HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Internationales Zentrum
für Ethik in den
Wissenschaften (IZEW)

Eberhard Karls
Universität
Tübingen

# Responsible Data Science

## Session 5: 29.05.2023, 13.30 – 16.45 h
## MA Seminar, SoSe 2024, Hasso-Plattner-Institute

Dr. Simon David Hirsbrunner

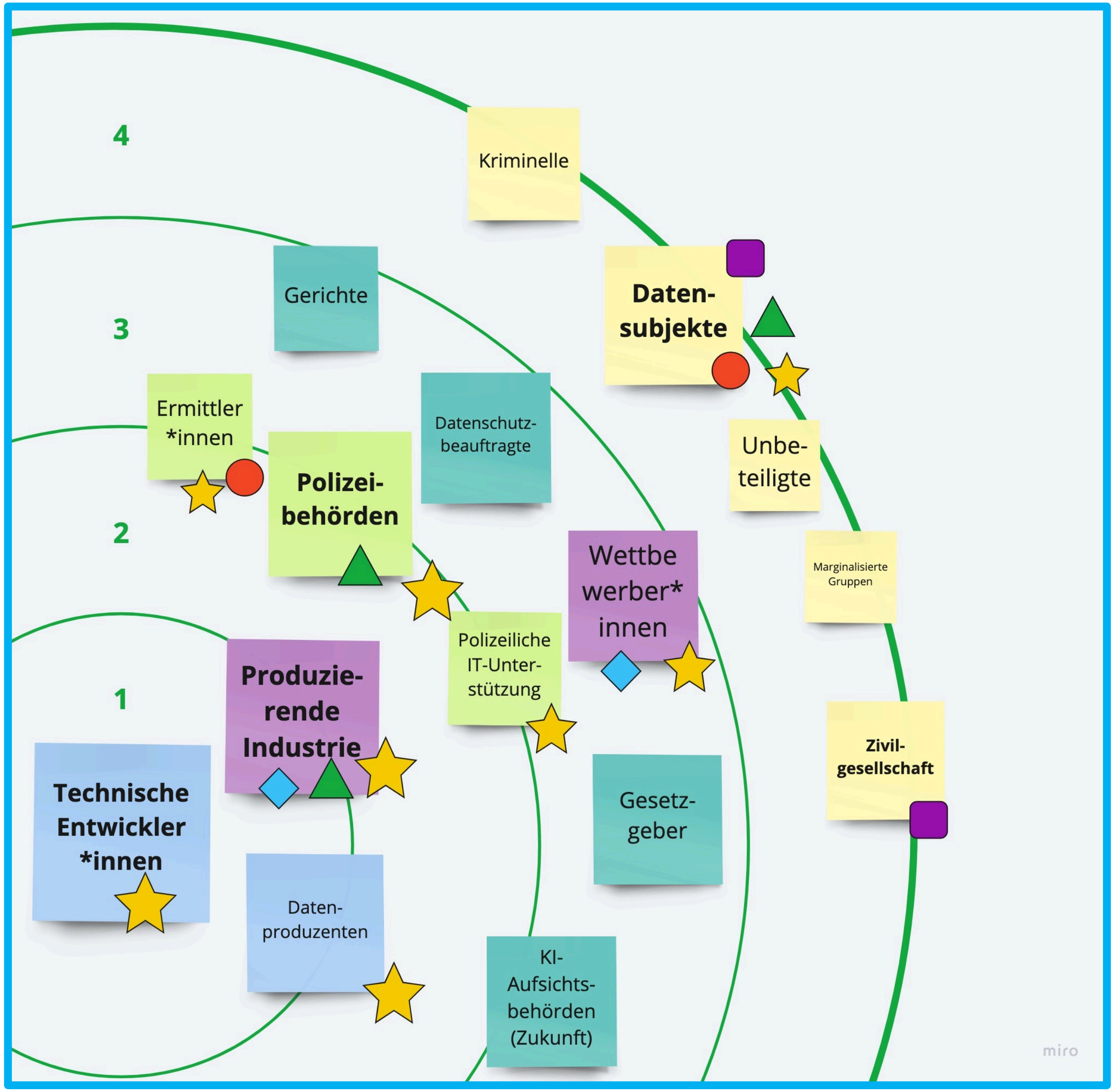
# Today

| topic | time |
|---|---|
| Introduction | 13h30 |
| AI risks – definitions, guidelines, identification methods | 13h40 |
| Guest input with use case (Alejandro Sierra-Múnera):<br>introducing a system for image captioning + Q&A | 14h00 |
| Exercise in small groups: identifying risks for the specified AI system | 14h30 |
| Sharing and discussing insights from small groups | 15h00 |
| —— Break —— | 15h35 |
| Student presentation: Deepfakes | 15h50 |
| Discussion | 16h10 |
| Assignments for next week | 16h30 |
| End | 16h45 |

