AI Factory Austria AI:AT

Grundlagen von Trustworthy AI

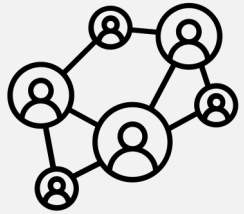
Pia Weinlinger
AI:AT

Michael Löffler
AI:AT

Warum brauchen wir die AI Factory Austria AI:AT?

Souveränität

Ethik und Trustworthiness



KI-Ökosystem

Unsere Mission
Von der Idee zur Innovation

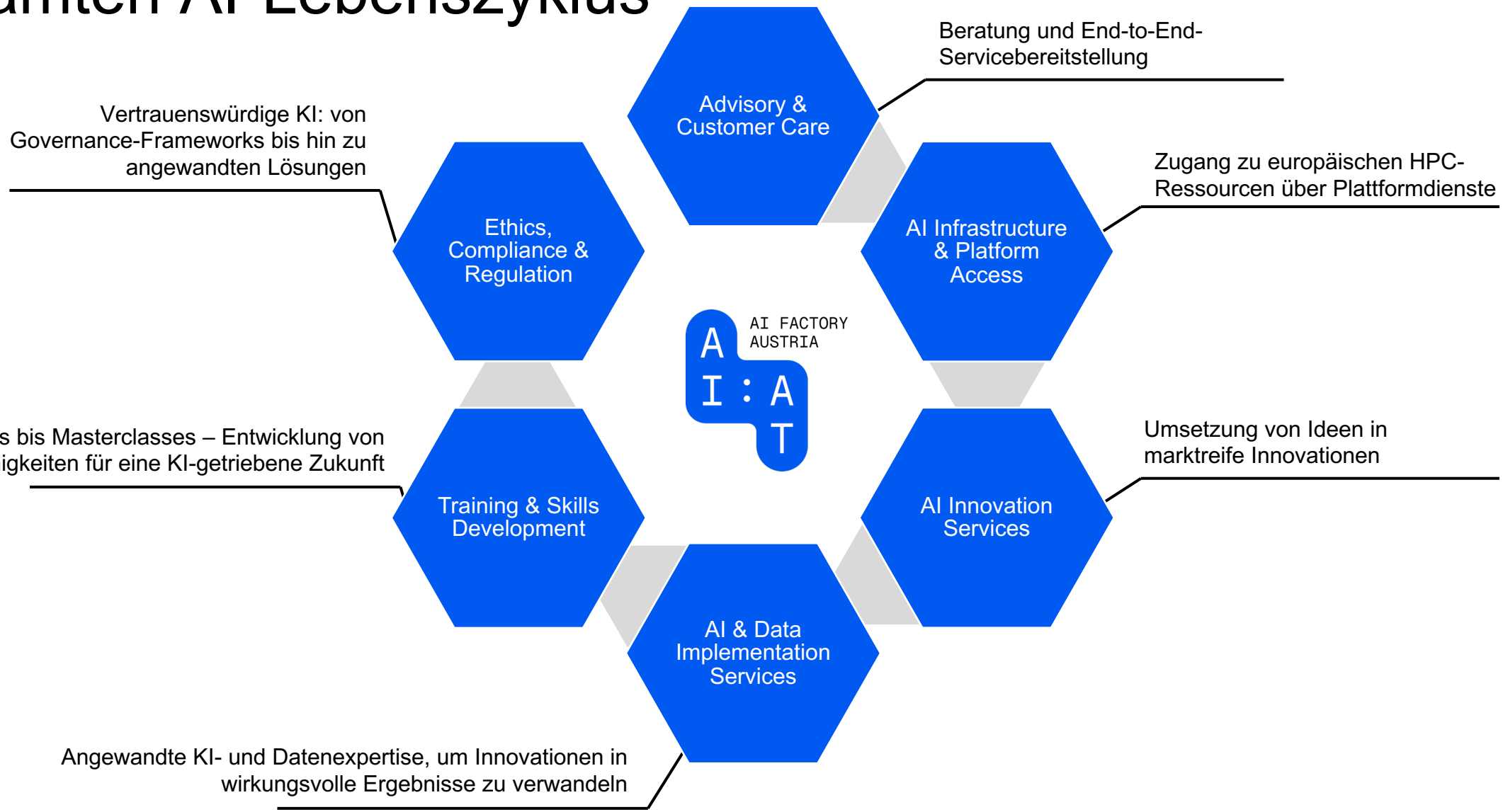
Etablierung eines One-Stop Shop für AI

Schließen des AI Ressourcen- und Wissens-Gap

Bewerbung von ethischer und trustworthy AI

Launchpad für AI-getriebene Innovation

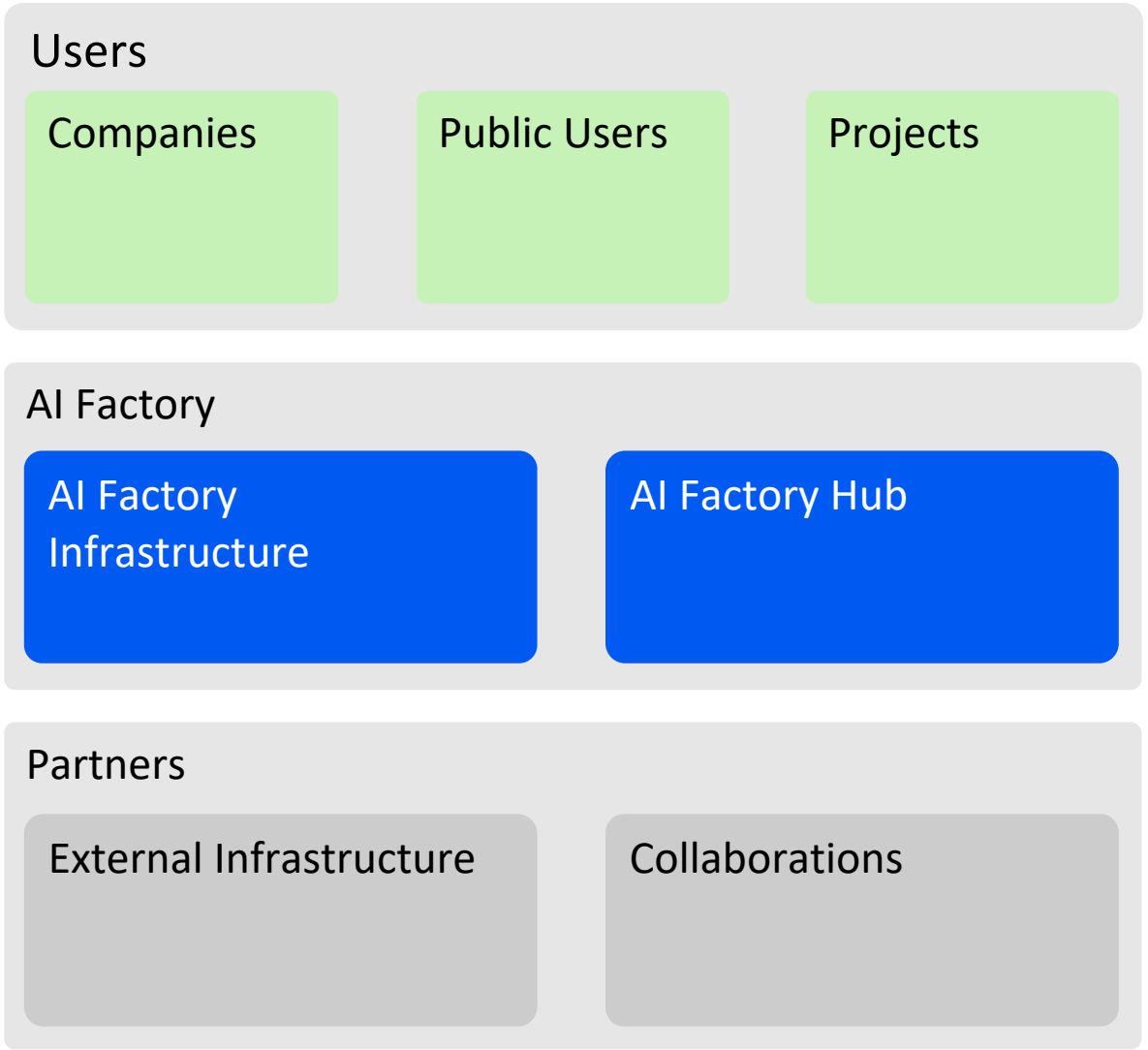
Unsere Services unterstützen über den gesamten AI-Lebenszyklus



Innovationsbereiche

Core Areas	Biotech	Industry Manufacturing	Public Administration	Physics
Additional Areas	Health	Fintech Lawtech	Environment & Sustainability	Other

AI:AT Network



AI:AT Key Strength

- **Together in one place**: spaces, programs, resources, and people.
- **Start-ups, Research, Companies, and Public Institutions** work here side by side.
- Technical development, legal security, and entrepreneurial thinking interlock seamlessly.
- We not only assist with implementation – we also support the **search for potential**.
- Included: Access to **Computing power, coaching, Funding Agencies, and Investors**.
- **Goal:** From the **first idea to a scalable product** – all under one roof, without detours.

AI Factory Austria AI:AT Consortium

Disclaimer:
The speakers are solely sharing their personal experiences. Therefore, this free seminar is not a substitute for professional/legal advice.

Beneficiaries

Affiliated Entities

AI Factory Austria AI:AT

Grundlagen von Trustworthy AI

Pia Weinlinger
AI:AT

Michael Löffler
AI:AT

Why do we need AI Factory Austria?

