Faire KI – (wie) kann sie funktionieren?

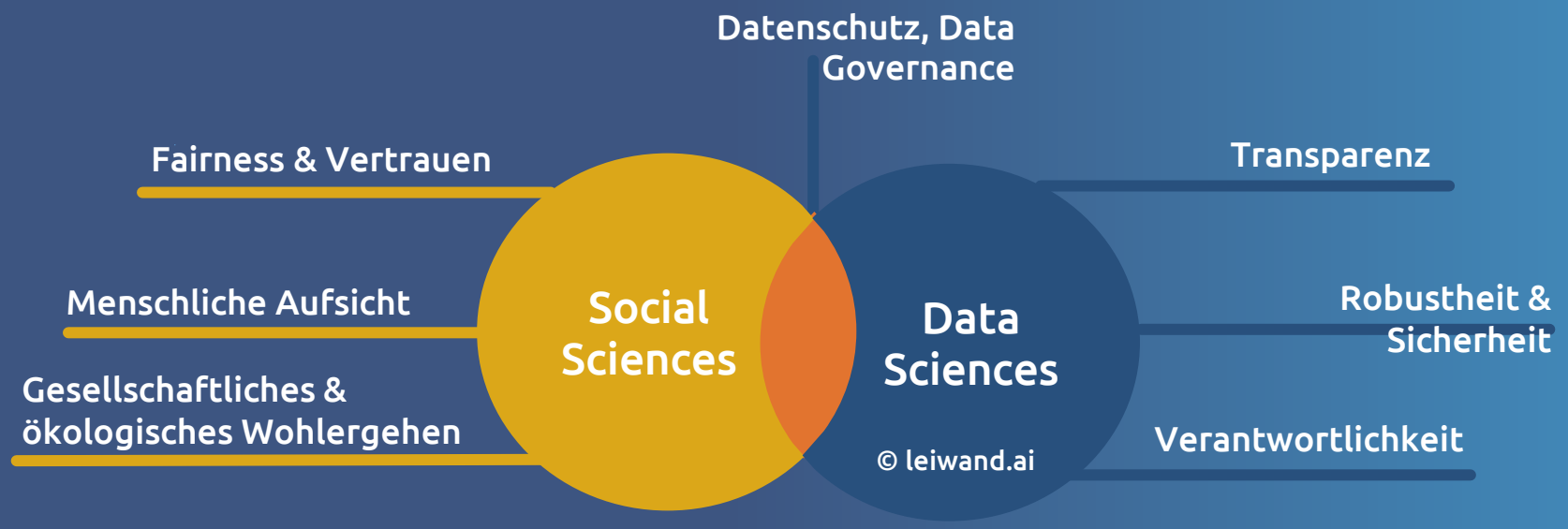
Dr. Gertraud Leimüller, MPA (Harvard)
CEO winnovation consulting gmbh & leiwand.ai

25. Jänner 2024
Laura Bassi 4.0 Netzwerkforum und Abschlusskonferenz: "Digital Fairness“

Die Mission von leiwand.ai

Faire und hochwertige KI zu einer Realität machen

leiwand AI entwickelt Werkzeuge, die Künstliche Intelligenz vertrauenswürdig machen (Trustworthy AI). Dabei nutzen wir unsere Expertise in Innovationsentwicklung (Open Innovation) und Alorithmic Fairness.



Wir machen KI leiwand

Rania Wazir

Expertin für Machine Learning, NLP und Trustworthy AI, Schwerpunkt auf algorithmischer Fairness; führende Rolle in internat. Standardisierung

Stanford-Absolventin

Co-founder und CTO

Gertraud Leimüller

Expertin für Open-Innovation-Strategien & Methoden, User-Integration, sozialwissen-schaftl. Expertin für Trustworthy AI

Harvard-Absolventin

Co-founder und CEO

Mira Reisinger
Linguistischer Hintergrund, NLP und ML Expertin

Data Scientist

Patrick Kosmider
philologischer Hintergrund; Redakteur, Multimedia Journalist, Analyst

