# ChatGPT war gestern
# Vertrauenswürdige KI ist die Zukunft

KNOW-CENTER GMBH
RESEARCH CENTER FOR DATA-DRIVEN BUSINESS
& BIG DATA ANALYTICS, GRAZ


DI HERMANN STERN
Business Area Manager – Digital Transformation Design

hstern@know-center.at
+43 (0)664 887 83 114
https://www.linkedin.com/in/hermann-stern



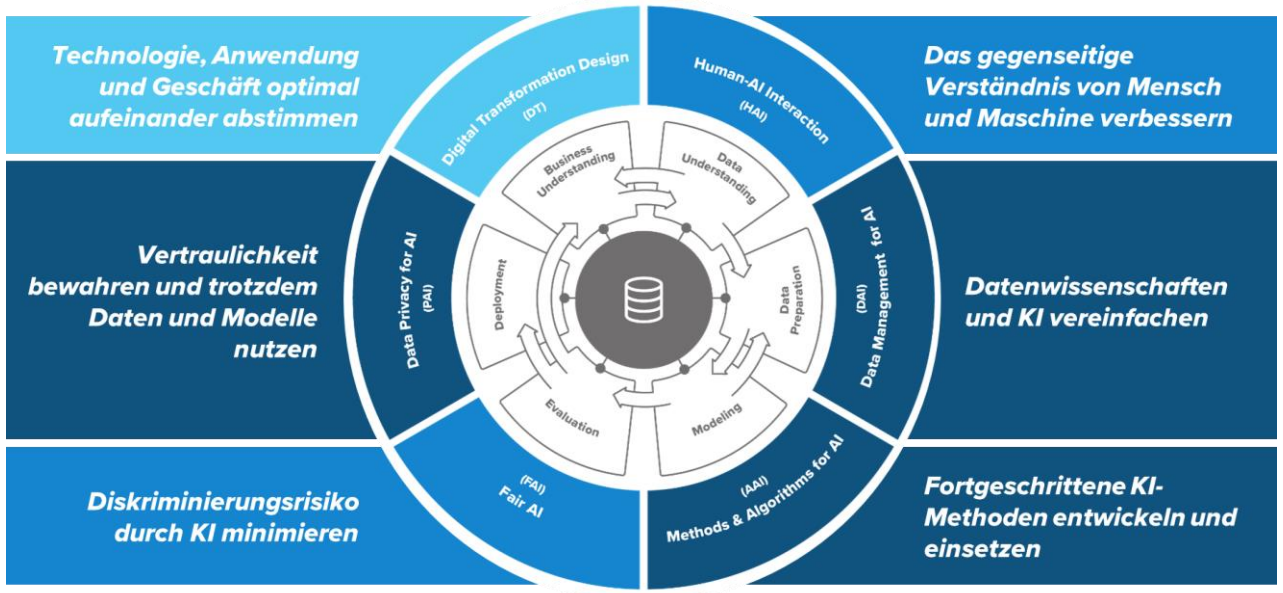
Hermann Stern, Roman Kern


AI & Medical Software in Healthcare
Regulatory Konferenz für Medizinprodukte und In-vitro Diagnostika
Wien, 17.10.2023

## Know-Center GmbH

- K1 COMET-Zentrum
- Gegründet 2001
- 140+ Mitarbeiter:innen
- TU Graz, Campus Inffeldgasse

*„Wir forschen, entwickeln und beraten entlang der Datenwertschöpfungskette zum Thema vertrauenswürdige KI und Data Science.“*



# DEFINITION VON KÜNSTLICHER INTELLIGENZ

# DEFINITION VON KÜNSTLICHER INTELLIGENZ

"Eine technische Lösung, die Menschen (in ihrer Arbeit) unterstützt"

*[…] Künstliche Intelligenz ist die Fähigkeit einer Maschine, menschliche Fähigkeiten wie logisches Denken, Lernen, Planen und Kreativität zu imitieren.*

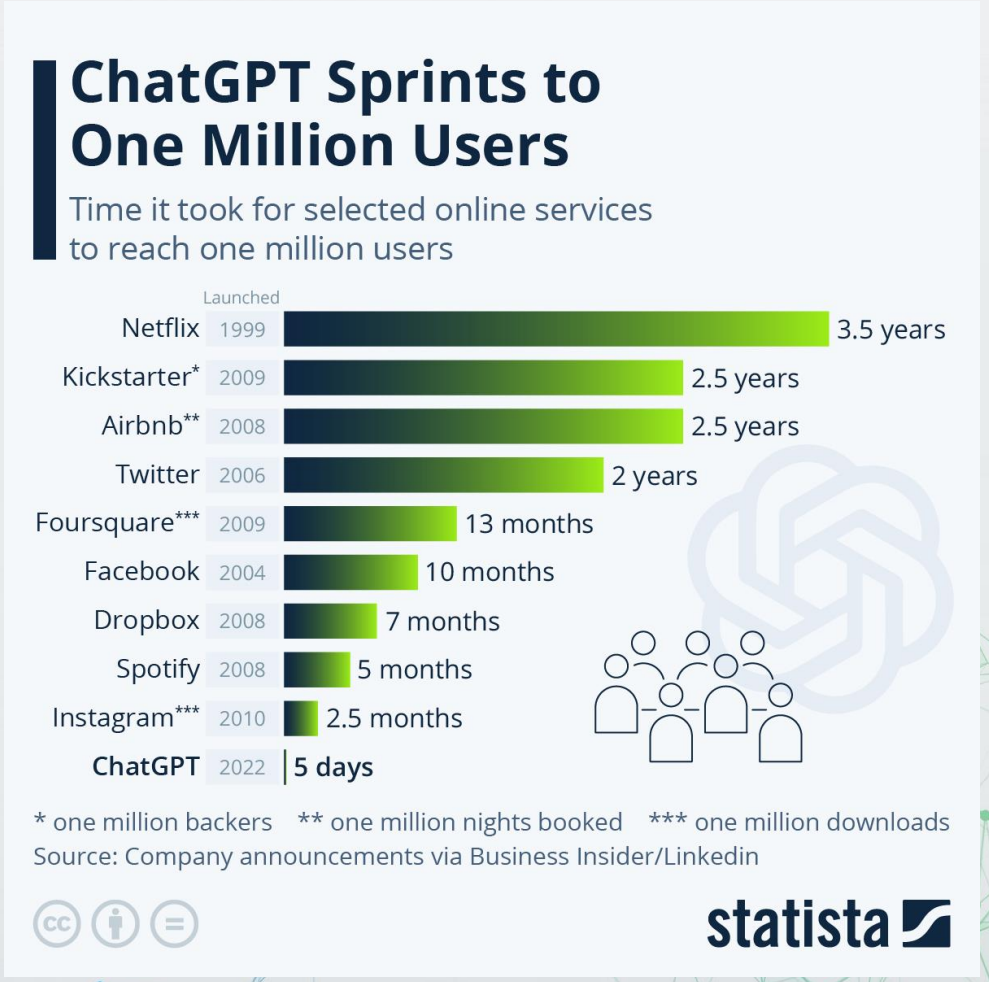
*KI ermöglicht es technischen Systemen, ihre Umwelt wahrzunehmen, mit dem Wahrgenommenen umzugehen und Probleme zu lösen, um ein bestimmtes Ziel zu erreichen. […]*