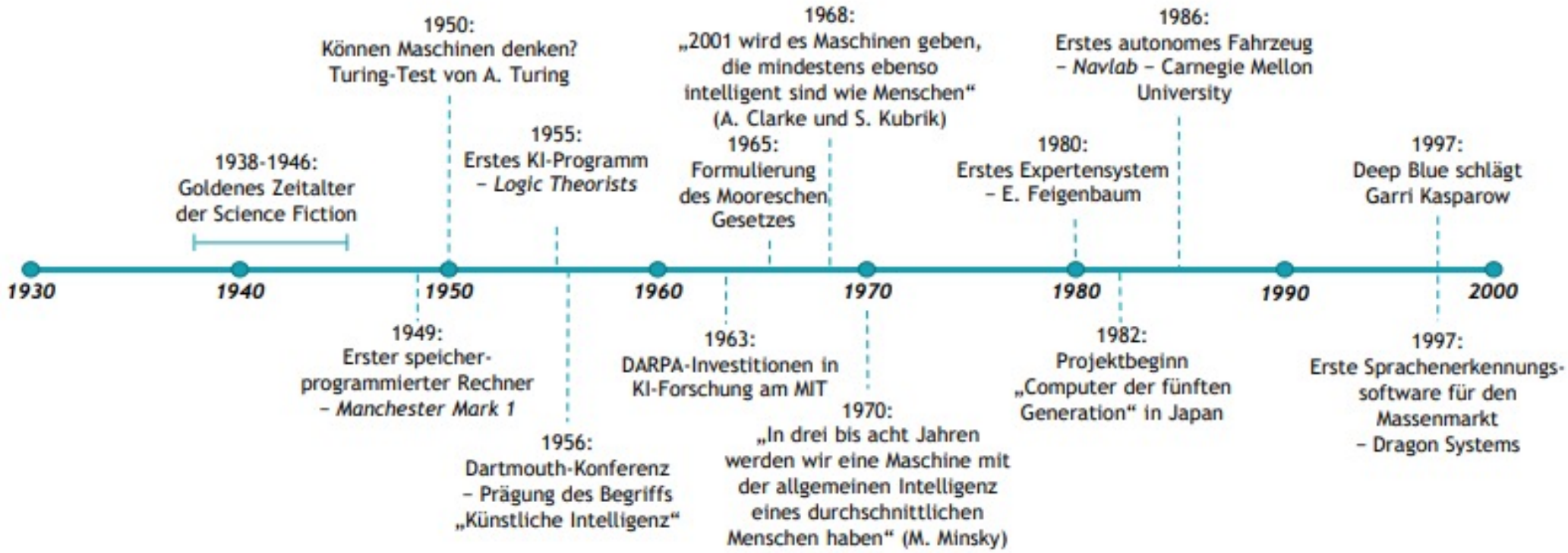
Sovereignty

Ethics and Trustworthiness

Connecting the Ecosystem

Eine kurze Geschichte der Künstlichen Intelligenz (KI)

- **Ziel:** Maschinen zu entwickeln, die wie Menschen lernen und denken



Quelle: OECD 2020: Künstliche Intelligenz in der Gesellschaft nach Anyoha (2017) https://sites.harvard.edu/sitn/2017/08/28/history-artificial-intelligence/

Wo stehen wir heute?

Stand der Technik

Starke KI	Schwache KI
• Systeme handeln, lernen und entwickeln sich selbständig weiter. • Können neue, bislang unbekannte Aufgabenstellungen ohne menschliche Intervention lösen. • Entwickeln eigene Lernstrategien und setzen Ziele eigenständig. • Verfügen über Fähigkeiten, die menschlicher Intelligenz ähneln, wie selbstständiges Planen und Problemlösen.	• Systeme lösen spezifische Aufgaben in einem klar definierten Anwendungsbereich. • Verarbeiten Eingabedaten, um eine vordefinierte Ausgabe zu generieren. • Haben keine Fähigkeiten zur selbständigen Weiterentwicklung oder Problemlösung außerhalb ihres spezifizierten Anwendungsbereichs.

https://oeffentlicherdienst.gv.at/wp-content/uploads/2024/09/250113_Leitfaden-Digitale-Verwaltung_2.0_A4.pdf

Blick in die Zukunft: Welche Auswirkungen hätte eine „starke KI“ auf unser Leben?

Dystopie oder Utopie?

An diesem Punkt reicht es nicht zu sagen: *„Die Technik funktioniert halt so.“* Wir müssen fragen: Ist das gerecht? Ist das nachvollziehbar? Wer trägt Verantwortung?

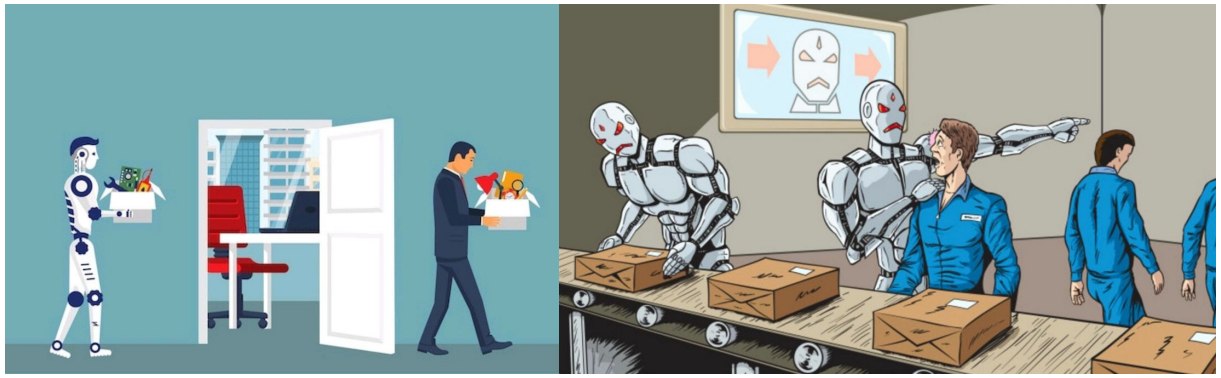
Wir sehen bereits reale Risiken von „schwacher KI“ – auch ohne „Superintelligenz“

- Wir schaffen viele leistungsstarke Black-Box-Systeme, die wir nicht wirklich verstehen.
- Wir machen KI-Systeme größer und leistungsfähiger, nicht verständlicher oder kontrollierbarer.
- Wir übertragen KI-Systemen immer mehr Verantwortung, als wir sicher überwachen können.