Communications Manager

Thomas Treml
Sozialwissenschaftlicher Hintergrund, Open-Source-Spezialist, NLP- und ML-Experte

Data Scientist – Head of MLOps

Silvia Wasserbacher-Schwarzer
Hintergrund in Politikwissenschaft und Kommunikation; Expertin für digitale Transformation und Open Innovation

Chief Strategist

Eine kurze Agenda

1. Begriffsklärung
2. Algorithmic Bias betrifft uns alle - Beispiele
3. Ursachen und Quellen von Algorithmic Bias
4. Die neue Rechtslage: EU AI Act und Fundamental Rights Assessment

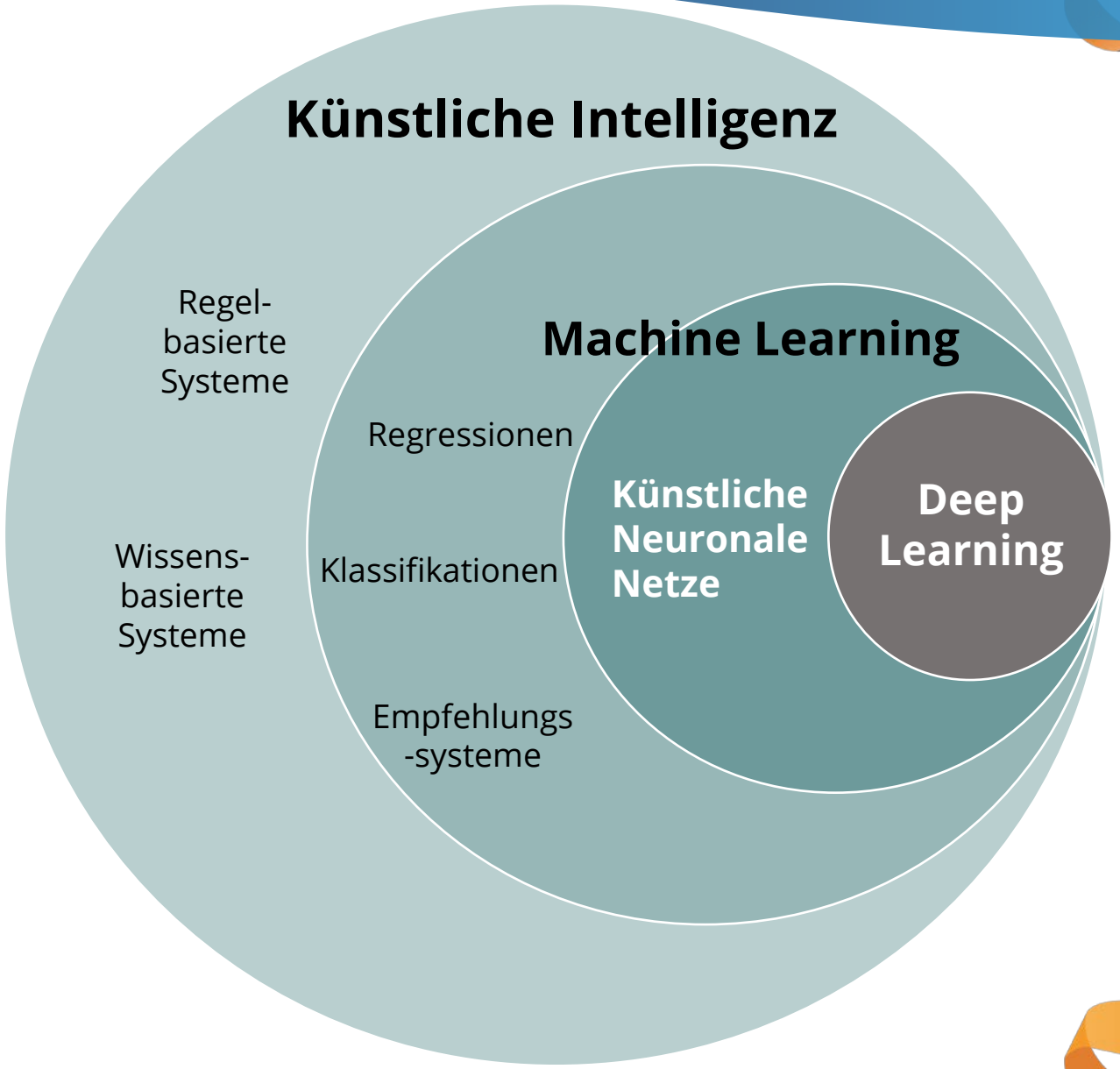
Algorithmic Bias Begriffsklärung

Algorithmischer Bias umfasst:

systematische und wiederholbare Fehler in einem Computersystem, die zu „unfairen“ Ergebnissen führen, wie etwa die „Privilegierung“ einer Kategorie gegenüber einer anderen auf eine Weise, die von der beabsichtigten Funktion des Algorithmus abweicht.

Führt zu Benachteiligung von Menschen im

1. **Zugang zu Ressourcen** (Beurteilung von Prüfungen, Gesundheit, Jobs, Kredit, etc.) - *Ressourcen-Fairness*
2. **Darstellung von Personen** – *Repräsentations-Fairness*
3. **Zugang zu Informationen und Prozessen** – *Prozessuale Fairness*



Algorithmic Bias betrifft uns alle

Beispiel 1: Übersetzungs-AI
Probieren Sie selbst aus!

Übersetzen Sie ins Deutsche:

My doctor is pregnant.

Übersetzen Sie ins Deutsche:

My secretary is on paternity leave.

Das Ergebnis zeigt eingebettete Stereotype

Übersetzen Sie ins Deutsche:

My doctor is pregnant.

Mein Arzt ist schwanger.

Übersetzen Sie ins Deutsche:

My secretary is on paternity leave.

Meine Sekretärin ist im Vaterschaftsurlaub.

Beispiel: Nicht-Muttersprachler werden benachteiligt

AI-Detektoren beurteilen Texte fälschlicherweise als AI-generiert

Sieben AI Detektoren wurden von Forschenden der Universität Stanford getestet. Resultat:

- Für Menschen, die in den USA geboren sind, arbeiten die Detektoren zuverlässig.
- Bei Essays von Schreibenden, deren Muttersprache nicht Englisch ist, ist die Zuverlässigkeit hingegen sehr gering:
  - 61.22% von TOEFL Aufsätzen wurden fälschlicherweise als AI-generiert ausgewiesen.
  - Alle 7 Detektoren besitzen diesen Bias.

⍰ Diskriminierungsgefahr für Menschen, deren Muttersprache nicht Englisch ist – Vorwurf des Schwindelns

Quelle: https://arxiv.org/abs/2304.02819

Beispiel: Halluzinationen führen zu Reputationsschaden

Chat GPT macht aus einem Korruptionsbekämpfer einen Beschuldigten

Brian Hood ist ein australischer Bürgermeister und ein Wistleblower, der gegen Korruption in einem Tochterunternehmen der Australischen Nationalbank gekämpft hat.

- ChatGPT dreht den Spieß um und behauptet fälschlicherweise, er sei wegen Bestechung im Gefängnis gesessen.
- Hood kündigte Klage an.

Quelle: https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/

Bild: unsplash.com, BBC

Beispiel: Falsches Design von Gesundheits-AI benachteiligt Menschen mit dunkler Hautfarbe