Quelle: https://www.europarl.europa.eu/news/de/headlines/society/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt

# DEFINITION VON KÜNSTLICHER INTELLIGENZ

**# KI ist gekommen, um zu bleiben!**

# ChatGPT – die KI funktioniert (plötzlich)!? - Bericht aus der Wissenschaft




- Fortschritte im Bereich **Machine Learning und Natural Language Processing („NLP“)**
- Immer mehr Aufgaben kann die Maschine **auf dem Niveau von Menschen** lösen
- ChatGPT hat sich angekündigt

# DEFINITION VON GENERATIVER KÜNSTLICHER INTELLIGENZ

*"Generative KI ist eine Art von künstlicher Intelligenz, die auf der Grundlage von Eingabedaten und Algorithmen **neue Outputs** wie Bilder, Musik oder Text erzeugt. Das Ziel der generativen KI ist es, Outputs zu erzeugen, **die realen Beispielen der gleichen Kategorie ähneln**."*

(ChatGPT, 04.02.2023)

# IST DIE (GENERATIVE) KI REIF FÜR DEN UNEINGESCHRÄNKTEN PRAKTISCHEN EINSATZ?

# IST DIE (GENERATIVE) KI REIF FÜR DEN UNEINGESCHRÄNKTEN PRAKTISCHEN EINSATZ?

**"Nein, die KI ist nicht vertrauenswürdig!"**

- Durchaus gemeint: das Vertrauen der Menschen (Benutzer:innen), vielleicht aber auch über Wertschöpfungsketten und Unternehmensgrenzen hinweg?
- Sind die Ergebnisse denn korrekt? fair? ethisch? sicher? transparent?
- Rechtlich einwandfrei verwendbar?
- Am Weg: Zertifiziert?

# Dialog mit GPT-3

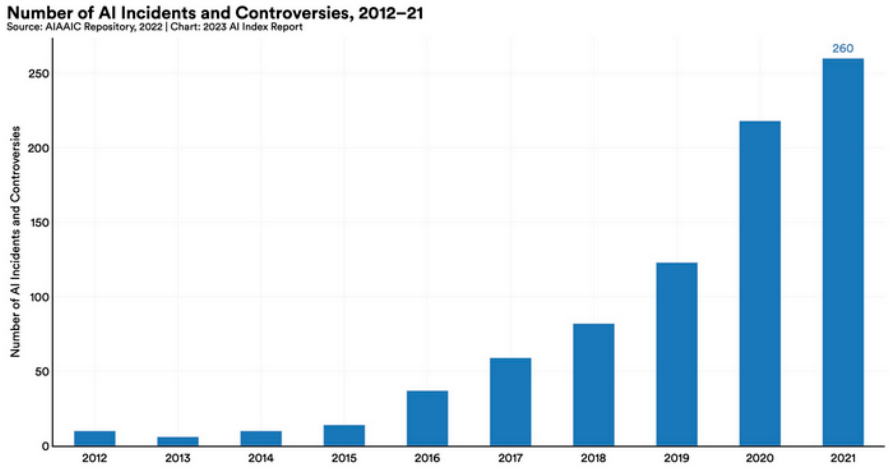
- Was ist die Nationalhymne der Schweiz?

# Dialog mit GPT-3

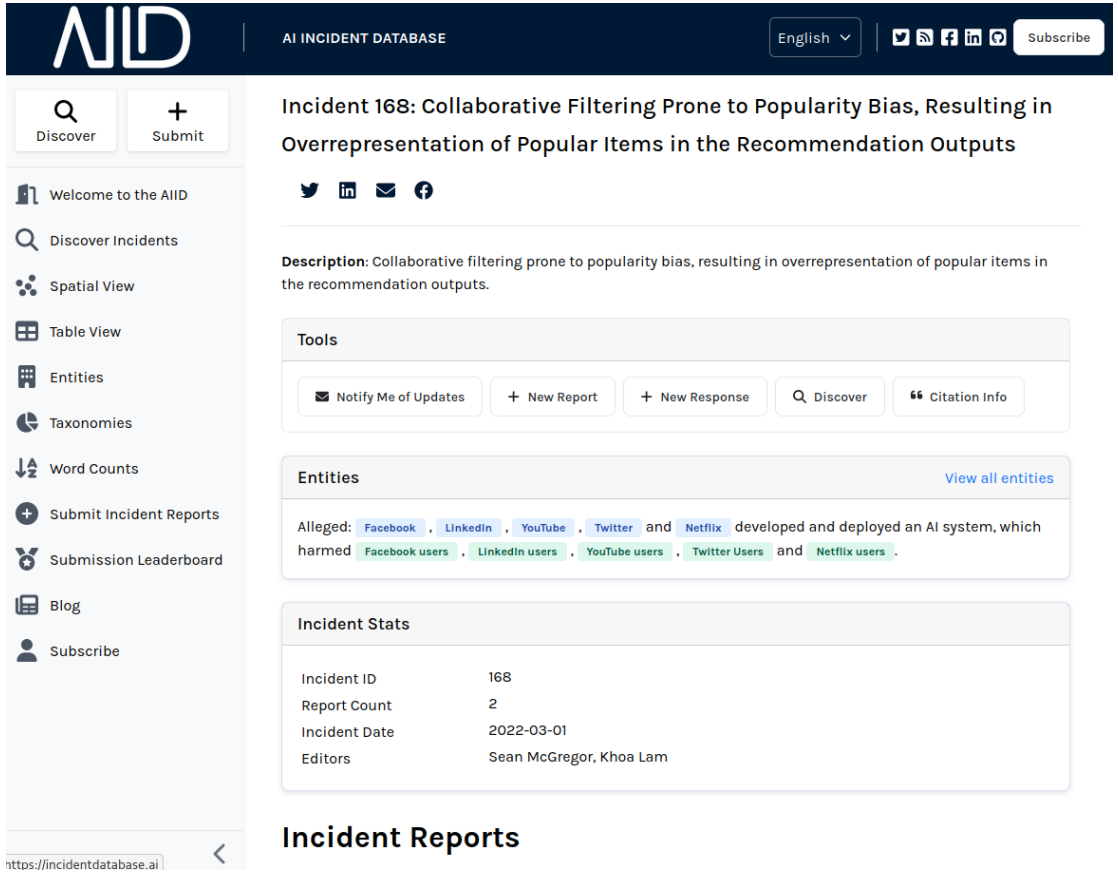
- Was ist die Nationalhymne der Schweiz?
- "Schweiz, Schweiz über alles!“
- ChatGPT “Halluziniert!”