Realistische Risiken von KI

Arbeitslosigkeit aufgrund von Automatisierung

Wenn wir sagen „Effizienz optimieren“, können Arbeitsplätze zu „Ineffizienzen“ werden.



Mangel an Erklärbarkeit und Transparenz

Tesla nutzte KI-Modelle, die mit Milliarden von gefahrenen Kilometern trainiert wurden, um das Fahrverhalten in einen „Safety Score“ zu übersetzen aber die meisten Fahrer:innen haben keine Ahnung, wie KI sie bewertet oder wie sie diese Bewertung anfechten können.

Realistische Risiken von KI

Bias und Diskriminierung

VERBORGENER BIAS

11.10.2025, 08:45 Uhr

Diskriminierung durch Daten: So unfair urteilt KI über Ostdeutsche

Künstliche Intelligenz soll objektiv urteilen, doch die Realität sieht meist anders aus. Eine Untersuchung der Hochschule München und Forschungsergebnisse der Cornell University zeigen: Chatbots wie ChatGPT können sogar regionale Vorurteile innerhalb Deutschlands reproduzieren.

https://www.ingenieur.de/technik/fachbereiche/kuenstliche-intelligenz/diskriminierung-durch-daten-so-unfair-urteilt-ki-ueber-ostdeutsche/

Privatsphäre und Datenschutz

TechScape: Clearview AI was fined £7.5m for brazenly harvesting your data – does it care?

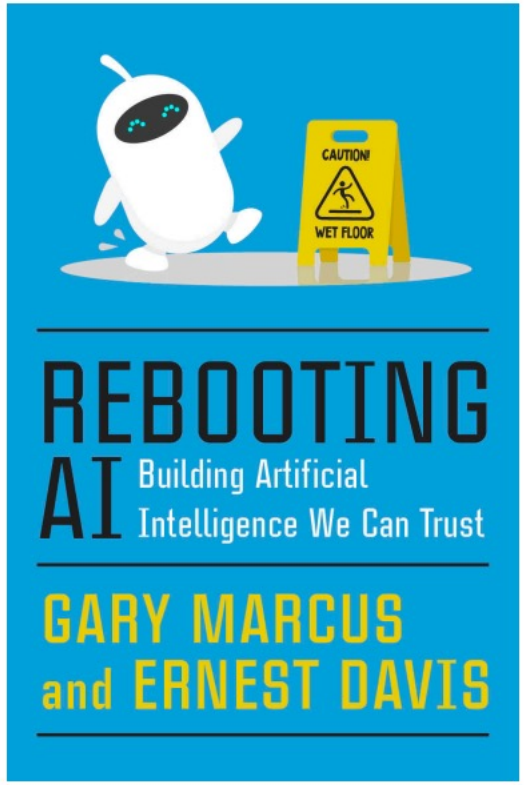
The facial recognition firm earned a heavy fine for scraping millions of pictures of people's faces from social media. But that doesn't mean it will change its ways

https://www.theguardian.com/technology/2022/may/25/techscape-clearview-ai-facial-recognition-fine?

Die zentrale Frage lautet: Wie können wir sicherstellen, dass die Vorteile von KI die potenziellen Risiken überwiegen?

Ein gesünderes Leben

Bildquelle. https://unsplash.com/de/s/fotos/ai-in-healthcare



Gesteigerte Produktivität und Wohlstand

Bildquelle: https://www.istockphoto.com/

Ein menschenzentrierter Ansatz für KI

“Artificial intelligence should treat all people fairly, empower everyone, perform reliably and safely, be understandable, be secure and respect privacy, and have algorithmic accountability. It should be aligned with existing human values, be explainable, be fair, and respect user data rights. It should be used for socially beneficial purposes, and always remain under meaningful human control.”

Tom Chatfield (2020) There’s No Such Thing As Ethical AI

– **KI-Ethik** bezieht sich auf moralische Grundsätze und Werte, die die Entwicklung und Nutzung von KI-Systemen leiten sollen.

Und wer entscheidet, welche moralischen Vorstellungen und Werte für KI als gut für uns Menschen gelten sollen?

