Fairness entlang der KI-Wertschöpfungskette: Was ändert sich mit dem AI Act?

Dr. Till Klein

14. Nov 2024 | CDR-Konferenz | Berlin

appliedAI institute for europe

A joint initiative

UNTERNEHMERTUM | IPAI

Vorstellung

Dr. Till Klein

- Head of AI Regulation
- Mitglied bei OECD.AI & GPAI
- Erfahrung zu Regularien in Medizintechnik, Drohnen & als Auditor

Values & Principles
Reliability, Impact, Collaborative
Community, Learners Mindset

initiative for applied artificial intelligence

Advancing Europe's industry to compete in the age of AI

appliedAI institute for europe

Empowering professionals to develop and apply latest AI technologies responsibly

UnternehmerTUM München

IP.AI (Innovationspark KI Heilbronn)

Agenda

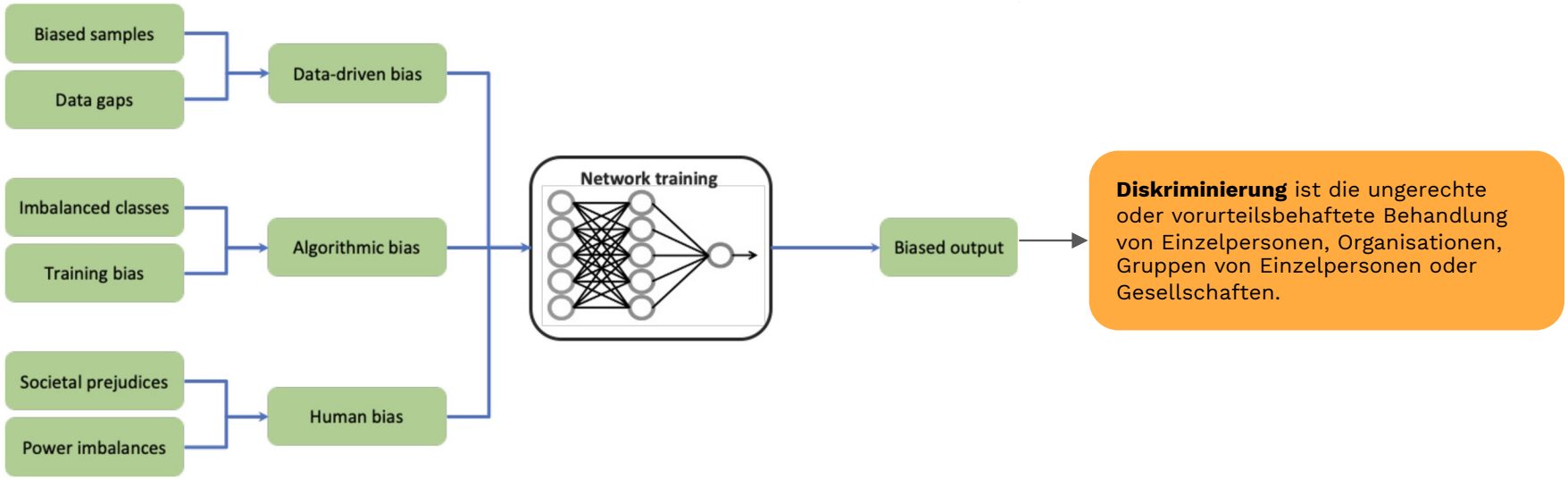
1. Warum ist Fairness wichtig für KI?
2. Fairness im AI Act: Was ändert sich?
3. Ausblick

Warum ist Fairness wichtig für KI?

Was ist Fairness? Wie kommt es in KI-Systemen zu einer Diskriminierung?

Fairness bedeutet, unparteiisch zu sein und sich ohne Bevorzugung oder Diskriminierung zu verhalten.

Voreingenommenheit (engl. bias): systematische unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen.