* Originaldialog auf Englisch, **funktioniert nicht mehr ;)**

# Auswirkungen durch den Einsatz der KI

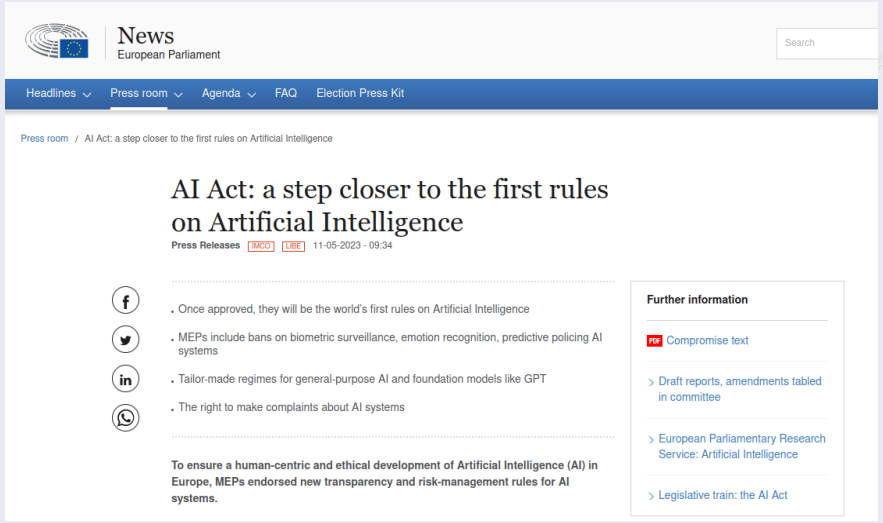



According to the AIAAIC database, which tracks incidents related to the ethical misuse of AI, the number of AI incidents and controversies has increased 26 times since 2012. Some notable incidents in 2022 included a deepfake video of Ukrainian President Volodymyr Zelenskyy surrendering and U.S. prisons using call-monitoring technology on their inmates. This growth is evidence of both greater use of AI technologies and awareness of misuse possibilities.