# What are stakeholders?

- “A stakeholder is **anyone who will be affected**, directly or indirectly, by the new system like the end users, the software staff, and the organization’s clients." (Shneiderman and Rose 1996: 92)

- Stakeholders “can be people, groups, neighborhoods, communities, organizations, institutions, or societies, and can also include past and future generations, nonhuman species, and other elements such as historic buildings or sacred mountaintops” (Friedman and Hendry 2019: 37)


4
Kriminelle
Daten-subjekte
3
Gerichte
Ermittler *innen
Datenschutz-beauftragte
Unbe-teiligte
Polizei-behörden
2
Wettbe werber* innen
Marginalisierte Gruppen
Polizeiliche IT-Unter-stützung
1
Produzie-rende Industrie
Zivil-gesellschaft
Technische Entwickler *innen
Gesetz-geber
Daten-produzenten
KI-Aufsichts-behörden (Zukunft)
miro

# AI risks for stakeholders

**Definition of risk**
„general probability of negative consequences to actions“

(see Cambridge Dictionary, 2022 in Lütge et al. 2022).

# AI risks for stakeholders

**Definition of risk**
„general probability of negative consequences to actions“

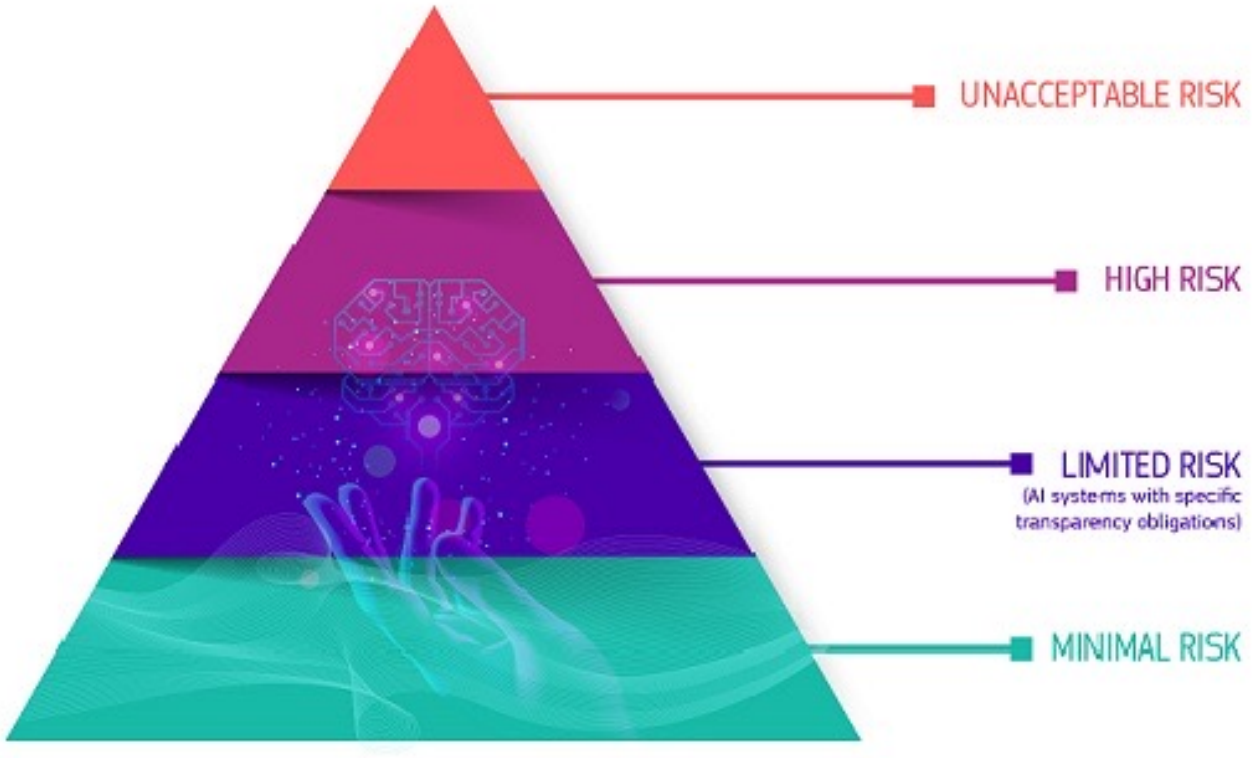
(see Cambridge Dictionary, 2022 in Lütge et al. 2022).