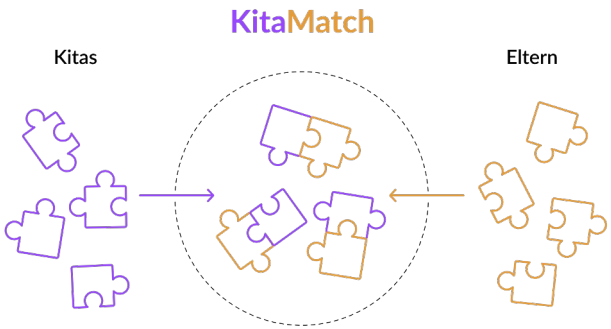
Diskriminierung ist die ungerechte oder vorurteilsbehaftete Behandlung von Einzelpersonen, Organisationen, Gruppen von Einzelpersonen oder Gesellschaften.

Quelle: Norori et. al, 2021, Addressing bias in big data and AI for health care: A call for open science, https://www.sciencedirect.com/science/article/pii/S2666389921002026

Quelle der Definitionen: IEC 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

Ob KI zu mehr oder weniger Fairness führt, ist eine Frage der Umsetzung im Kontext. Inklusion und Transparenz sind dabei zentral.

Kita Plätze per KI



"(...) Die Vergabe von Kita-Plätzen per Algorithmus ist ein gutes Beispiel dafür. Der neue Leitfaden erleichtert es insbesondere Kita-Leitungen und Jugendämtern, vorhandene Kita-Plätze fairer, schneller und transparenter zu vergeben"

Quelle: https://kitamatch.com/

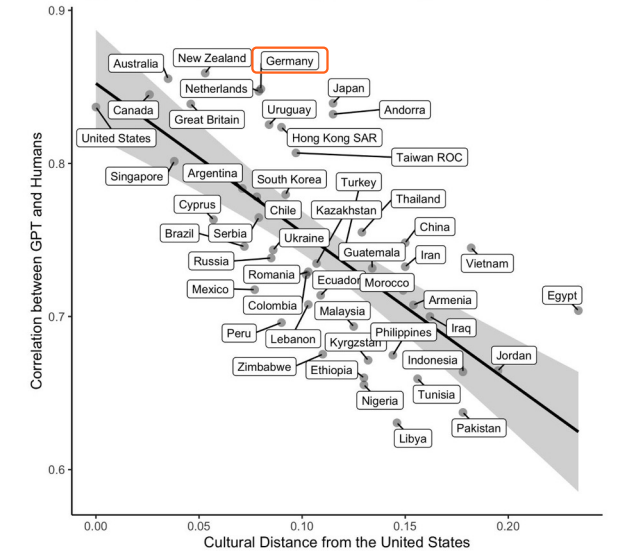
Uni Plätze per KI

"(...) The system at national level matches the requests to the respective intake capacities of (public and private) higher education institutions. (...) concerns among the plaintiffs of Discrimination against less wealthy prospective students or those from suburbs."

Quelle: Federal Anti-Discrimination Agency, 2019,
Risks of Discrimination through the Use of Algorithms

Verzerrungen in Basismodellen finden sich in nachgelagerten KI-Systemen und müssen entlang der Wertschöpfungskette mitigiert werden.

Text



"These results point to a strong WEIRD (Western, Educated, Industrialized, Rich, and Democratic) bias in GPT's responses to questions about cultural values, political beliefs, and social attitudes."