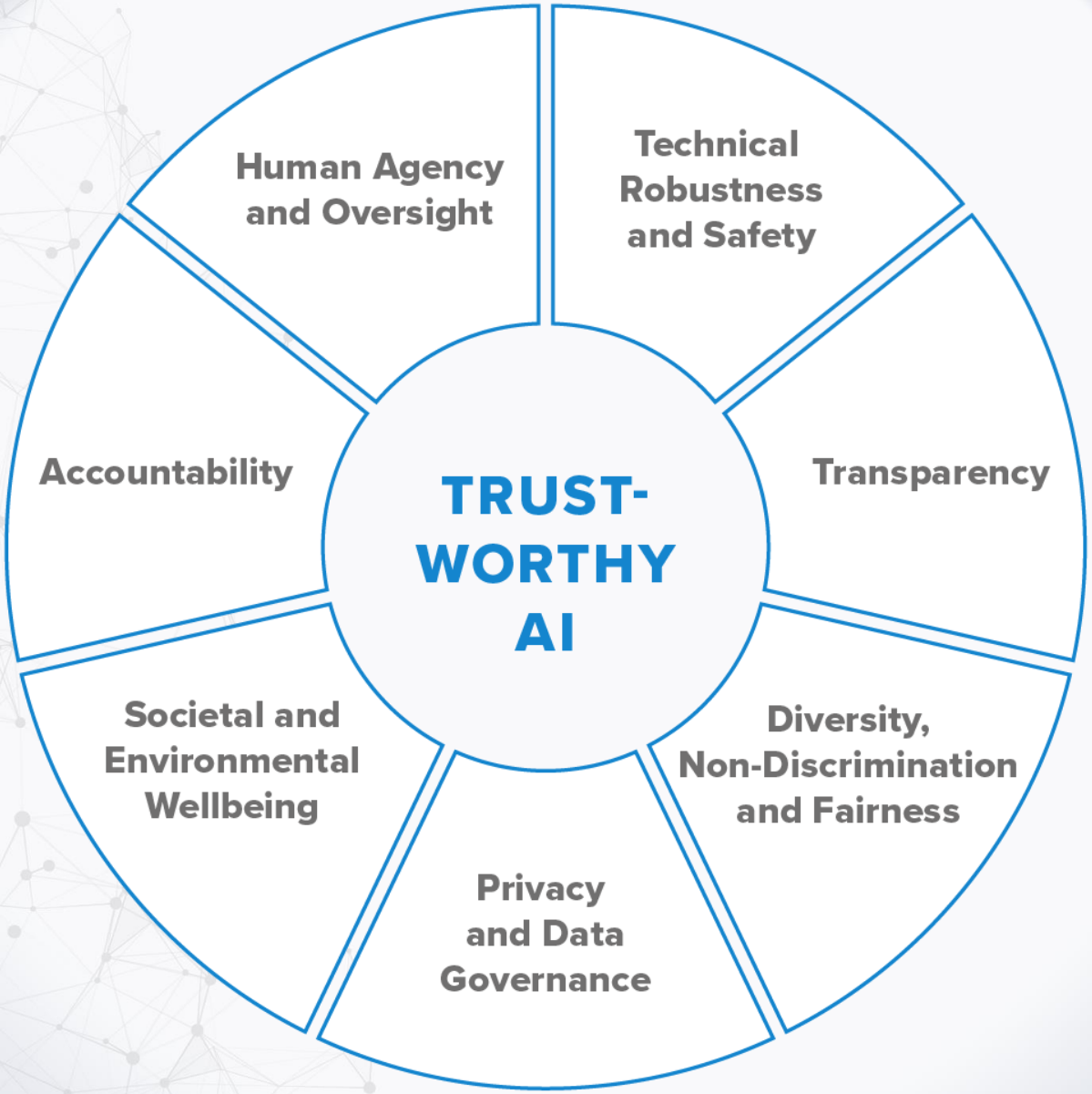


# WIE NUN “VERTRAUEN” SCHAFFEN?

# EU INITIATIVEN

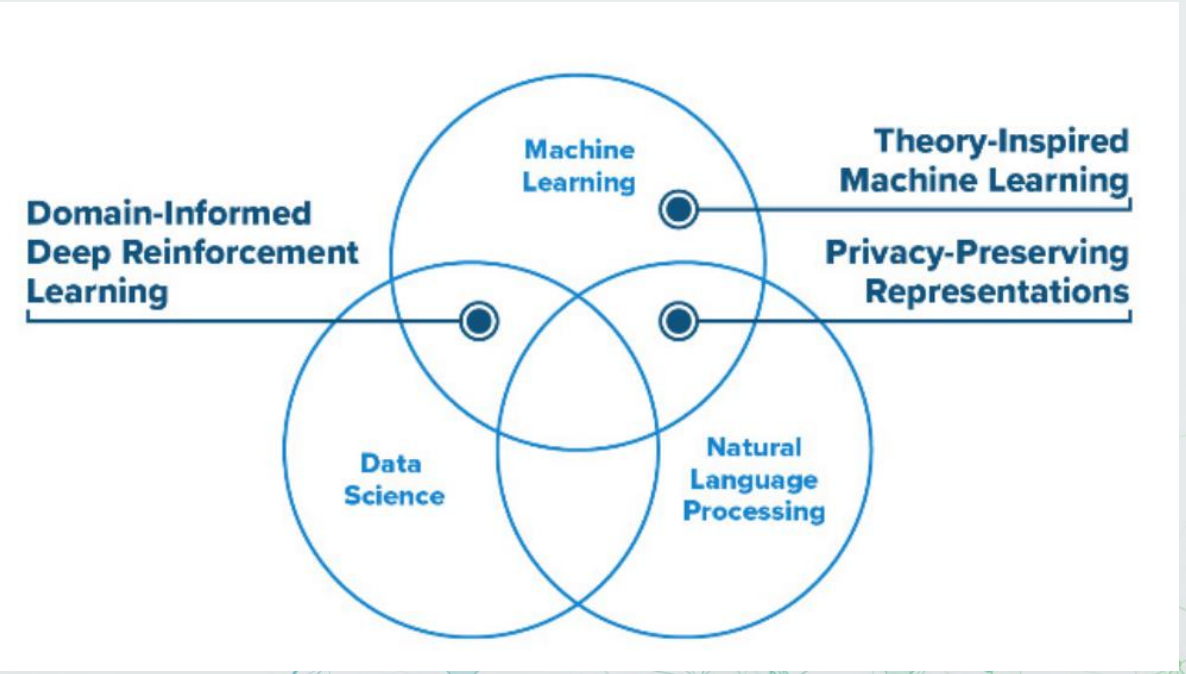


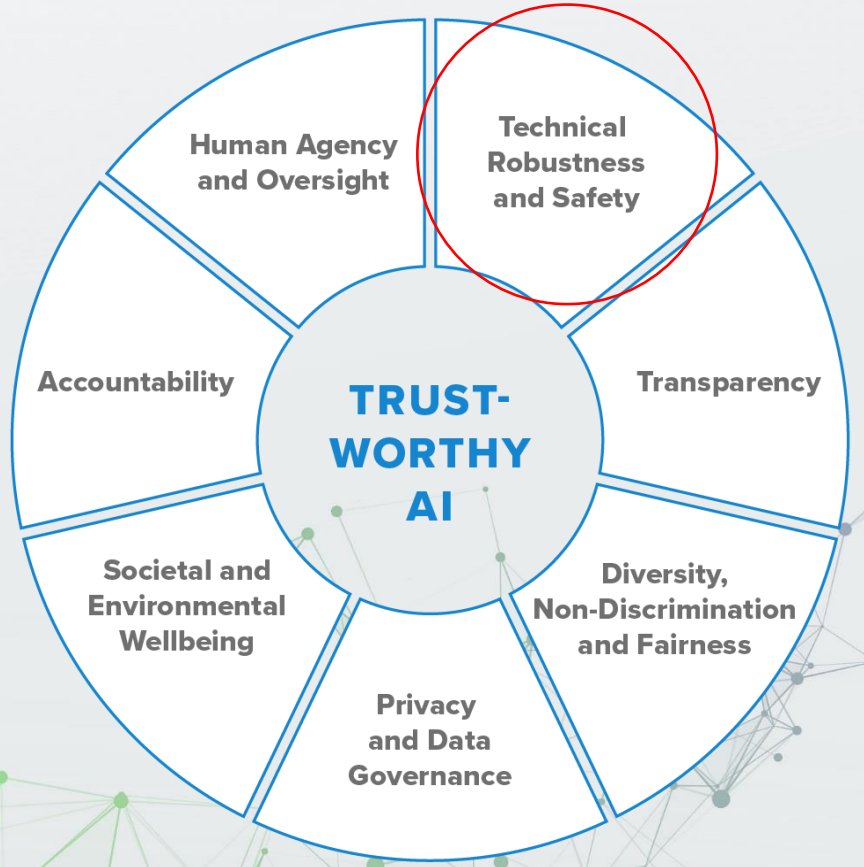

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai



# KI „korrekt“ machen - Hallucination und Strategien in der Forschung

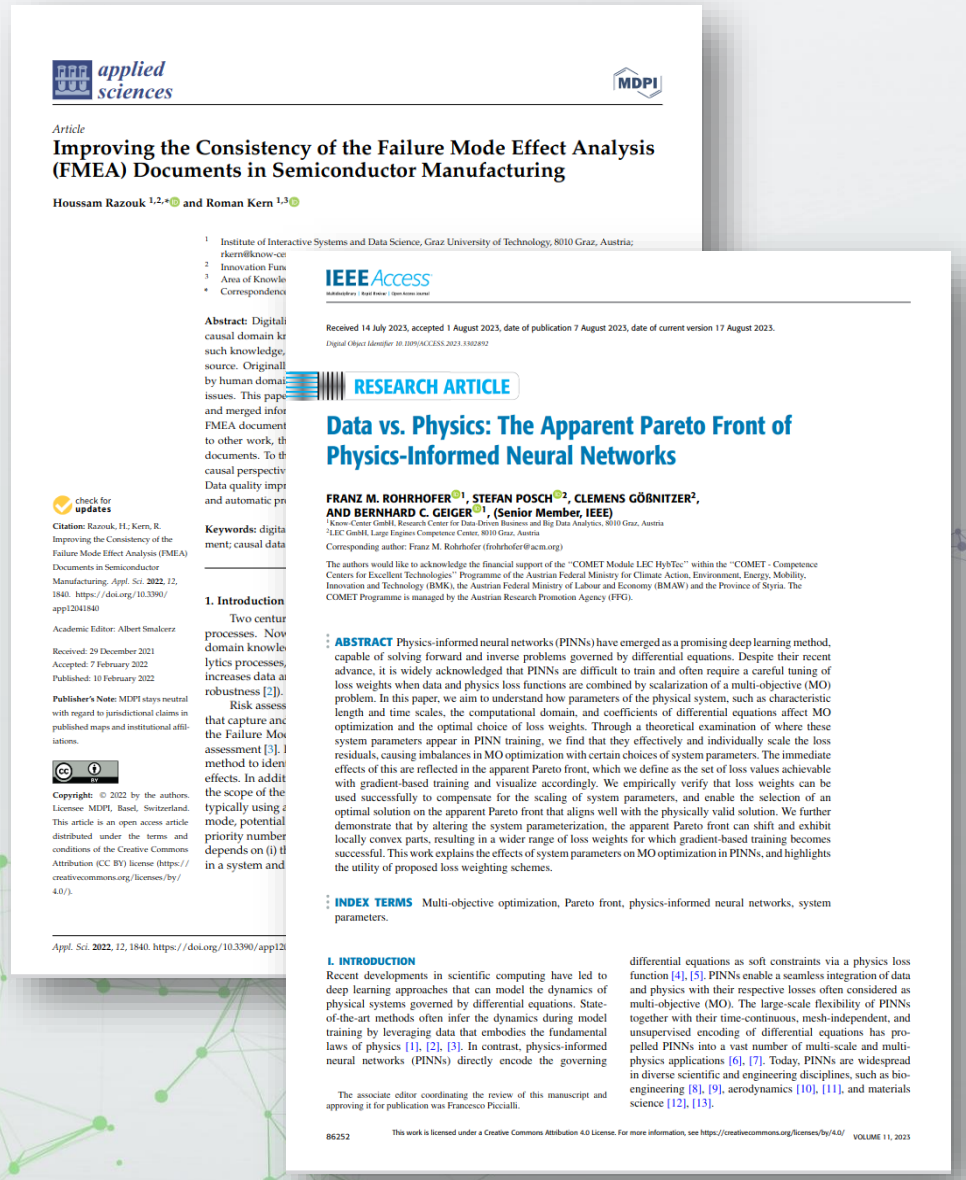
- **Grundlagenforschung**
  - Der Maschine das Denken beibringen



# Hallucination und Strategien in der Forschung

- **Angewandte Forschung**
  - **Vorhandenes Wissen** nutzen
    - … beispielsweise **FMEA** und **technische Reports**
      - via Knowledge Graphs
    - … **Domänenwissen** in die Modellierung miteinbeziehen
      - via PINNs (wissenschaftliche Modelle und physikalische Gesetze in neuronale Netzwerke einbeziehen)



applied sciences

MDPI

Article

Improving the Consistency of the Failure Mode Effect Analysis (FMEA) Documents in Semiconductor Manufacturing

Houssam Razouk [1,2,*] and Roman Kern [1,3]

IEEE Access

Received 14 July 2023, accepted 1 August 2023, date of publication 7 August 2023, date of current version 17 August 2023.

RESEARCH ARTICLE

Data vs. Physics: The Apparent Pareto Front of Physics-Informed Neural Networks

FRANZ M. ROHRHOFER, STEFAN POSCH, CLEMENS GÖßNITZER, AND BERNHARD C. GEIGER, (Senior Member, IEEE)

ABSTRACT Physics-informed neural networks (PINNs) have emerged as a promising deep learning method, capable of solving forward and inverse problems governed by differential equations. Despite their recent advance, it is widely acknowledged that PINNs are difficult to train and often require a careful tuning of loss weights when data and physics loss functions are combined by scalarization of a multi-objective (MO) problem. In this paper, we aim to understand how parameters of the physical system, such as characteristic length and time scales, the computational domain, and coefficients of differential equations affect MO optimization and the optimal choice of loss weights. Through a theoretical examination of where these system parameters appear in PINN training, we find that