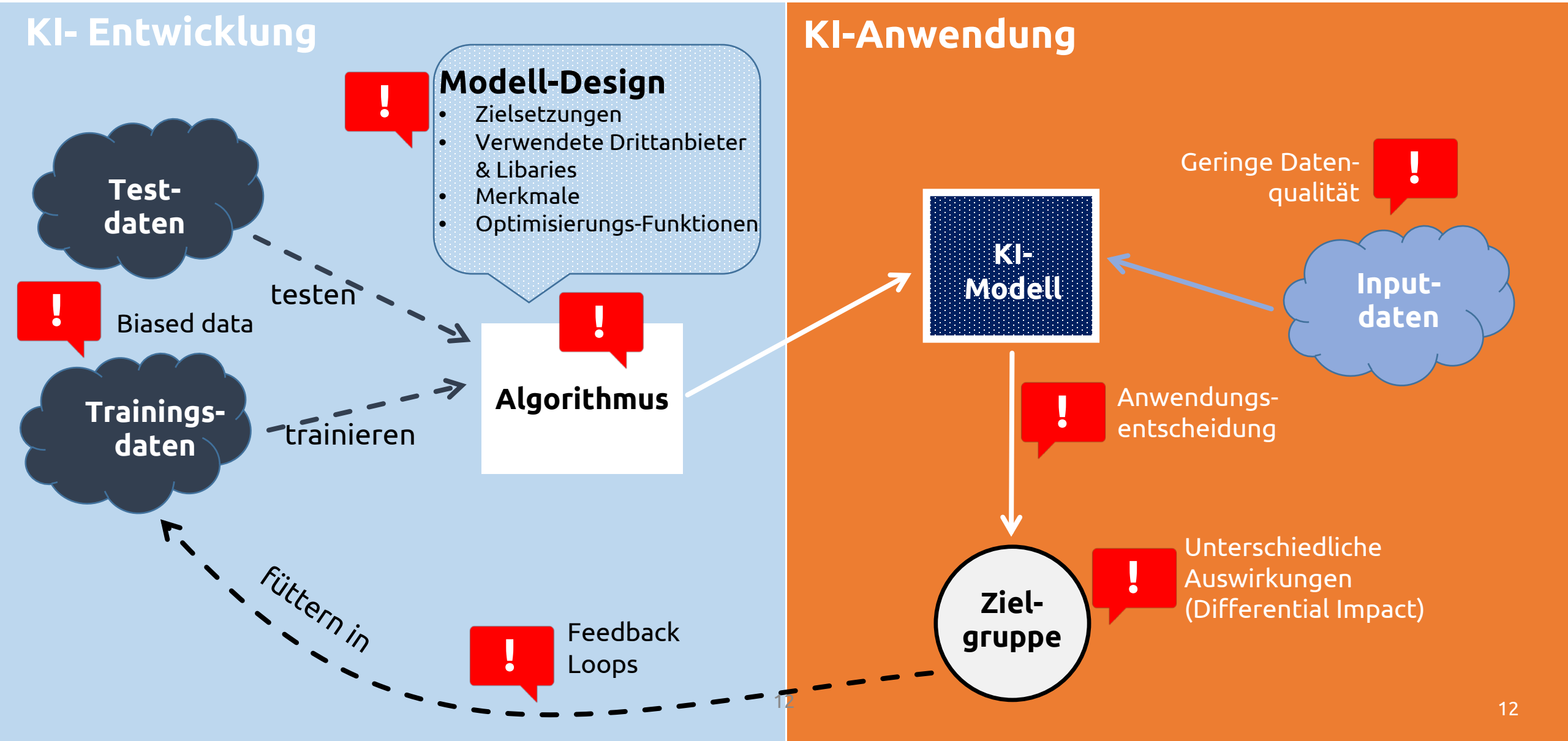
Bild: unsplash.com, CDC

Hochrisiko-Patient:innen bekommen in den USA eine **zusätzliche Versorgung** angeboten.

- Ein Algorithmus benachteiligt Personen mit dunkler Hautfarbe, weil er **bisherige Gesundheitskosten als Proxy für Bedarf an zusätzlicher Gesundheitsversorgung** heranzieht und damit den Bedarf von Unterversorgten unterbewertet.
- Gemessen am tatsächlich Risiko müsste der Anteil von Afro-Amerikaner:innen an Hochrisikopatient:innen 46% statt 18% betragen.