**Algorithm-related risks for stakeholders**
❏ Access to goods, benefits, or services ❏ Financial ❏ Property/material resources ❏ Reputation ❏ Emotional ❏ Life/security ❏ Privacy ❏ Liberty ❏ Rights/intellectual property

(see City and County of San Francisco's Ethics and Algorithms Toolkit)

# The risk-based approach

## of the EU AI regulation



Source: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

# High-risk application areas for AI

## according to the EU AI Act

- critical **infrastructures;**
- **educational** or vocational training;
- **safety** components of products;
- **employment,** management of workers and access to self-employment;
- essential private and **public services;**
- **law enforcement** that may interfere with people's fundamental rights;
- **migration,** asylum and border control management;
- administration of **justice** and democratic processes.

Source: Annex III. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

# Prohibited application areas for AI

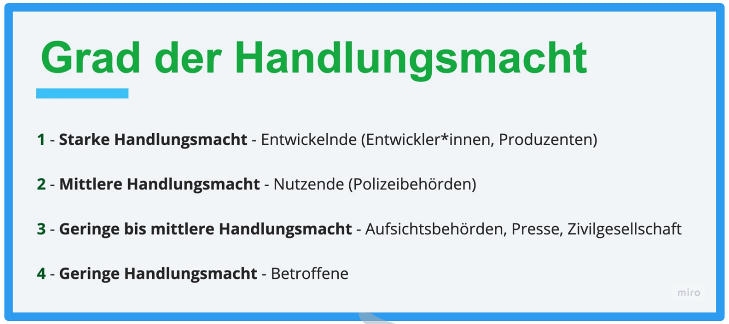
## according to the EU AI Act

- **Social scoring**;
- **Exploitation of vulnerabilities** of persons, use of subliminal techniques;
- **Real-time remote biometric identification** in publicly accessible spaces by law enforcement, subject to narrow exceptions;
- **Biometric categorisation of natural persons** based on biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs or sexual orientation.
- Individual **predictive policing;**
- **Emotion recognition** in the workplace and education institutions;
- Untargeted **scraping of internet or CCTV for facial images** to build-up or expand databases.

Source: https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683

# Tool

# Stakeholder and risk mapping



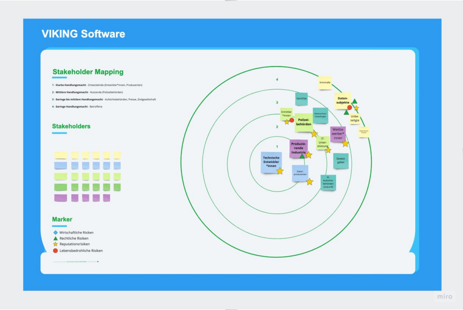
## Grad der Handlungsmacht

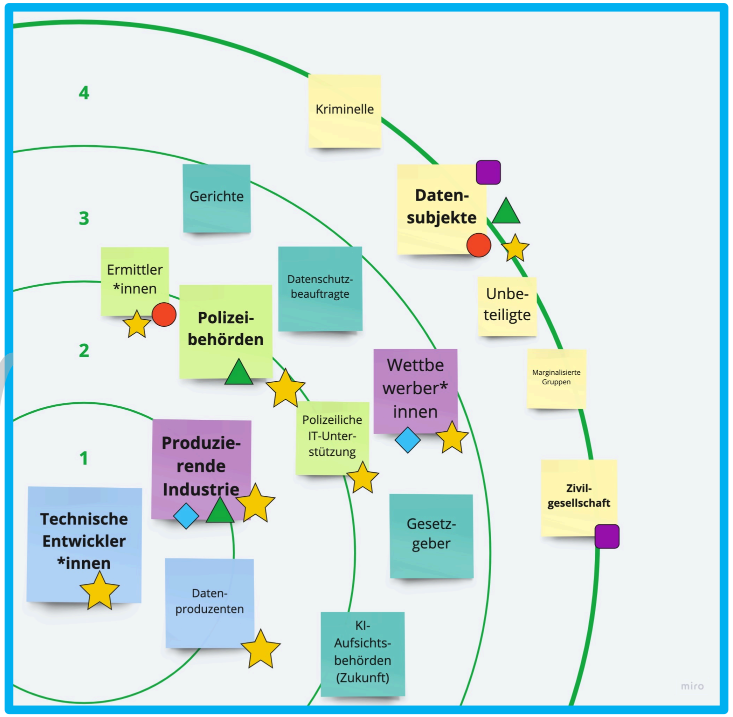
1 - **Starke Handlungsmacht** - Entwickelnde (Entwickler*innen, Produzenten)