they effectively and individually scale the loss residuals, causing imbalances in MO optimization with certain choices of system parameters. The immediate effects of this are reflected in the apparent Pareto front, which we define as the set of loss values achievable with gradient-based training and visualize accordingly. We empirically verify that loss weights can be used successfully to compensate for the scaling of system parameters, and enable the selection of an optimal solution on the apparent Pareto front that aligns well with the physically valid solution. We further demonstrate that by altering the system parameterization, the apparent Pareto front can shift and exhibit locally convex parts, resulting in a wider range of loss weights for which gradient-based training becomes successful. This work explains the effects of system parameters on MO optimization in PINNs, and highlights the utility of proposed loss weighting schemes.

INDEX TERMS Multi-objective optimization, Pareto front, physics-informed neural networks, system parameters.

I. INTRODUCTION

Recent developments in scientific computing have led to deep learning approaches that can model the dynamics of physical systems governed by differential equations. State-of-the-art methods often infer the dynamics during model training by leveraging data that embodies the fundamental laws of physics [1], [2], [3]. In contrast, physics-informed neural networks (PINNs) directly encode the governing differential equations as soft constraints via a physics loss function [4], [5]. PINNs enable a seamless integration of data and physics with their respective losses often considered as multi-objective (MO). The large-scale flexibility of PINNs together with their time-continuous, mesh-independent, and unsupervised encoding of differential equations has propelled PINNs into a vast number of multi-scale and multi-physics applications [6], [7]. Today, PINNs are widespread in diverse scientific and engineering disciplines, such as bio-engineering [8], [9], aerodynamics [10], [11], and materials science [12], [13].

