vorurteile (unabsichtlich) in Programme einbauen

- Durch ethische Regeln wird versucht, dass kein Mensch ausgeschlossen und diskriminiert wird



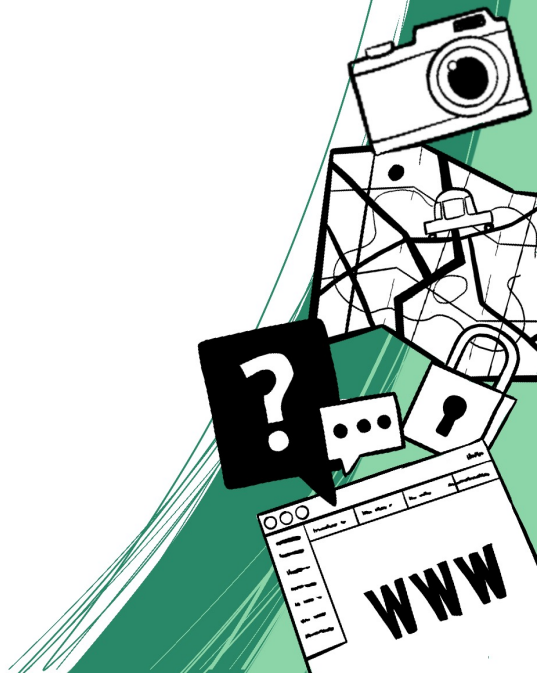
# *Ethikrichtlinien für eine vertrauenswürdige KI*

- **Fairness**
- **Achtung der menschlichen Autonomie**
- **Schutz vor Schaden**
- **Nachvollziehbarkeit**



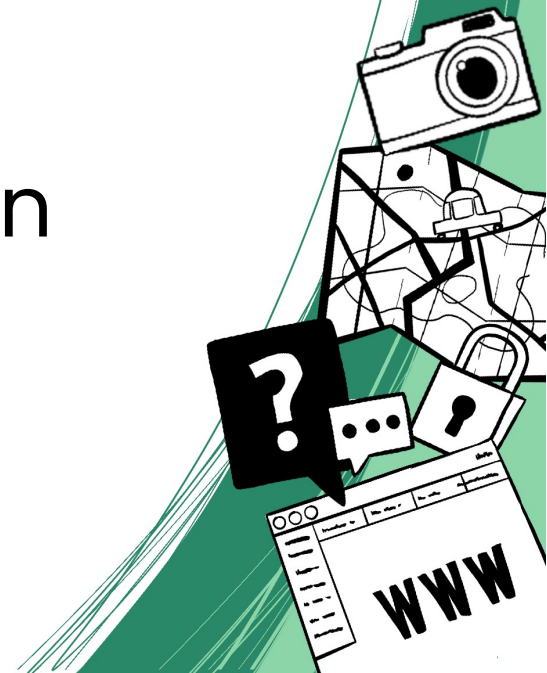
# Fairness

- vor Diskriminierung schützen
- Chancengleichheit
- Menschen dürfen nicht getäuscht werden
- KI-Systeme müssen transparent sein



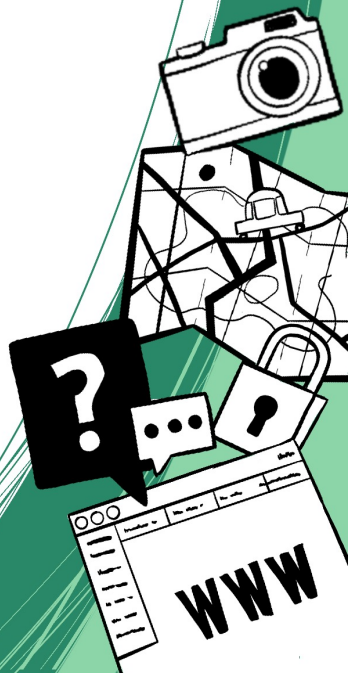
# Achtung der menschlichen Autonomie

- Selbstbestimmung (Ich darf über mich selbst entscheiden)
- Grundrechte ausleben
- KI-Systeme sollen Menschen stärken und fördern
- Aufsicht von KI-Systemen durch Menschen