Quelle: Atari et. al (2023), Which Humans?, Harvard University

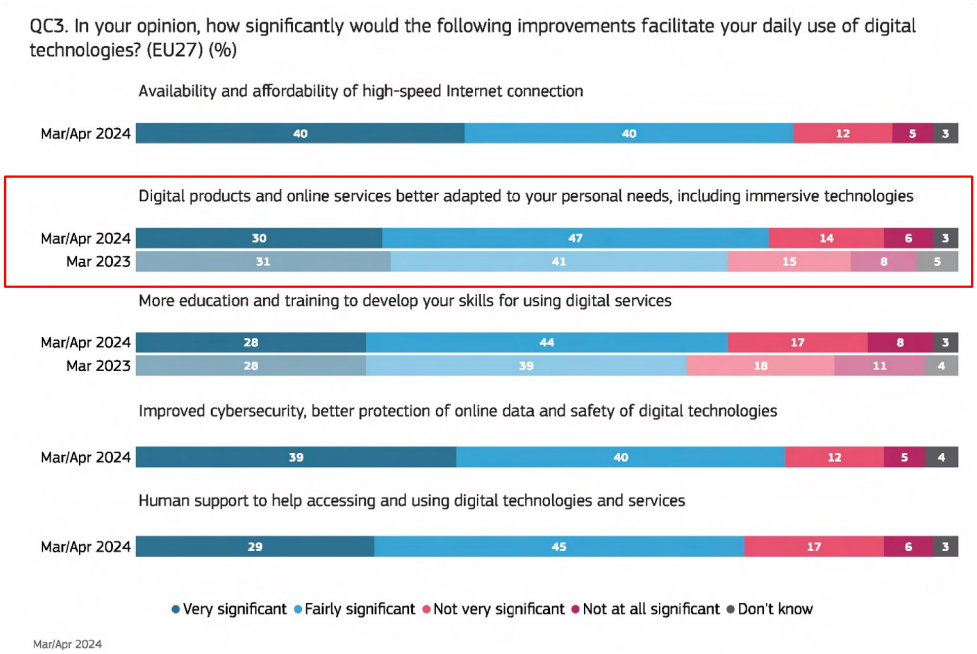
Bild

the American Smile (in a selfie)

Quelle: Jenka at Medium (2023), AI and the American Smile; Wen et. al (2024), Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation

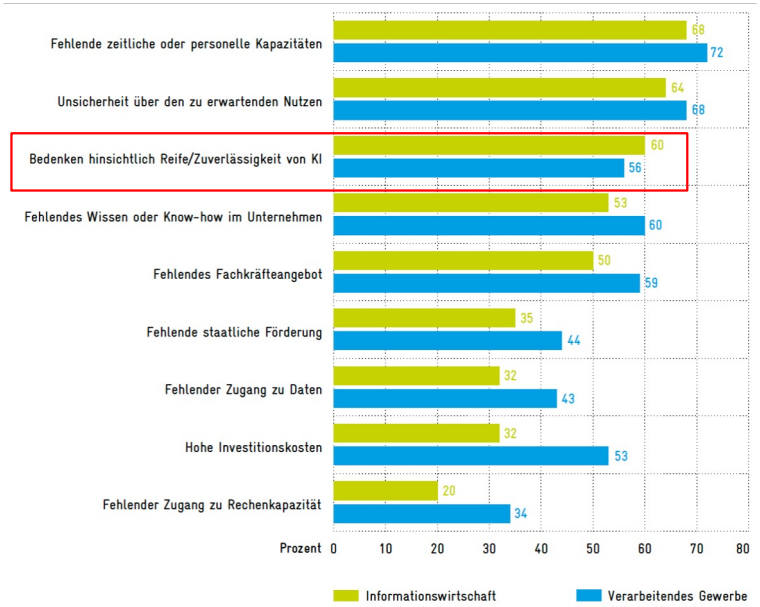
No trust, no use!
More trust, more use?

Fokus auf persönliche Bedürfnisse kann die Nutzung von KI steigern



Quelle: Eurobarometer, The Digital Decade, 2024, https://europa.eu/eurobarometer/surveys/detail/3174

Unzuverlässigkeit von KI ist ein Hindernis für die Nutzung in Unternehmen



Quelle: ZEW & EFI, EFI-Gutachten 2024, https://www.zew.de/presse/pressearchiv/bedenken-und-unsicherheit-hemmen-ki-einsatz-in-unternehmen

Fairness im AI Act:
Was ändert sich entlang der KI Wertschöpfungskette?

Die KI-Verordnung baut auf Ethik-Leitlinien auf. Jetzt gilt es, die neuen Regularien in die Praxis zu bringen.