# Hallucination und Strategien in der Forschung

- **Angewandte Forschung**
  - **Tool-Augmented LLMs**
    - Agents, Plugins, Extensions

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

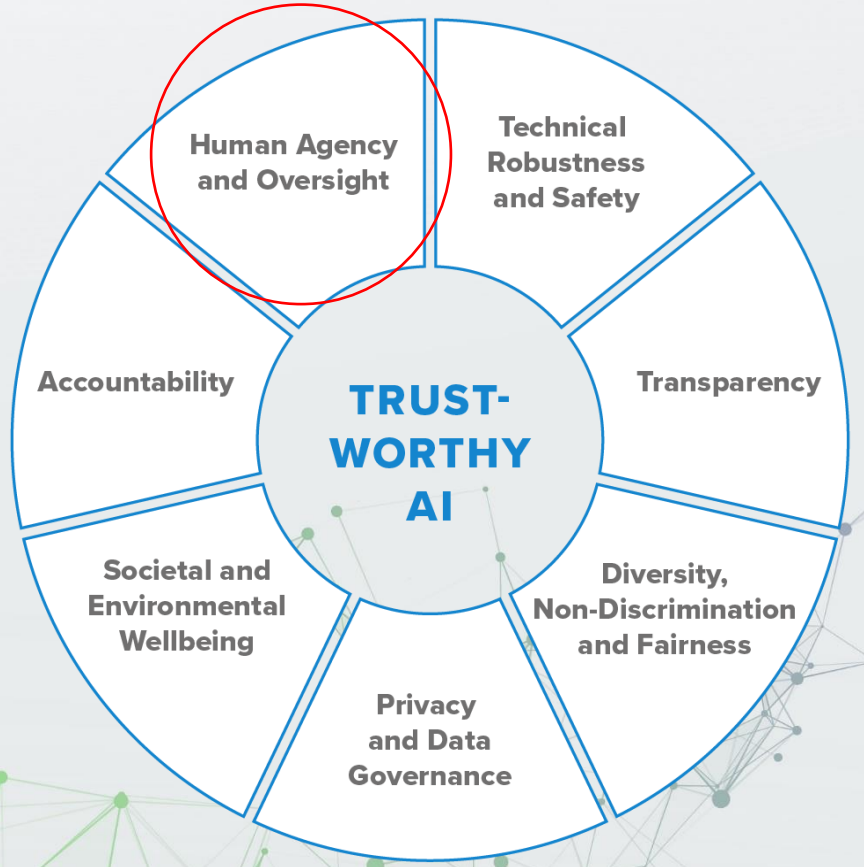
The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

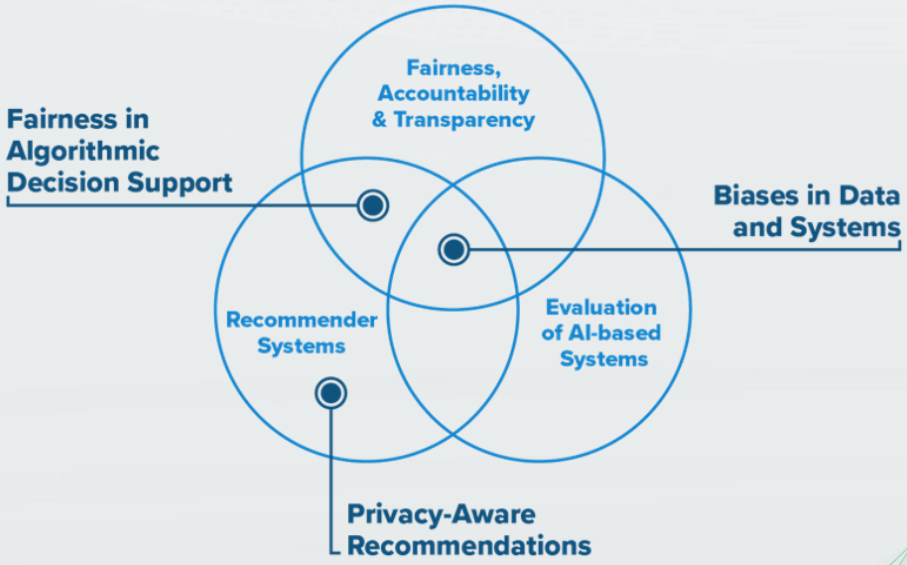
https://arxiv.org/pdf/2302.04761.pdf

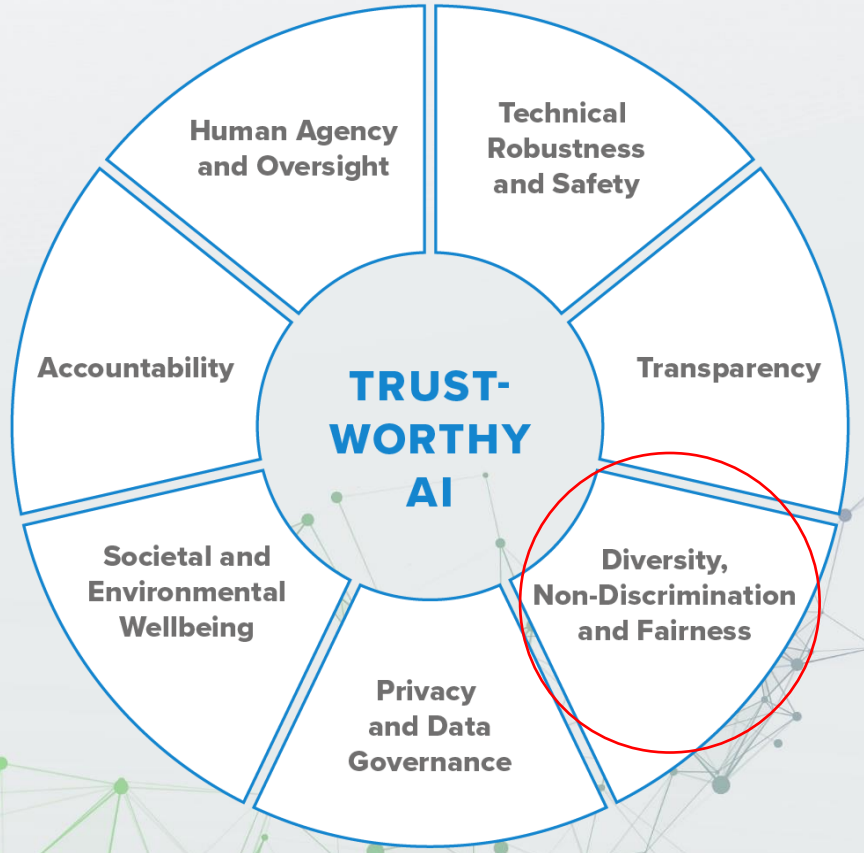
# Wahrnehmung von ChatGPT (KI) ändern

- **In der Gesellschaft**
  - "Terminator"
  - Monopol von Technologieriesen
  - Missinformation, Fake News
- **Im Unternehmen**
  - Angst vor Jobverlust
  - Veränderung des Arbeitsumfelds
  - Unklarheiten der Potentiale