# Schutz vor Schaden

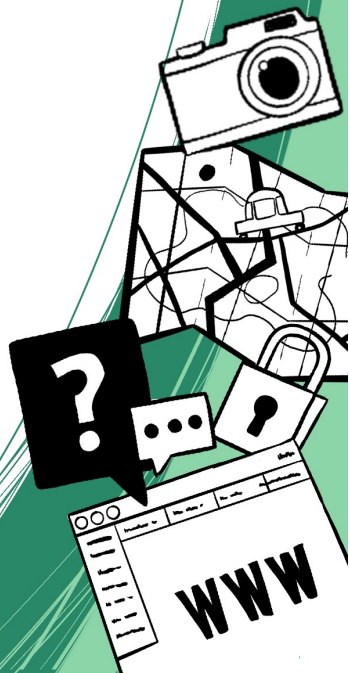
- KI darf keine Schäden verursachen noch diese verschlimmern (geistige und körperliche Unversehrtheit)
- KIs müssen technisch robust sein
- Rücksicht auf schutzbedürftige Personen (Kinder, beeinträchtigte Personen…)
- ungleiche Macht- oder Informationsverteilungen (Bsp. Staat und Bürger*innen)



# Nachvollziehbarkeit

- Prozesse sollen transparent sein

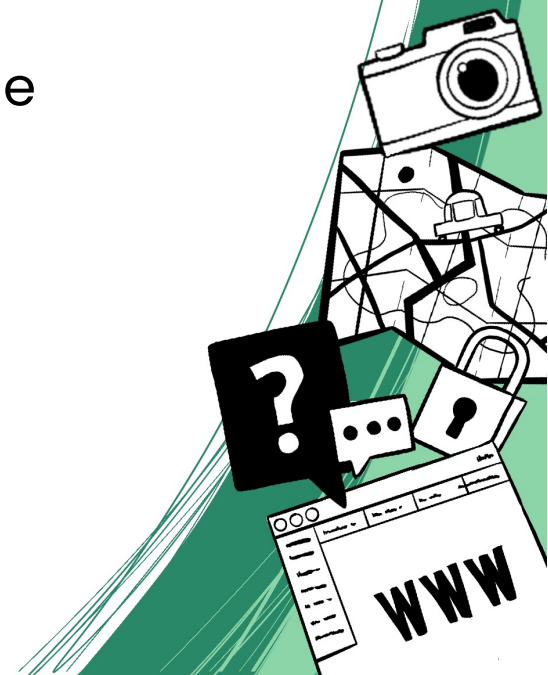
- Achtung bei **„Blackbox-Algorithmen“** (Hier ist nicht ganz klar ist, wie ein System zum jeweiligen Ergebnis kommt)



# ACHTUNG!

- Manchmal lassen sich diese vier Bereiche nicht vereinen!

- Bsp. „vorausschauende Polizeiarbeit“
  Spezielle Überwachsungsmaßnahmen können dann zwar bei der Verbrechensbekämpfung helfen, schränken dabei aber zugleich die eigenen Freiheits- und Datenschutzrechte ein.



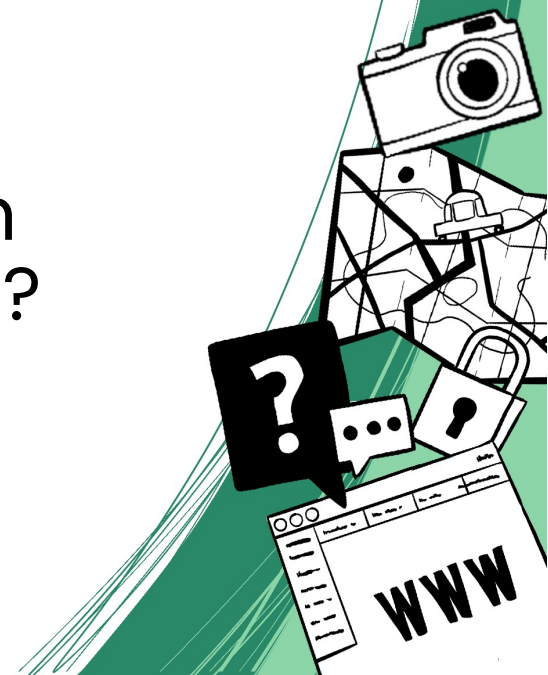
# Wie können Bias-Fehler vermieden werden?

## 1. Selbstreflexion

Einnehmen unterschiedlicher Perspektiven: Erwische ich mich bei Schubladendenken? Hab ich Vorurteile gegenüber anderen?

## 2. Aktive Kommunikation

Fällt mir etwas auf in einem Programm? Fühle ich mich dadurch ausgeschlossen oder diskriminiert?
→ direkt darauf ansprechen



# Wie können Bias-Fehler vermieden werden?

## 1. Selbstreflexion

Einnehmen unterschiedlicher Perspektiven: Erwische ich mich bei Schubladendenken? Hab ich Vorurteile gegenüber anderen?

## 2. Aktive Kommunikation

Fällt mir etwas auf in einem Programm? Fühle ich mich dadurch ausgeschlossen oder diskriminiert?
→ direkt darauf ansprechen

## Habt ihr noch weitere Ideen?