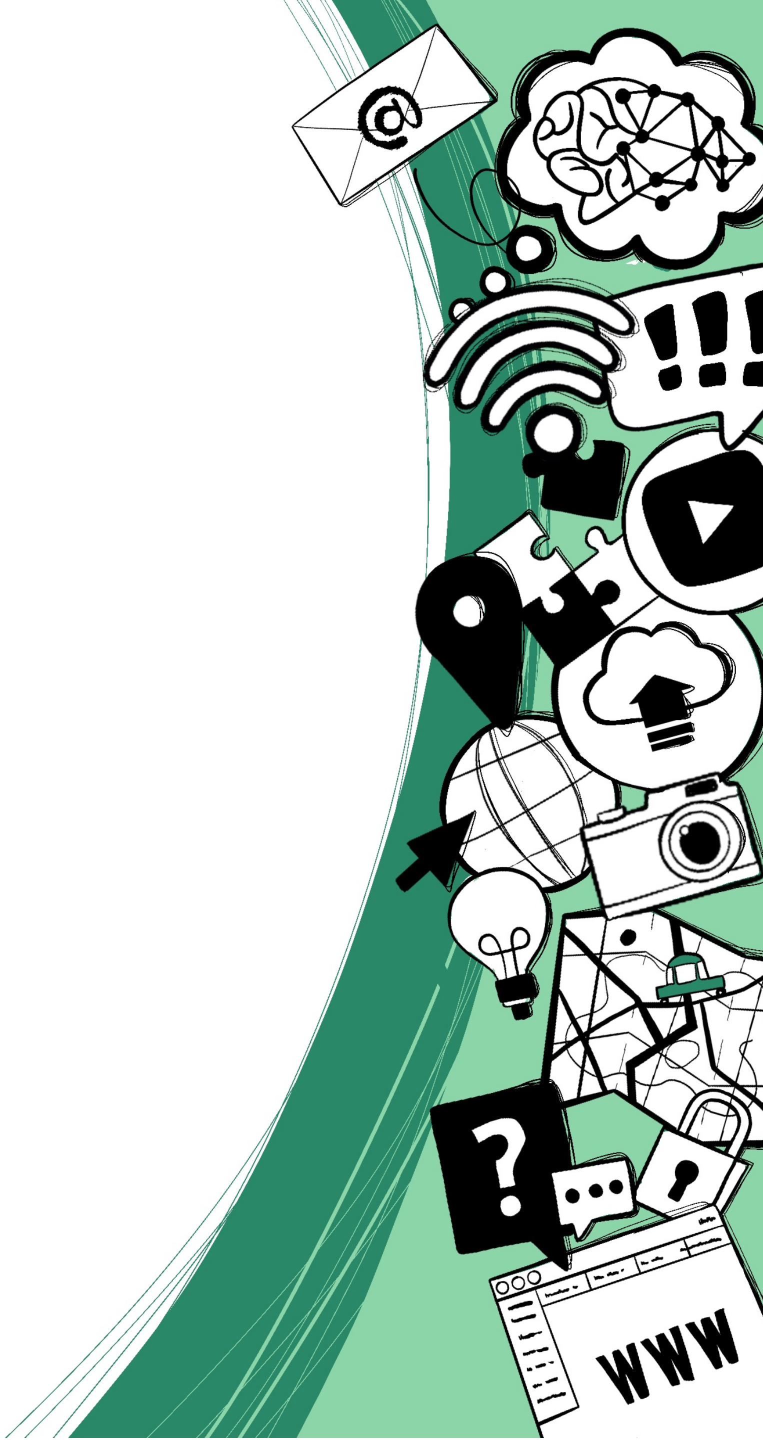
ENARIS

Bias errors in AI systems



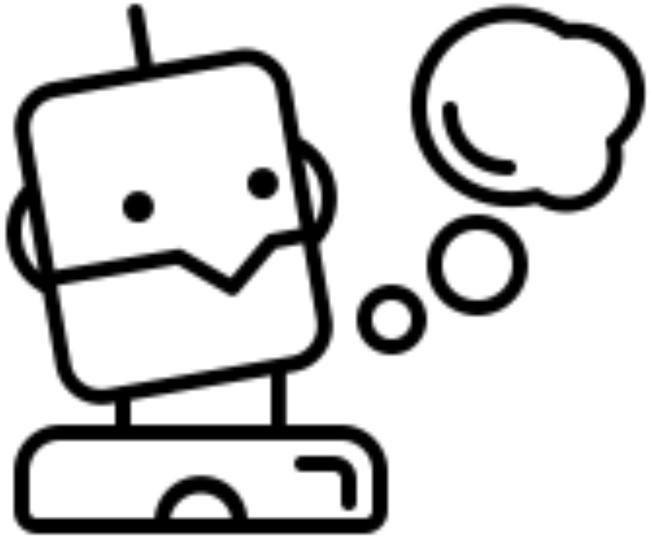
mobilis
interaktív kiállítási központ

TU Graz

OESTERREICHISCHE COMPUTER GESELLSCHAFT ®
AUSTRIAN COMPUTER SOCIETY

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

What are bias errors anyway?

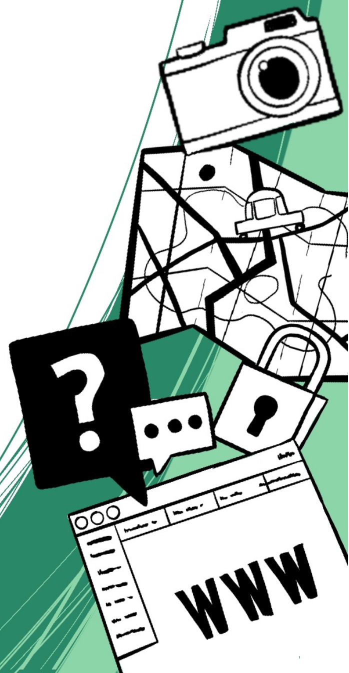


What are bias errors anyway?