Quelle: Science 2019, Vol 366, Issue 6464, https://www.science.org/doi/10.1126/science.aax2342

Woher kommt algorithmischer Bias?
Es sind nicht nur die Daten ....



Wie können AI Fairness-Probleme vermieden werden? (1)

Wir brauchen einen Paradigmen-Wechsel in der Entwicklung von AI-Modellen: Weg von einer reinen Informatik-Brille, hin zu **interdisziplinären kollaborativen Innovationsprozessen** mit:

1. Data Scientists, Data Engineers mit hoher Gender-Diversität! Nur 16% aller AI-Beschäftigten sind in Europa weiblich.
2. Domänen-Expert*innen (z.B. HR, Medizin, …)
3. potenziellen Nutzer*innen und Betroffenen
4. Sozialwissenschaftliche Expert*innen und
5. Rechtsexpert*innen

Wie können AI Fairness-Probleme vermieden werden? (2)

Sieben zentrale Erkenntnisse aus dem Forschungsprojekt „fAIr by design“: https://www.fairbydesign.eu/

1. **Strategische Zielsetzungen** und „Job to be done“ zu **Projektbeginn** zu Beginn sauber definieren, Problemlösungs-Kapazität von AI kritisch diskutieren
2. **User-zentriertes Vorgehen:** Use-Kontext analysieren, **Fairness in diesem Kontext definieren (Was beurteilen die User als Fairness-Risiko?)**, Risiko-Analyse durchführen – Groups at Risk, Art der Diskriminierungs-Risiken
3. Performance Measurement basierend auf vorherigen Erkenntnissen definieren, **Kombination mehrerer Kennzahlen** für Training und Monitoring verwenden

fAIr by design

Wie können AI Fairness-Probleme vermieden werden? (3)

4. **Rohdaten auf möglichen unerwünschten Bias überprüfen**, bearbeiten und gg.falls ergänzen
5. **Limitationen** von AI-Modellen und Daten verstehen und kommunizieren
6. **Laufend testen**, auch mit unterschiedlichen Testdaten
7. System **entlang des Lebenszyklus** monitoren und weiterentwickeln

fAIr by design

Trustworthyness ist auch ein wirtschaftlicher Vorteil

Bild: upsplash, NASA

Das Marktforschungsunternehmen Gartner erwartet, dass jene Unternehmen, die bis 2026 AI-Transparenz, Vertrauen und Sicherheit operationalisieren,

sich beim Einsatz von AI, dem Erreichen von Businesszielen und der Akzeptanz für AI um 50% verbessern werden.