Der Blick in die KI-Verordnung

Artikel 1 (KI-VO), Gegenstand

(1) Zweck dieser Verordnung ist es,

- die Einführung einer auf den Menschen ausgerichteten und vertrauenswürdigen künstlichen Intelligenz (KI) zu fördern und gleichzeitig
- ein hohes Schutzniveau in Bezug auf Gesundheit, Sicherheit und die in der Charta verankerten Grundrechte, (...) zu unterstützen.

Charta der Grundrechte der EU

Artikel 21, Nichtdiskriminierung
Artikel 23, Gleichheit von Frauen und Männern
Artikel 24, Rechte des Kindes
Artikel 25, Rechte älterer Menschen
Artikel 26, Integration von Menschen mit Behinderung

Der Blick auf die KI-Verordnung

Empfehlungen der High-Level Expert Group

Vier Grundsätze für vertrauenswürdige KI:

- Achtung der menschlichen Autonomie
- Schadensverhütung
- Fairness
- Erklärbarkeit

Integration der Empfehlungen in die KI-VO

Guidelines as have impact on:

- the legal requirements for high-risk AI systems
- the list of prohibited AI practices

Guidelines serve as:

- a normative compass for Europe's AI regulation
- a basis for (...) voluntary codes of conduct

Quelle: KI Verordnung: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=OJ:L_202401689#d1e39-1-1
Charta der Grundrecht der EU: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=OJ:L_202401689#d1e1900-1-1

HLEG Empfehlungen: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
Smuha, Nathalie A., The Work of the High-Level Expert Group on AI as the Precursor of the AI Act (October 01, 2024). Available at SSRN: https://ssrn.com/abstract=

Szenario: Jemand geht in eine Bank ...

... und hat ein seltsames Gefühl bei der Antwort auf einen Kreditantrag.

Finanzberaterin
Betreiber
(verwendet KI für creditscoring)