- Distortion or Prejudice

  → arise when collecting the data necessary for the system

  → can lead to serious problems in the system and discrimination

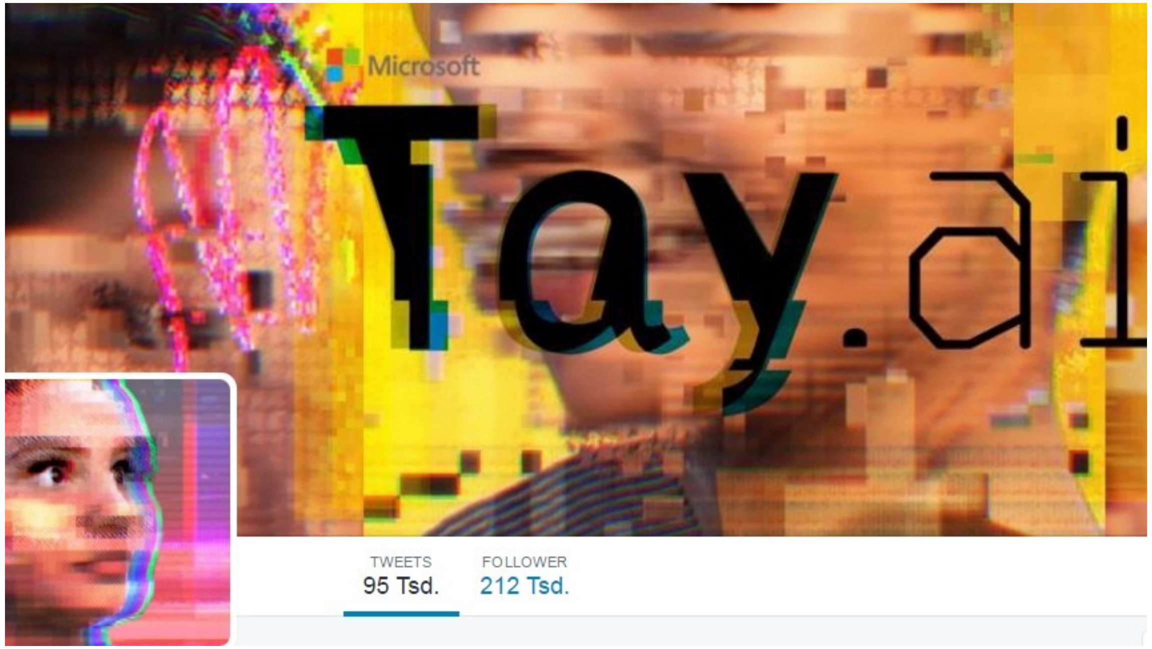


What are bias errors anyway?

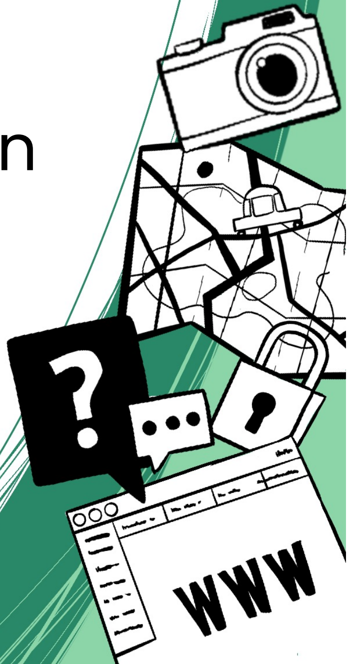
Example: **Chatbot Tay** (2016)

Twitter chatbot that "learned" from real users' tweets

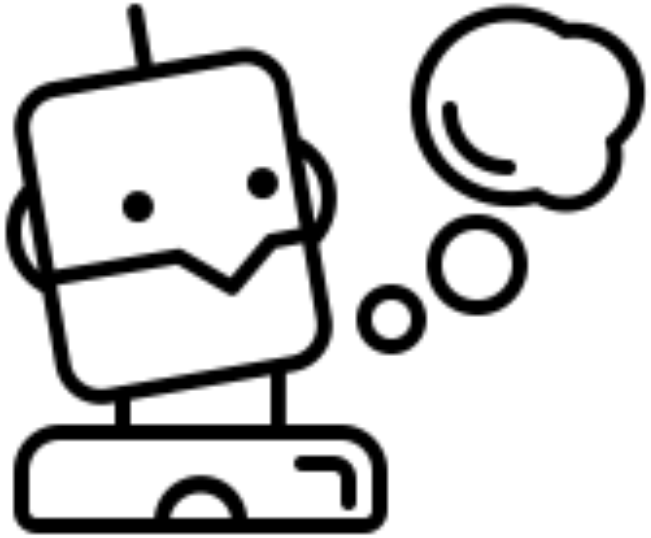
→ began posting discriminatory messages very quickly



→ after only 16 hours the chatbot was taken offline again



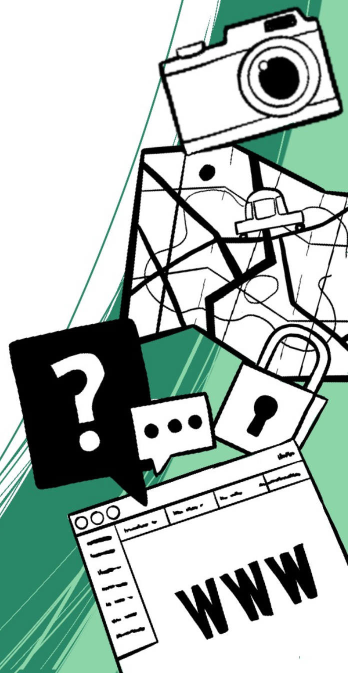
What types of bias errors are there?



What types of bias errors are there?