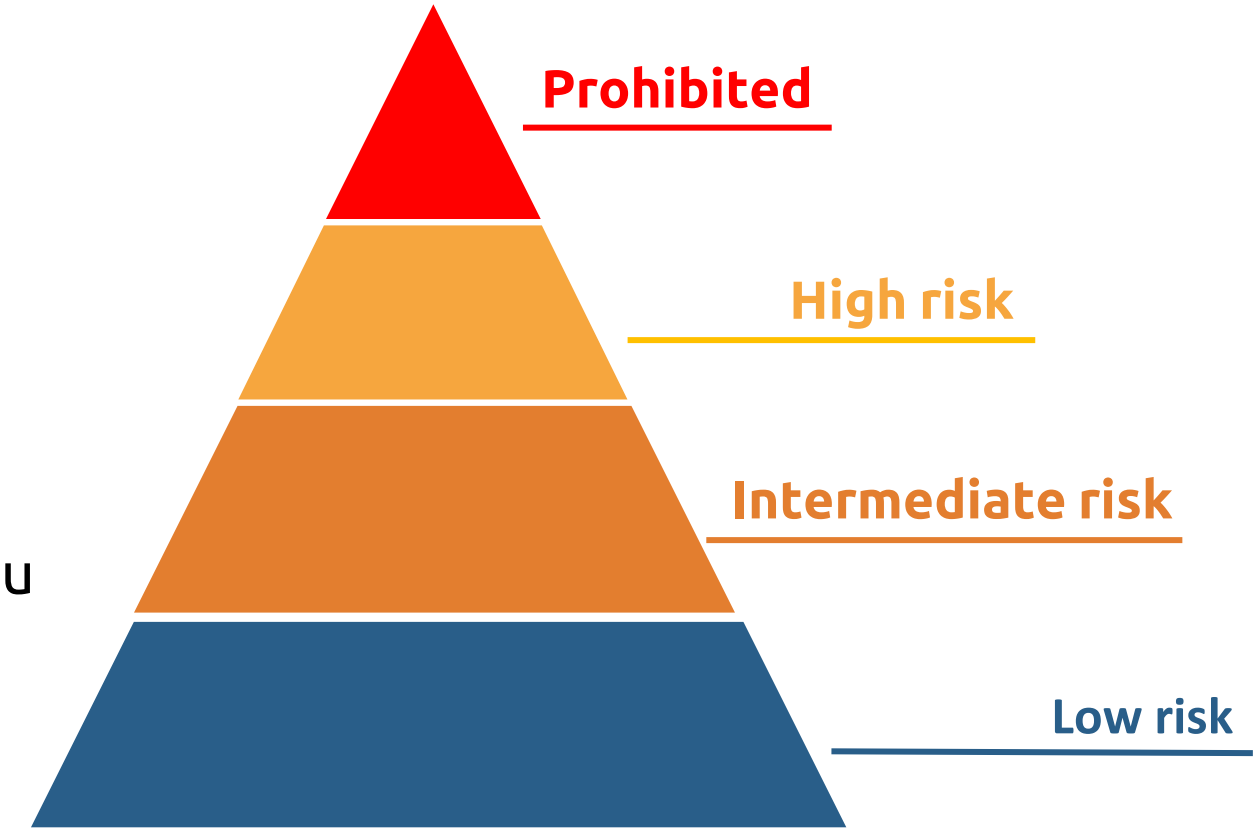
Quelle: 2023 Gartner® Top 10 Strategic Technology Trends: https://www.centreon.com/a-look-at-gartner-technology-trends-2023/

Die neue Rechtslage

Neue Regulierungen und Standards werden die Landschaft gründlich verändern

Beispiele:

- **GDPR**
- **EU AI Act**
- **EU Medical Devices Regulation**
- **EU AI Liability Directive**
- **ISO Standards** im KI-Bereich (z.B. zu Transparenz, NLP Accuracy



Fairness & Non-Diskriminierung als zentrale Parameter für AI

- **Trustworthy AI als Kernanforderung gemäß EU HLEG: u.a. Vielfalt, Nichtdiskriminierung und Fairness** (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai)

- **Fundamental Rights Impact Assessment (wahrsch.) verpflichtend für High Risk Anwendungen im AI Act**

Mandatory fundamental rights impact assessments

Following the advocacy efforts of ECNL and broader civil society, the European Parliament agreed that harms related to AI systems not only arise from how the system has been designed, but also from how and in which context it is deployed. Therefore, the text now includes an obligation for all deployers of high-risk systems, both from the private and the public sector, to assess the impact of the AI system on fundamental rights and specify how they are planning to mitigate risks, prior to putting it into use. When conducting the fundamental rights impact assessment, the deployer should also gather feedback from civil society and representatives of persons likely to be affected by the system. Thankfully, despite our concerns, technical standardisation bodies will not be able to impose rules for how to conduct this assessment.

Quelle: https://ecnl.org/news/big-win-fundamental-rights-european-parliament-adopts-ai-act

„Algorithmen sind menschliche Kreationen und unterliegen wie jedes andere menschliche Unterfangen auch Fehlern.“

US Richter Stephen Smith

R. Brauneis & E. P. Goodman, Algorithmic Transparency for the Smart City, 20 YALE J. L. & TECH. 103 (2018), FN63; https://yjolt.org/sites/default/files/20_yale_j._l._tech._103.pdf

leiwand.ai

Machen wir **AI**
gemeinsam **leiwand**!

Stay in touch!