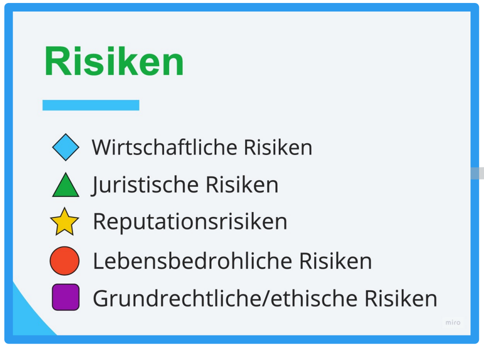
2 - **Mittlere Handlungsmacht** - Nutzende (Polizeibehörden)

3 - **Geringe bis mittlere Handlungsmacht** - Aufsichtsbehörden, Presse, Zivilgesellschaft

4 - **Geringe Handlungsmacht** - Betroffene

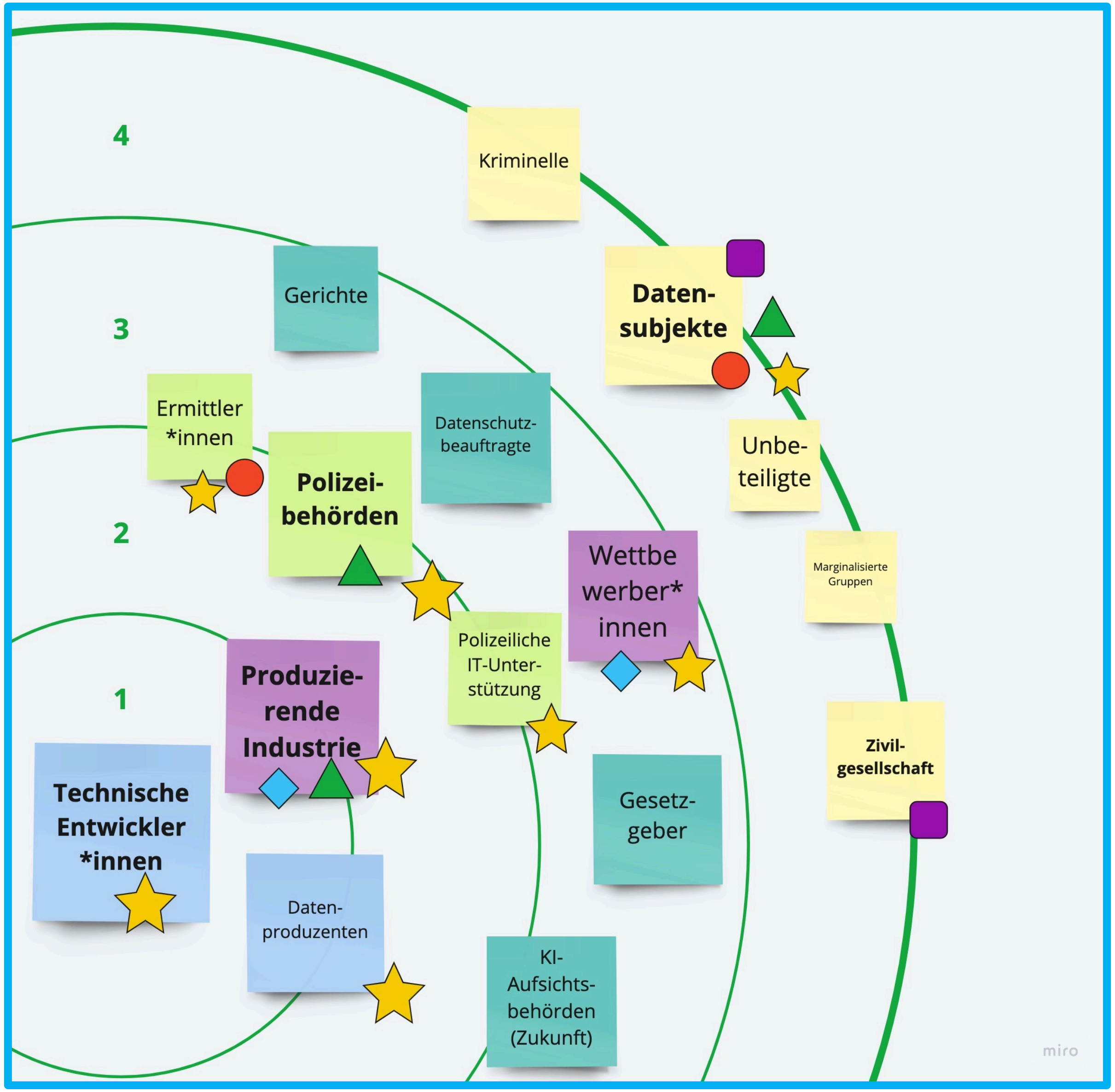






## Risiken

- ◆ Wirtschaftliche Risiken
- ▲ Juristische Risiken
- ☆ Reputationsrisiken
- ● Lebensbedrohliche Risiken
- ■ Grundrechtliche/ethische Risiken



4

3

2

1

Kriminelle

Gerichte

Daten-subjekte

Ermittler *innen

Polizei-behörden

Datenschutz-beauftragte

Unbe-teiligte

Marginalisierte Gruppen

Wettbewerber* innen

Polizeiliche IT-Unter-stützung

Produzie-rende Industrie

Technische Entwickler *innen

Daten-produzenten

Zivil-gesellschaft

Gesetz-geber

KI-Aufsichts-behörden (Zukunft)

# Risks for different stakeholder groups

In the event of faulty or otherwise problematic functioning,

**Technical developers**

- can lose their scientific / professional reputation;

**Companies** offering or selling the software or service,

- may be banned from a market;
- be sued and lose money and reputation.

**Users**

- make mistakes that result in lost time, personal or material damage;
- get sued or fired for using the software (incorrectly).

**Data subjects**

- are wrongly suspected of having committed a crime or intending to commit a crime in the future;
- can be discriminated;
- be subjected to threats and violence;
- be unjustly persecuted, arrested, injured or killed.

**Institutional actors**

- lose confidence in AI technology and stop supporting it (e.g., governments, trade associations, corporations, private and institutional funders, mass media);

(Civil) **society** actors are threatened in their exercise of fundamental rights.