Wissenschaft & Zivilgesellschaft

- **Ada Lovelace Institute (UK)**
- **Data & Society (USA)**

Unternehmen

- **Microsoft** „Responsible AI Standard“, 6 KI-Prinzipien (Fairness, Zuverlässigkeit/Sicherheit, Datenschutz & Sicherheit, Inklusivität, Transparenz, Rechenschaft)
- **Google/Google DeepMind** eigene „AI Principles“ (u.a. Fairness, Sicherheit, Verantwortlichkeit)
- **IBM** „Trustworthy AI“ (Transparenz, Erklärbarkeit, Robustheit, Fairness)

High Level Policy & Regulierung

- **UNESCO Recommendation on the Ethics of AI (2021):** Menschenrechte, Menschenwürde, Transparenz, Fairness, menschliche Aufsicht, Nachhaltigkeit
- **UN-Generalversammlung – Resolution A/RES/78/265 (März 2024)** „Seizing the opportunities of safe, secure and trustworthy AI systems for sustainable development.” „The same rights that people have offline must also be protected online, including throughout the life cycle of artificial intelligence systems.”
- **EU**: **EU AI Act + HLEG-Prinzipien** für vertrauenswürdige KI (u.a. Robustheit/Sicherheit, Transparenz, Diversität, Fairness,Nichtdiskriminierung, gesellschaftliches & ökologisches Wohlergehen)

KI-Ethik in der Praxis umsetzen

Abstrakte ethische Prinzipien sind leichter zu vereinbaren, als in der Praxis umzusetzen:

Beispiel: Fairness

Viele Arten von bias/Verzerrungen gleichzeitig

- **Repräsentationsbias** (Trainingsdaten sind nicht repräsentativ)
- **Historische/gesellschaftliche Bias** (frühere Diskriminierung, die in den Daten verankert ist)
- **Algorithmischer Bias** (Modell verstärkt bestehende Muster)
- **usw……..**

Verschiedene Verzerrungen, die bei jedem Prozessschritt der ML-Pipeline auftreten können

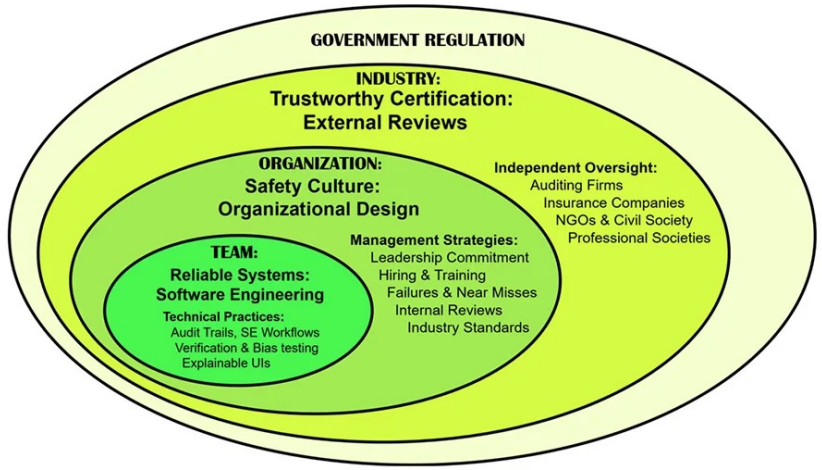
Kheya, T. A., Bouadjenek, M. R., & Aryal, S. (2024). The pursuit of fairness in artificial intelligence models: A survey (arXiv Preprint No. 2403.17333). arXiv.

KI-Ethik in der Praxis umsetzen

- **Governance frameworks innerhalb von Organisationen:**
  - „Responsible AI Governance“
- **Übersetzung ethischer Prinzipien in Fragestellungen:**
  - Ethik-Checklisten sind z.B. nützlich, aber nur, wenn sie

  (a) mit echter Entscheidungsmacht verknüpft sind (Projekte können gestoppt oder neu gestaltet werden),

  (b) sie in Rechts-, Sicherheits- und Risikomanagementfunktionen integriert sind.
- **Impact und Risikoassessments:**
  Diese Instrumente helfen uns, uns vorzustellen, was passiert, wenn das jeweilige KI-System tatsächlich im großen Maßstab eingesetzt wird.



Shneiderman, B. (2020). *Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems.* ACM Transactions on Interactive Intelligent Systems, 10(4).

Leitfaden digitale Verwaltung: KI, Ethik und Recht

Take-Home Messages

- **Ethik ist komplex und bleibt wichtig**
  - Es gibt viele offene Forschungsfragen → wir sind mittendrin, nicht am Ende.
- **Derzeitige KI-Systeme sind oft ein Spiegel von uns Menschen**
- **Zwei Ebenen der KI-Ethik:**
  - Kurzfristig „KI-Ethik von heute”:
    - Bias, Deepfakes, Gesichtserkennung, Urheberrecht usw.
  - Langfristig Fragen zur „Superintelligenz”:
    - Was passiert, wenn KI intelligenter wird als wir selbst?
- **Rolle des AI Factory Ethics Teams**
  - Unterstützt bei Fragen zur heutigen KI-Ethik
  - Entwickelt Rahmenbedingungen und Leitlinien für vertrauenswürdige KI
  - Unterstützt Stakeholder dabei, vertrauenswürdige KI-Systeme zu entwickeln und einzusetzen

Why do we need AI Factory Austria?