- **algorithmic AI bias** or “**data bias”**: Bias error caused by data fed in (statistical distortion of the data)

- **societal AI bias:** norms indoctrinated by society; but stereotypes also create blind spots or prejudices (social prejudices)

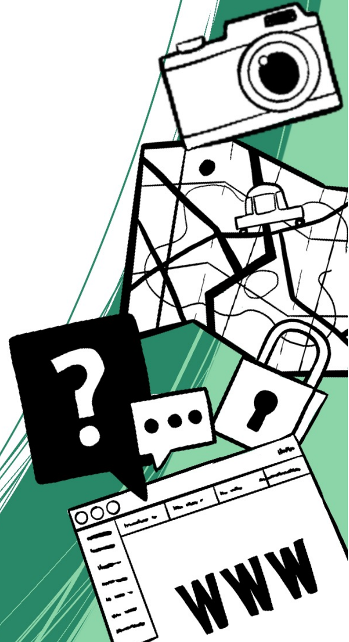


What types of bias errors are there?

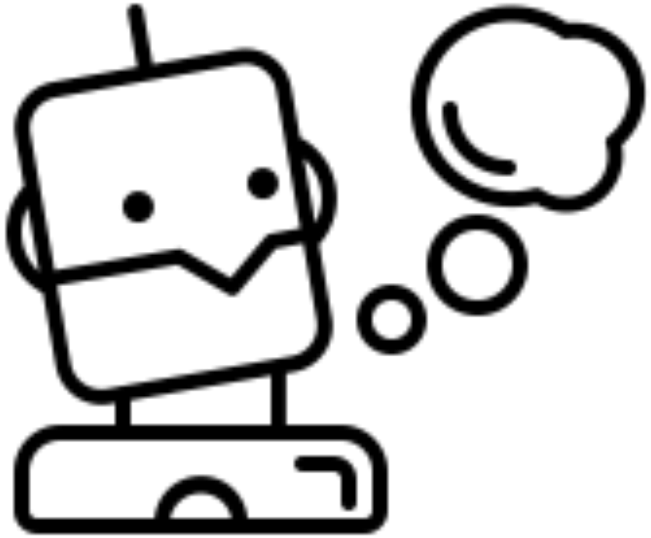
- **algorithmic AI bias or “data bias”: Bias error caused by data fed in (statistical distortion of the data)**

- **societal AI bias:** norms indoctrinated by society; but stereotypes also create blind spots or prejudices (social prejudices)

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



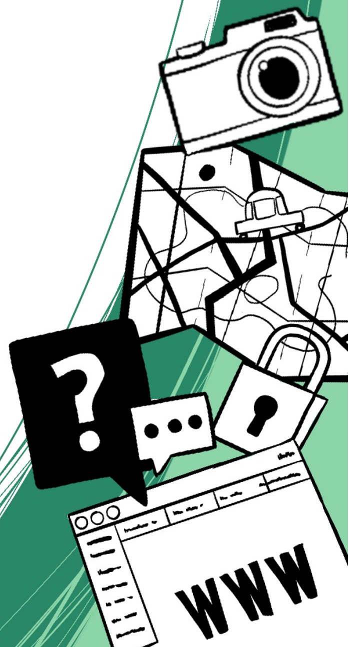
Why do we need ethical rules for artificial intelligence?



Why do we need ethical rules for artificial intelligence?

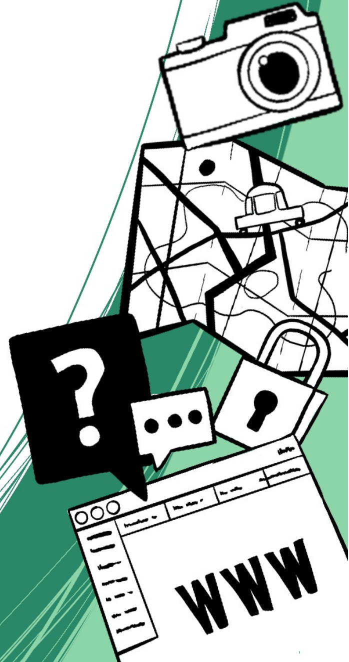
- Programmers can (unintentionally) build their own prejudices into programs