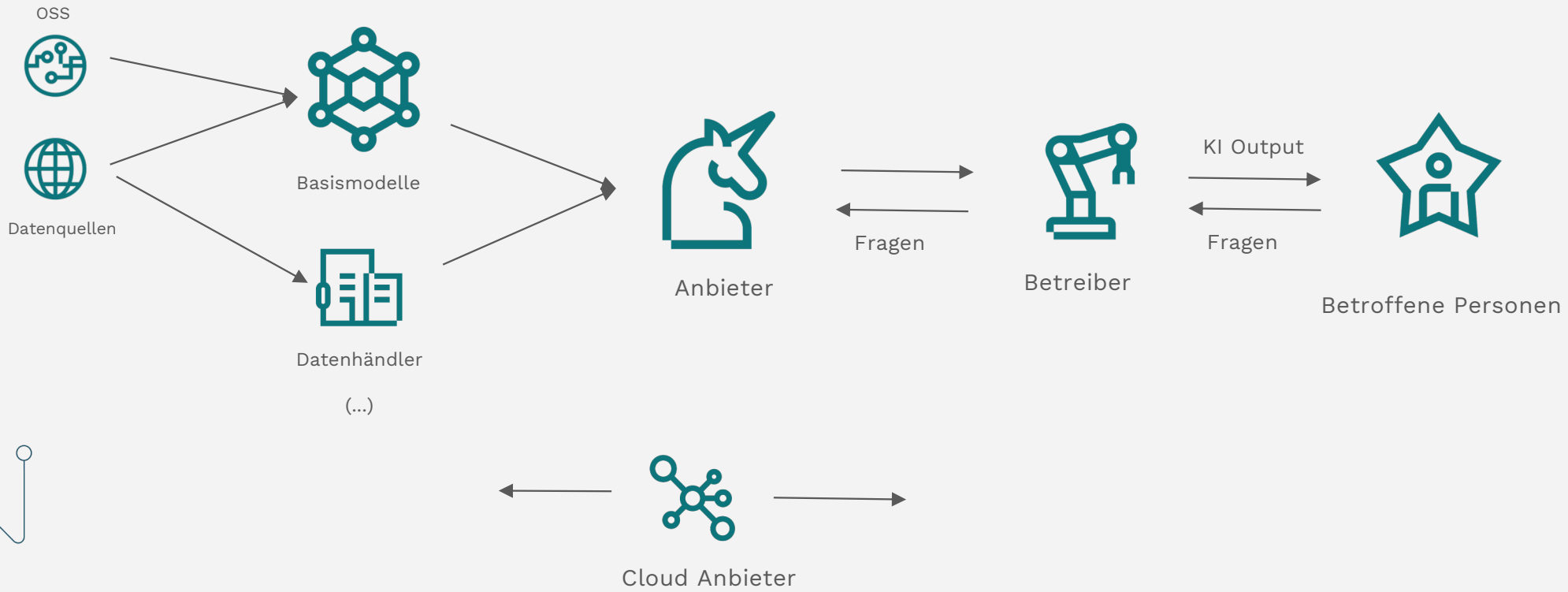
betroffene Personen

Ihre Kreditwürdigkeit ist X!

Wie sind Sie auf diesen Wert gekommen?

Bild: Mikhail Nilov@Pexels

Zur Erreichung der Schutzziele der KI-Verordnung müssen Informationen entlang der **KI-Wertschöpfungskette** geteilt werden



Was ändert sich für KI **betroffene Personen** bezüglich **Fairness**?

Was ist neu?

- Recht auf eine Erläuterung des Betreibers (Artikel 86)
- Recht bei der Marktüberwachungsbehörde eine Beschwerde einzulegen (Artikel 85)

Was ist offen?

- Wer ist eine betroffene Person (nicht definiert)?
- Wie erhalten betroffene Personen Kenntnis über ihre Recht?
- An welche Behörde kann man sich wenden?

Bild: Mikhail Nilov@Pexels

Was ändert sich für **Betreiber** bezüglich **Fairness**?

Was ist neu?

- Verpflichtende Grundrechte-Folgenabschätzung für bestimmte Hochrisiko-KI-Systeme (Artikel 27)
- Pflichten für Betreiber von Hochrisiko-KI-Systemen (Artikel 25)
  - Nutzung gemäß Betriebsanleitungen
  - Benennung von menschlicher Aufsicht
  - Meldung von Vorfällen an Anbieter und Behörden
- Informationen an Arbeitnehmervertreter vor Inbetriebnahme
- Eingabedaten müssen zweckbestimmt und repräsentativ sein

Was ist offen?

- Eine geeignete Methode für Grundrechte-Folgenabschätzung
- Schulungsangebote für die KI-Aufsicht
- Erfahrung mit der Erkennung meldepflichtiger Vorfälle

Bild: Mikhail Nilov@Pexels

Was ändert sich für **Anbieter von KI Systemen** bezüglich **Fairness**?

Was ist neu?

Hochrisiko-KI-Systeme:

- Risiko Management inkl. Fokus auf Grundrechte und auf Minderjährige und schutzbedürftige Gruppen (Artikel 9)
- Data Governance Traningsdaten müssen “relevant, repräsentativ, fehlerfrei und vollständig” sein; Kontext & Region
- Pflichten für Anbieter von Hochrisiko-KI-Systemen (Artikel 16)
- Meldepflicht schwerwiegender Vorfällen Behörden (Artikel 72)

Mittel-Risiko-KI-Systeme: Transparenz

Niedrigrisiko-KI-Systeme: KI-Kompetenz, freiwilliger Kodex (Art. 4 & 95)

Was ist offen?

- Die harmonisierten Normen werden (erst) H2/2025 erwartet
- Methoden zur korrekten Umsetzung der Normen
- Metriken zur “quantifizierung” der Risiken zu Grundrechten
- Die Behörden zur Marküberwachung in DE sind zu benennen