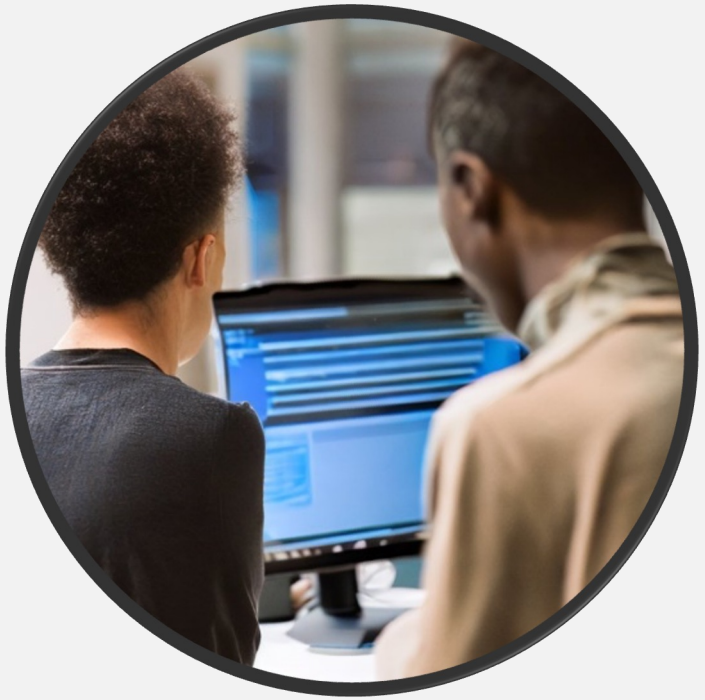
Sovereignty

Ethics and Trustworthiness

Connecting the Ecosystem

Was ist „künstliche Intelligenz“?

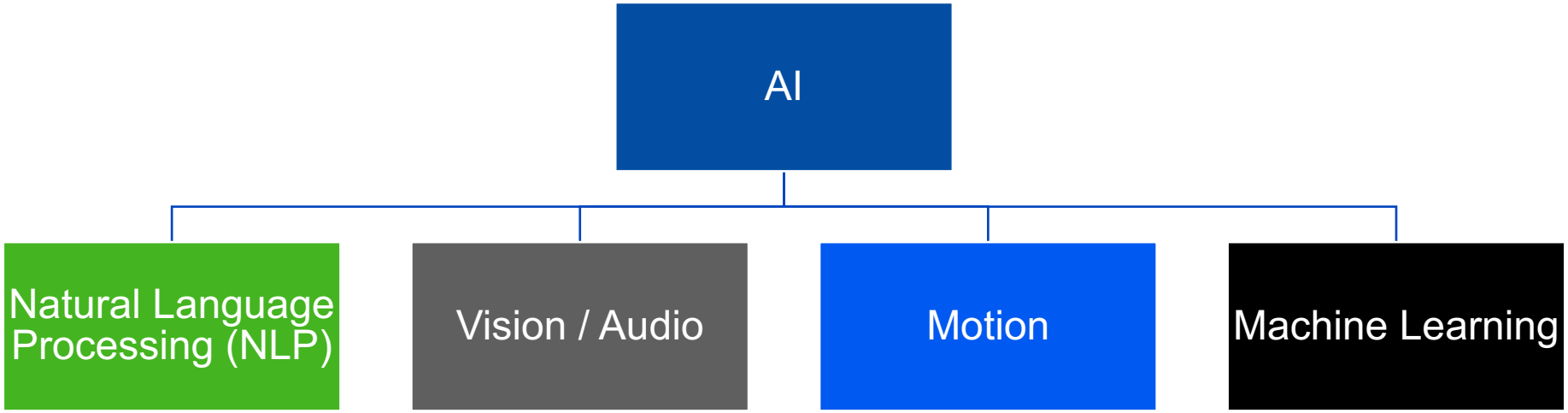
- Ein System, dass beim Chatten nicht von einem Menschen unterschieden werden kann („Turing-Test“)



„Turing-Test“, von Alan Turing 1950 formuliert.

Bildquelle: Stable-Diffusion 2.1.

Was ist „künstliche Intelligenz“?



Vgl: IBM Technology, AI vs Machine Learning (https://www.youtube.com/watch?v=4RixMPF4xis).