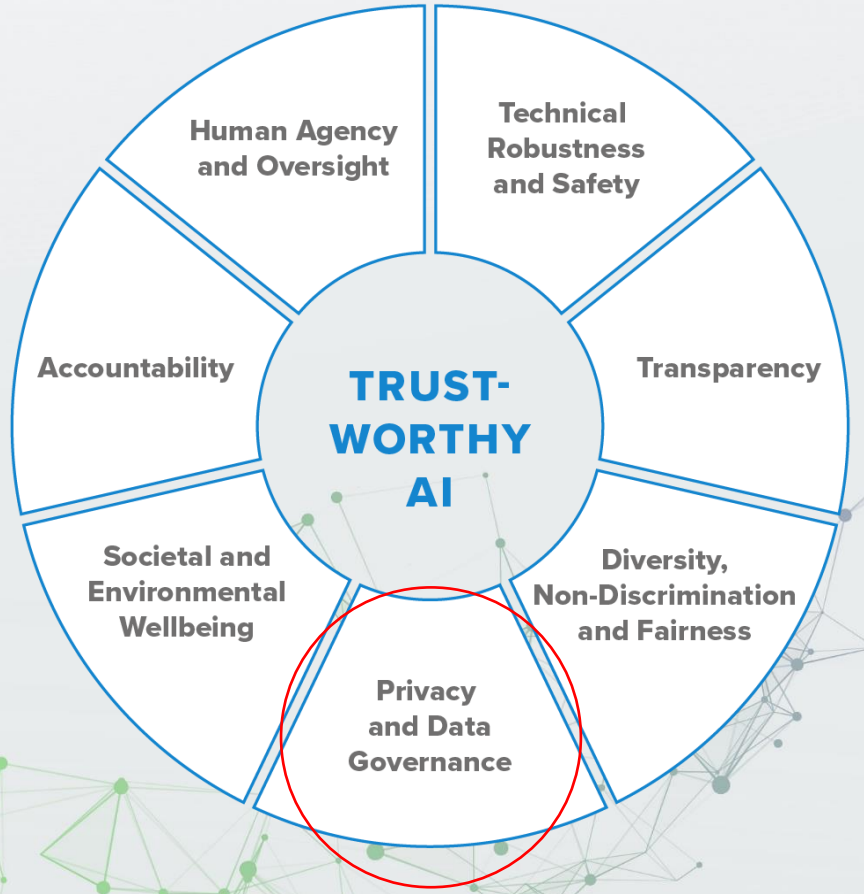
# KI “fair” machen – in Gestaltung, Entwicklung und Betrieb



# KI “sicher” machen

- Austausch von sensiblen und vertraulichen Daten
- Schutz der Privatsphäre
- Ansatz: On-Premises LLMs




Tech Artificial Intelligence

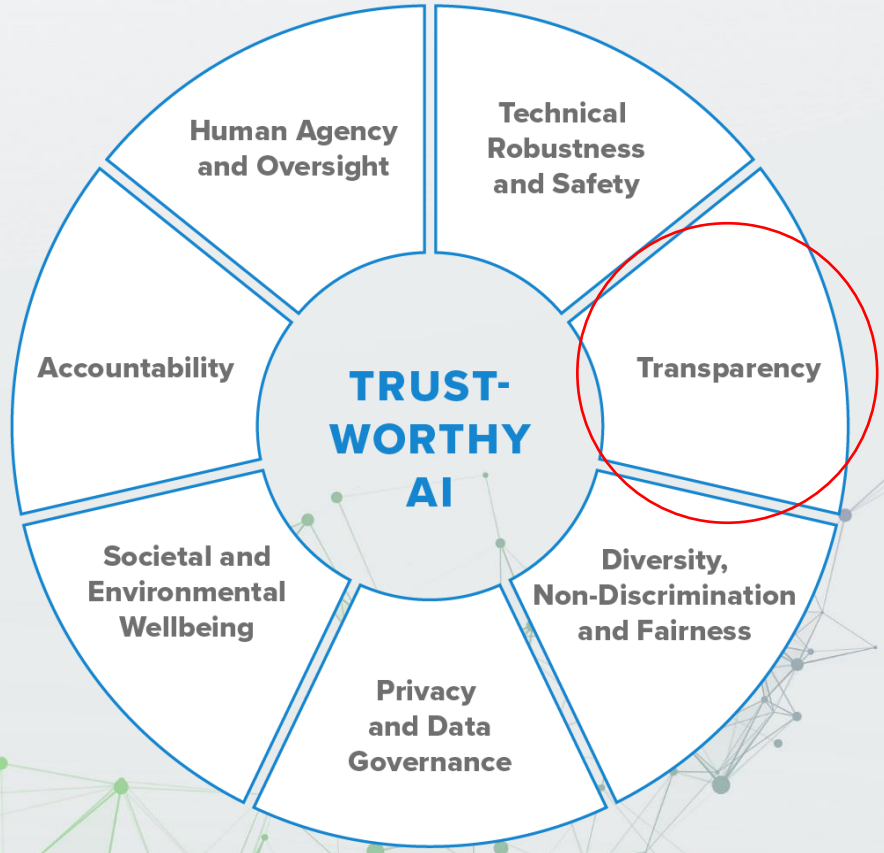
**Whoops, Samsung workers accidentally leaked trade secrets via ChatGPT**

ChatGPT doesn't keep secrets.