- Ethical rules try to ensure that no one is excluded and discriminated against



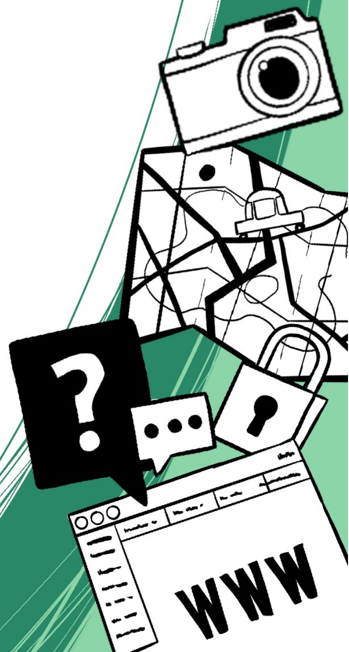
Ethical guidelines for a trustworthy AI

- Fairness
- Respect for human autonomy
- Protection from harm
- Traceability



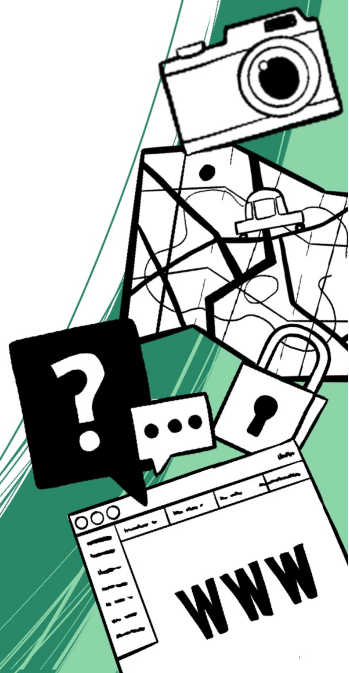
Fairness

- protect against discrimination
- equal opportunity
- People must not be deceived
- AI systems have to be transparent



Respect for human autonomy

- Self-determination (I can decide about myself)
- Living basic rights
- AI systems are designed to empower and encourage people
- Human oversight of AI systems



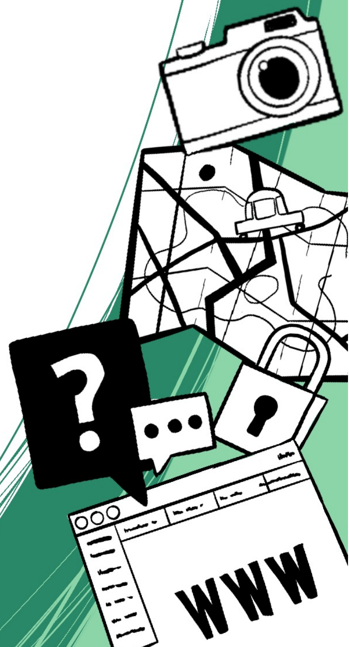
Protection from harm

- AI must neither cause nor aggravate damage (mental and physical integrity)

- AIs have to be technically robust

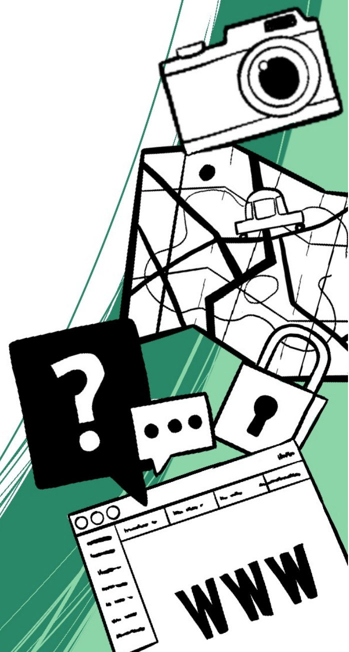
- Consideration for vulnerable people (children, disabled people …)

- unequal distribution of power or information (e.g. state and citizens)