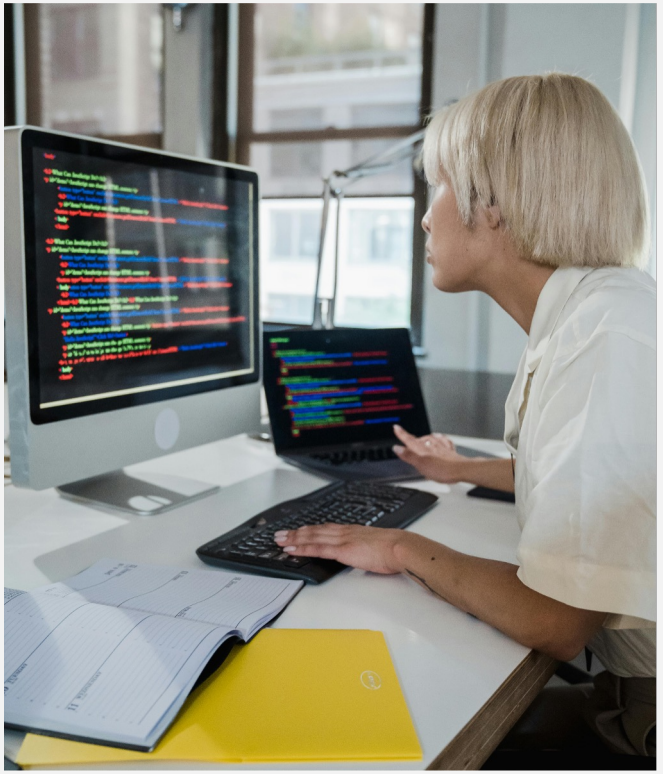


Bild: Mizuno K@Pexels

Was ändert sich für **Anbieter von Basismodellen** bezüglich **Fairness**?

Was ist neu?

- Einstufung von KI-Modellen mit allgemeinem Verwendungs- zweck mit/ohne systemischem Risiko (Artikel 51)
- Transparenzpflichten ggü. nachgelagerten Anbietern (Artikel 53)
  - Einhaltung von Urheberrecht
  - Zusammenfassung der Trainingsdaten
  - Technische Dokumentation

Was ist offen?

- Der Praxisleitfaden für Basismodelle soll April 2025 veröffentlicht werden
- Die Reaktion der vornehmlich US-Anbieter auf die KI-VO
- Umsetzung von Verträgen entlang der Wertschöpfungskette