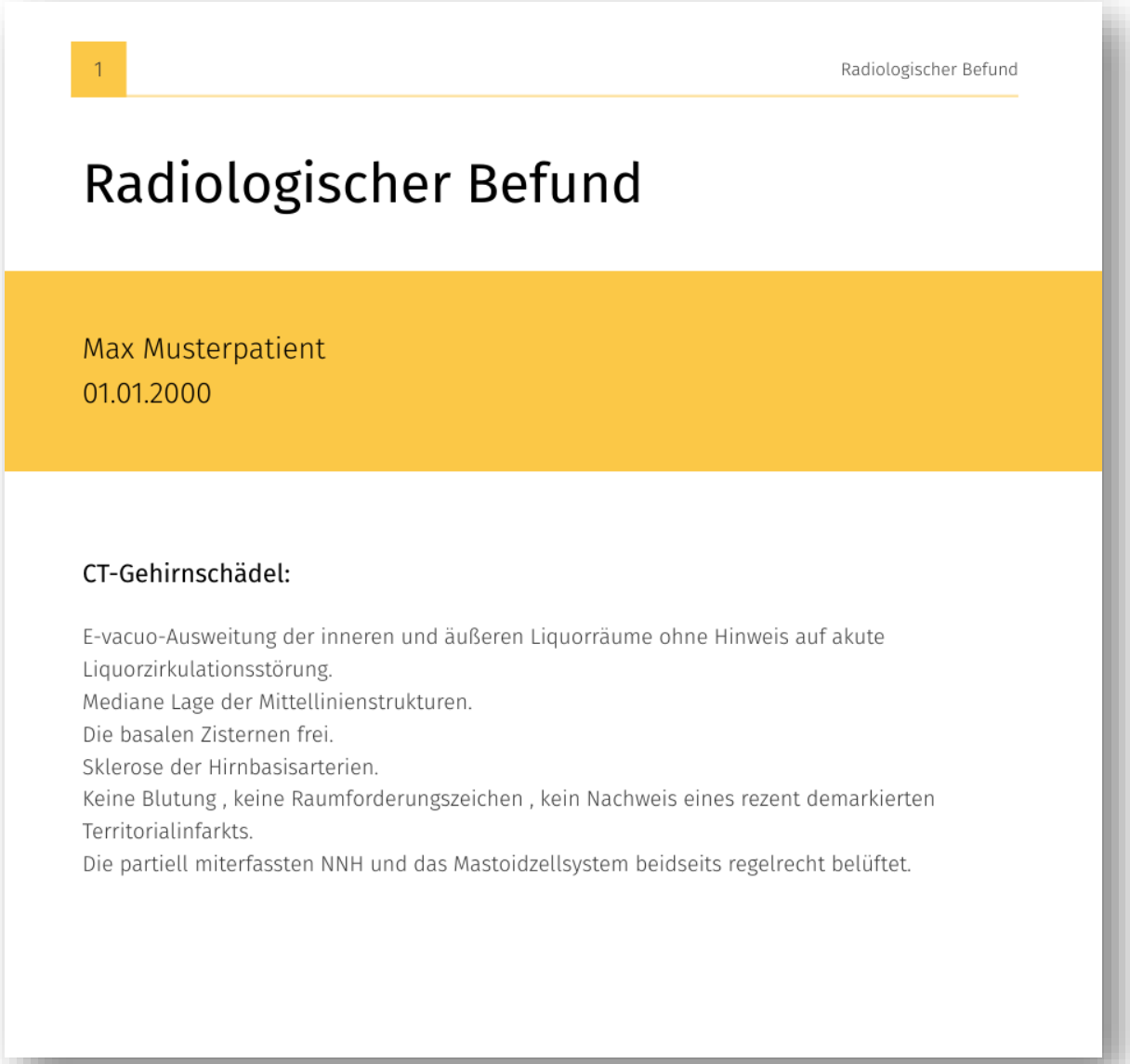
By Cecily Mauran on April 6, 2023

# KI “verständlich” machen

- Schutz KI als **Black Box**
- **eXplainable AI** (XAI) als Forschungsfeld
- Domänenexperten die AI **erklärbar** machen



# Beispiel: simplifAI



1 Radiologischer Befund

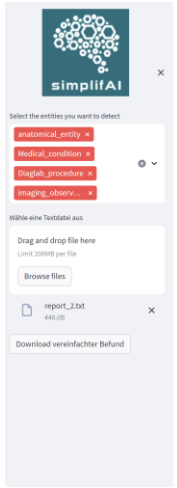
Radiologischer Befund

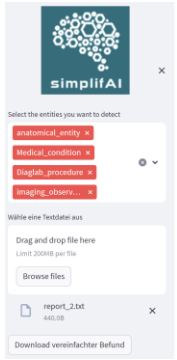
Max Musterpatient
01.01.2000

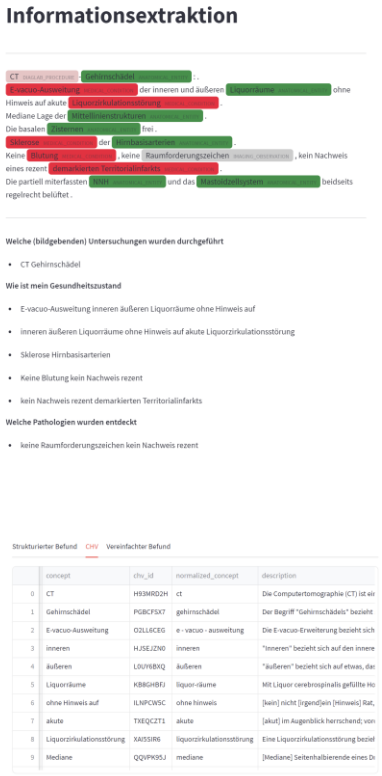
**CT-Gehirnschädel:**

E-vacuo-Ausweitung der inneren und äußeren Liquorräume ohne Hinweis auf akute Liquorzirkulationsstörung.
Mediane Lage der Mittellinienstrukturen.
Die basalen Zisternen frei.
Sklerose der Hirnbasisarterien.
Keine Blutung , keine Raumforderungszeichen , kein Nachweis eines rezent demarkierten Territorialinfarkts.
Die partiell miterfassten NNH und das Mastoidzellsystem beidseits regelrecht belüftet.

# Beispiel: simplifAI






## Radiologischer Befund

Max Musterpatient
01.01.2000

**CT-Gehirnschädel : .**
Die Computertomographie (CT) ist eine Art der Bildgebung, bei der mithilfe von Röntgenstrahlen detaillierte Bilder des Körpers erstellt werden. Im vorliegenden Fall wird der Gehirnschädel untersucht, also der knöcherne Teil des Schädels, der das Gehirn umgibt und schützt.

**E-vacuo-Ausweitung der inneren und äußeren Liquorräume ohne Hinweis auf akute Liquorzirkulationsstörung .**
Die E-vacuo-Ausweitung bezieht sich auf eine Erweiterung der inneren und äußeren Liquorräume ohne Hinweis auf eine akute Liquorzirkulationsstörung. Das bedeutet, dass sich die mit Liquor cerebrospinalis gefüllten Hohlräume im Gehirn und Rückenmark ausgedehnt haben, aber es gibt keine Anzeichen für eine akute Störung des normalen Flusses von Liquor cerebrospinalis.

**Mediane Lage der Mittellinienstrukturen .**
Die medianen Mittellinienstrukturen beziehen sich auf anatomische Strukturen im zentralen Teil des Gehirns, die entlang der mittleren Linie des Gehirns liegen.

# ZUSAMMENFASSUNG

1. KI funktioniert (in vielen Bereichen)
2. KI muss vertrauenswürdig (Trustworthy) werden
3. KI muss den Menschen dienen

# Vielen Dank!


DI HERMANN STERN
Business Area Manager – Digital Transformation Design

hstern@know-center.at
+43 (0)664 887 83 114
https://www.linkedin.com/in/hermann-stern



Hermann Stern, Roman Kern


AI & Medical Software in Healthcare
Regulatory Konferenz für Medizinprodukte und In-vitro Diagnostika
Wien, 17.10.2023