Traceability

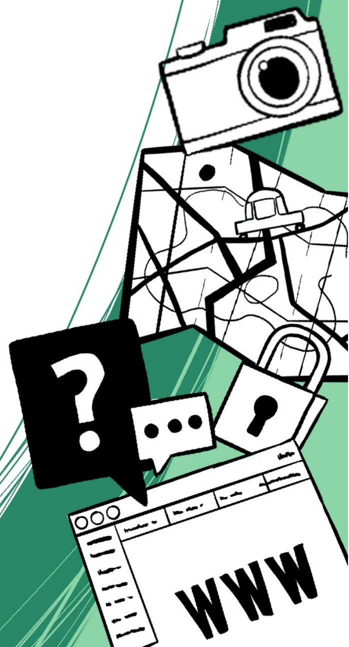
- Processes should be transparent
- Beware of "**black box algorithms**" (it is not entirely clear here how a system comes to the respective result)



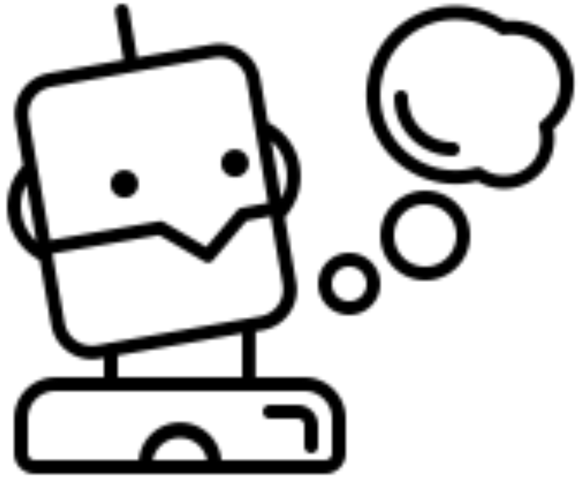
Watch out!

- Sometimes these four areas cannot be combined!

- E.g. "predictive police work"
  Special surveillance measures can then help in the fight against crime, but at the same time limit one's own freedom and data protection rights.



How can bias errors be avoided?



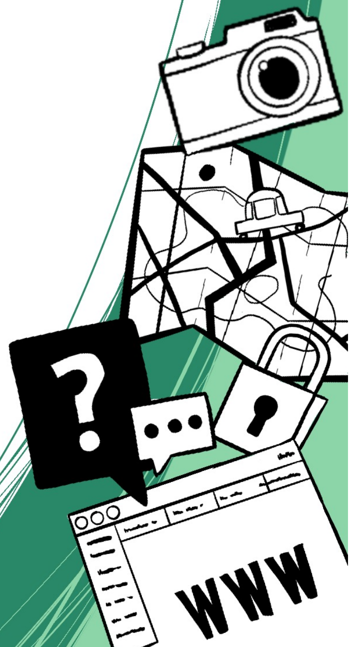
How can bias errors be avoided?

1. **Self Reflection**

Taking different perspectives: Do I catch myself thinking in stereotypes? Do I have prejudices against others?

2. Active communication

Do I notice something in a program? Do I feel excluded or discriminated against as a result?
→ address it directly



How can bias errors be avoided?

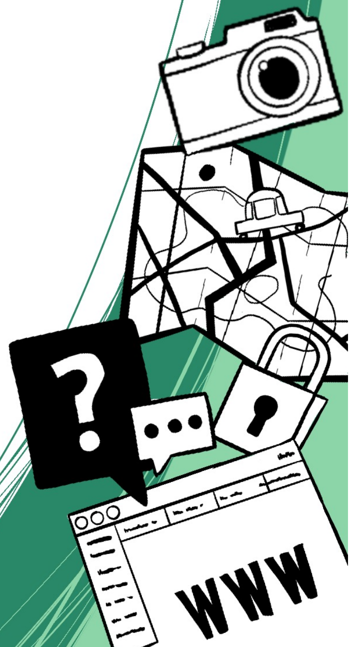
1. **Self Reflection**

Taking different perspectives: Do I catch myself thinking in stereotypes? Do I have prejudices against others?

2. Active communication

Do I notice something in a program? Do I feel excluded or discriminated against as a result?
→ address it directly

Do you have any other ideas?



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?