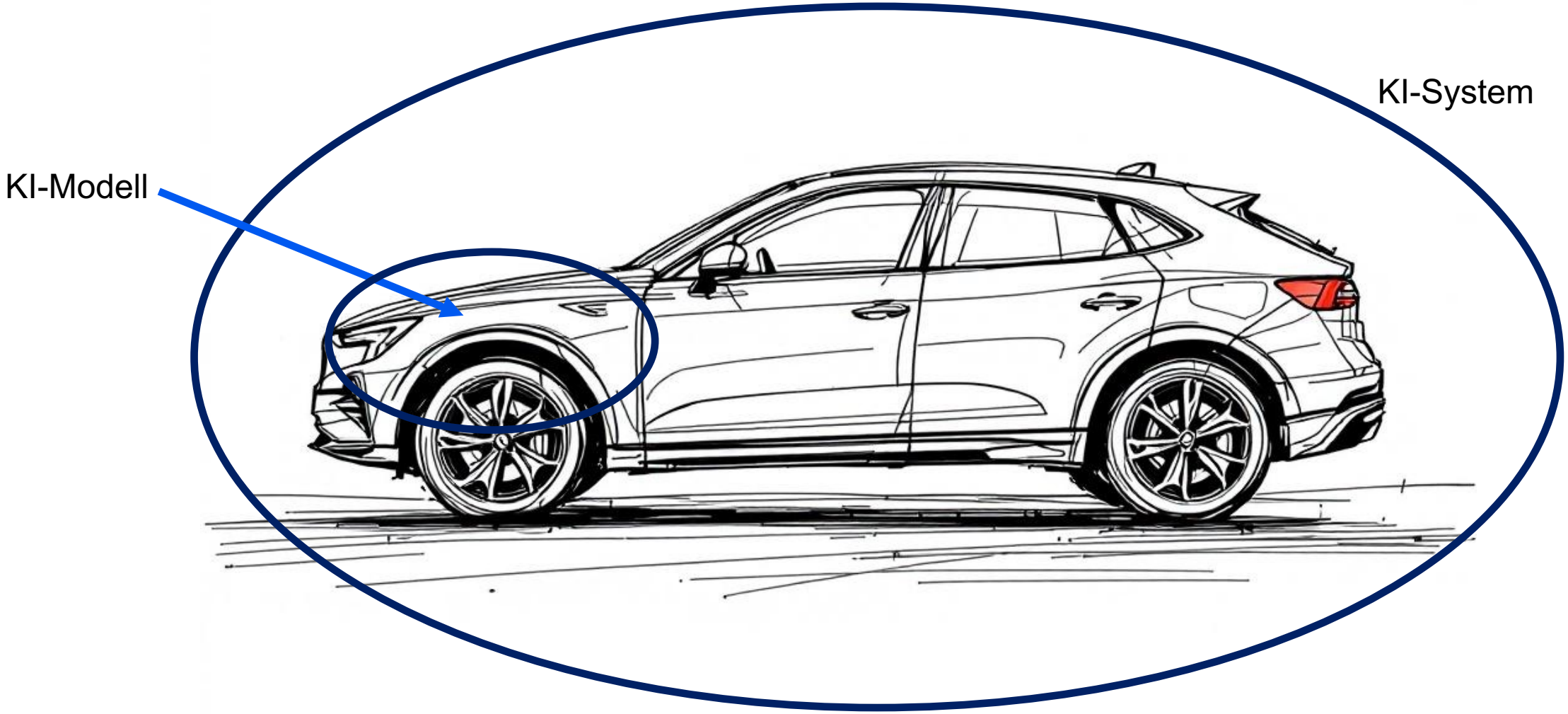
Was ist „künstliche Intelligenz“?

KI-System ist[1]

- ein **maschinengestütztes** System,
- das für einen in unterschiedlichem Grade **autonomen Betrieb** ausgelegt ist,
- das nach seiner Betriebsaufnahme **anpassungsfähig** sein kann und das aus den erhaltenen Eingaben für **explizite oder implizite Ziele ableitet**,
- wie **Ausgaben** wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen hervorgebracht werden,
- die physische oder virtuelle Umgebungen **beeinflussen können**.

1: Art 3 Z 1 Verordnung über künstliche Intelligenz (KI-VO) bzw *European Commission (EC)*, Commission Guidelines on the definition of an artificial intelligence system established by Regulation (EU) 2024/1689 (AI Act), 06.02.2025.

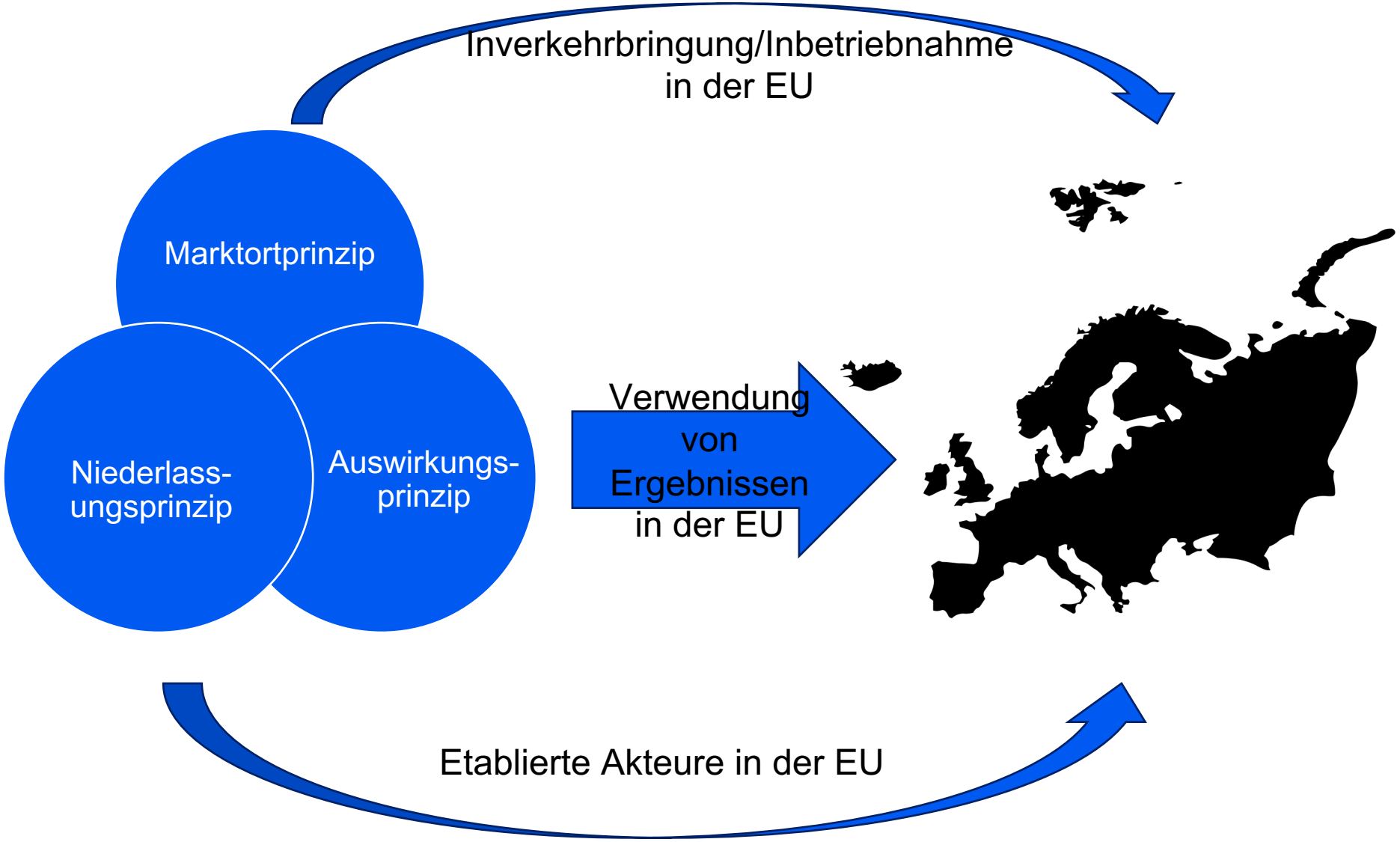
Wie verhalten sich KI-Modelle zu KI-Systemen