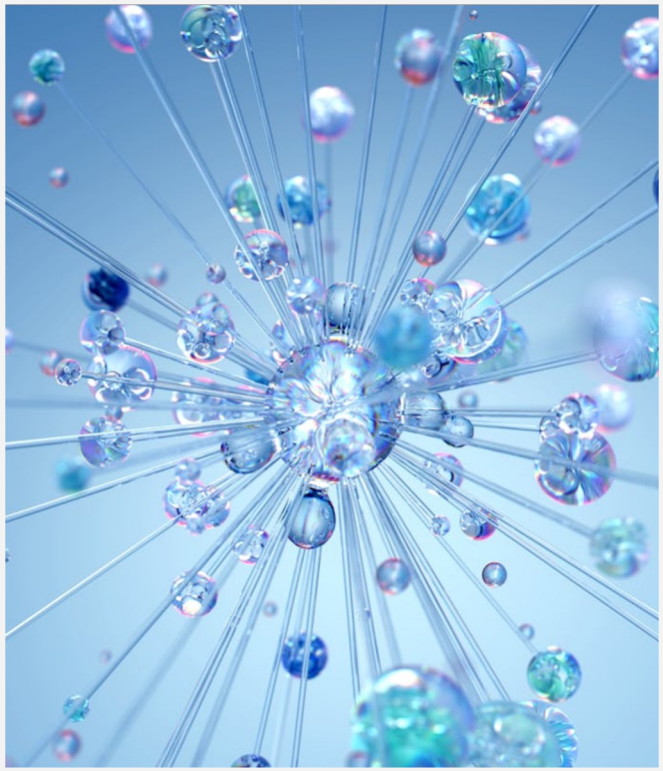
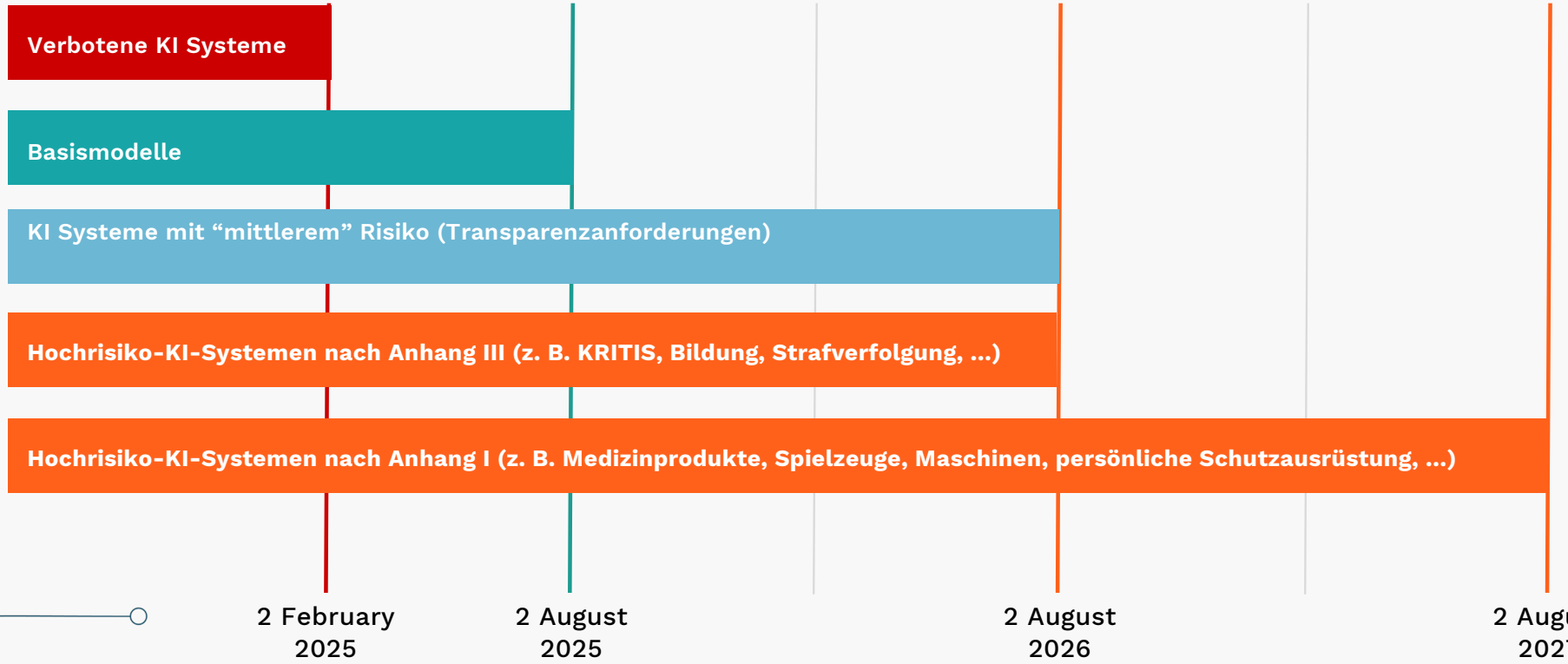


Bild: Google DeepMind@Pexels

Ausblick

Fristen in der KI-Verordnung

Die KI-Verordnung wird 24 Monate nach seinem Inkrafttreten vollständig anwendbar sein, wobei für jede Risikoklasse und Basismodelle der folgende abgestufte Ansatz gilt:



Ausblick
Oder: Was können wir heute machen?

1. Aktive Teilnahme bei Ausgestaltung der KI Verordnung (AI Office, Standards, Nationale Umsetzung).
2. Experimentieren mit “doppeltem Boden”.
3. Nutzen Sie CDR-Initiative als Plattform fürs gemeinsame lernen.

Vielen Dank!

Dr. Till Klein
Head of AI Regulation

t.klein@appliedai-institute.de

Office
Munich Urban Colab
Freddie-Mercury-Straße 5
D-80797 München