# Input and use case by Alejandro Sierra-Múnera (HPI): an AI-based system for art work image captioning

# 2 Exercice

- Split into small groups;
- Consider a scenario in which the software presented just before is repurposed and deployed for classification tasks in a predictive policing scenario;
- Think about and visualize the stakeholders and their risk exposure.

# Student presentation and discussion

Godulla A, Hoffmann C and Seibert D (2021) Dealing with deepfakes - An interdisciplinary examination of the state of research and implications for communication studies. Studies in Communication and Media.

# Take note: assignment for 10.06

- Develop an abstract (300 words) for your seminar paper, incl.
  - Specify author (one or group of two)
  - Motivation
  - Use case (technology and application field)
  - Main topic (ethical issue)
  - Methods (e.g. stakeholder and risk mapping, scenario techniques)

- Send the abstract to the lecturer until the evening of 10.06.24

Your abstract will be shared with another student who will develop a critical review of your ideas. The lecturer will also comment on the abstracts and we will discuss and develop them further in the last seminar session on 26.6.

# See you on 05.06.!

simon.hirsbrunner@uni-tuebingen.de

**Reading**

Lyons H, Velloso E and Miller T (2021) Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1): 1–25. https://dl.acm.org/doi/abs/10.1145/3449180?casa_token=Bq1MV-xhxzYAAAAA:7LcnY3lzS_36vFSnDmwaNlPyMLEbnnre_tEBNrxtDnfhwmIVWBH-gmt5WX_BBLPpaXwQyRrOmWEF

# Sources

The entire bibilography for the course can be found on Github here: https://github.com/simonsimson/responsible-data-science/blob/main/slides/Bibliography-of-the-entire-course.pdf

**Image sources**
Most sources are cited on the relevant slide. Slide 1: © Adobe Stock / kras99