 Vgl. *Almada*, Law & Compliance in AI Security & Data Protection, 12.2024.

Bildquelle: Le Chat

Wann gilt die KI-VO?

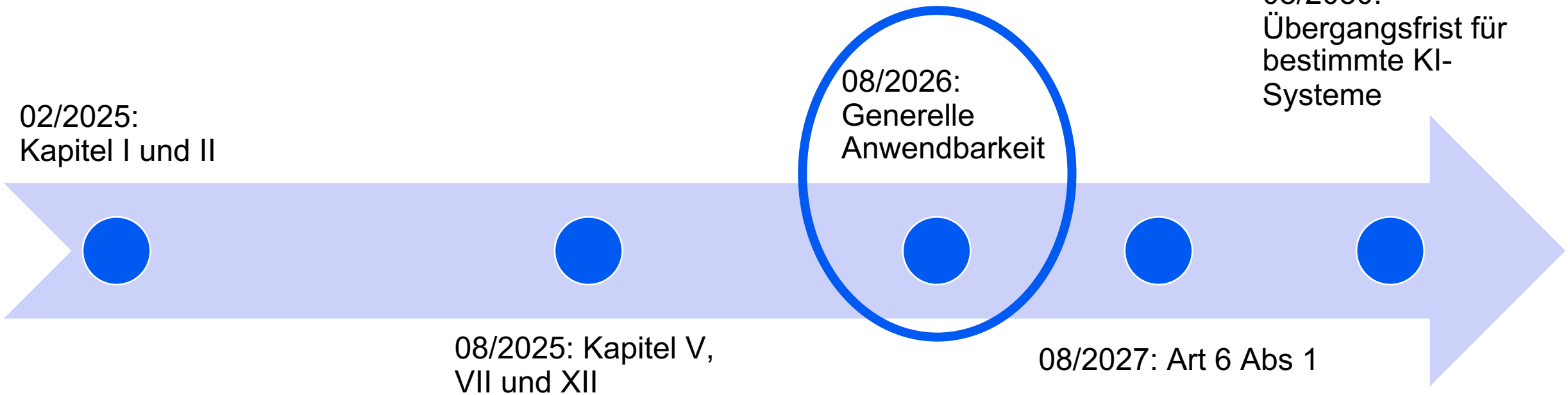


Wann gilt die KI-VO nicht?

- Die KI-VO gilt unter anderem **nicht** für:
- ...
- für KI-Systeme oder KI-Modelle, die **ausschließlich** für die **wissenschaftliche Forschung und Entwicklung** entwickelt und in Betrieb genommen werden;
- für KI-Systeme, die unter freien und **quelloffenen Lizenzen** bereitgestellt werden, es sei denn, sie sind Verbotene-KI-Systeme, Hochrisiko-KI-Systeme oder Art-50-KI-Systeme;
- …

1: Art 2 Verordnung über künstliche Intelligenz.

Ab wann gilt die KI-VO?



Ab wann gilt die KI-VO?

Zeitpunkt:	02/2025	08/2025	08/2026	08/2027
Zentrale Regeln (Auszug):	- KI-Kompetenz - Verbote	- Behörden-wesen - allg. KI-Modelle	- Anhang-III-Hochrisiko-KI-Systeme	- Anhang-I-Hochrisiko-KI-Systeme

Ab wann gilt die KI-VO?

Kommissionsvorschlag (Digital Omnibus on AI):

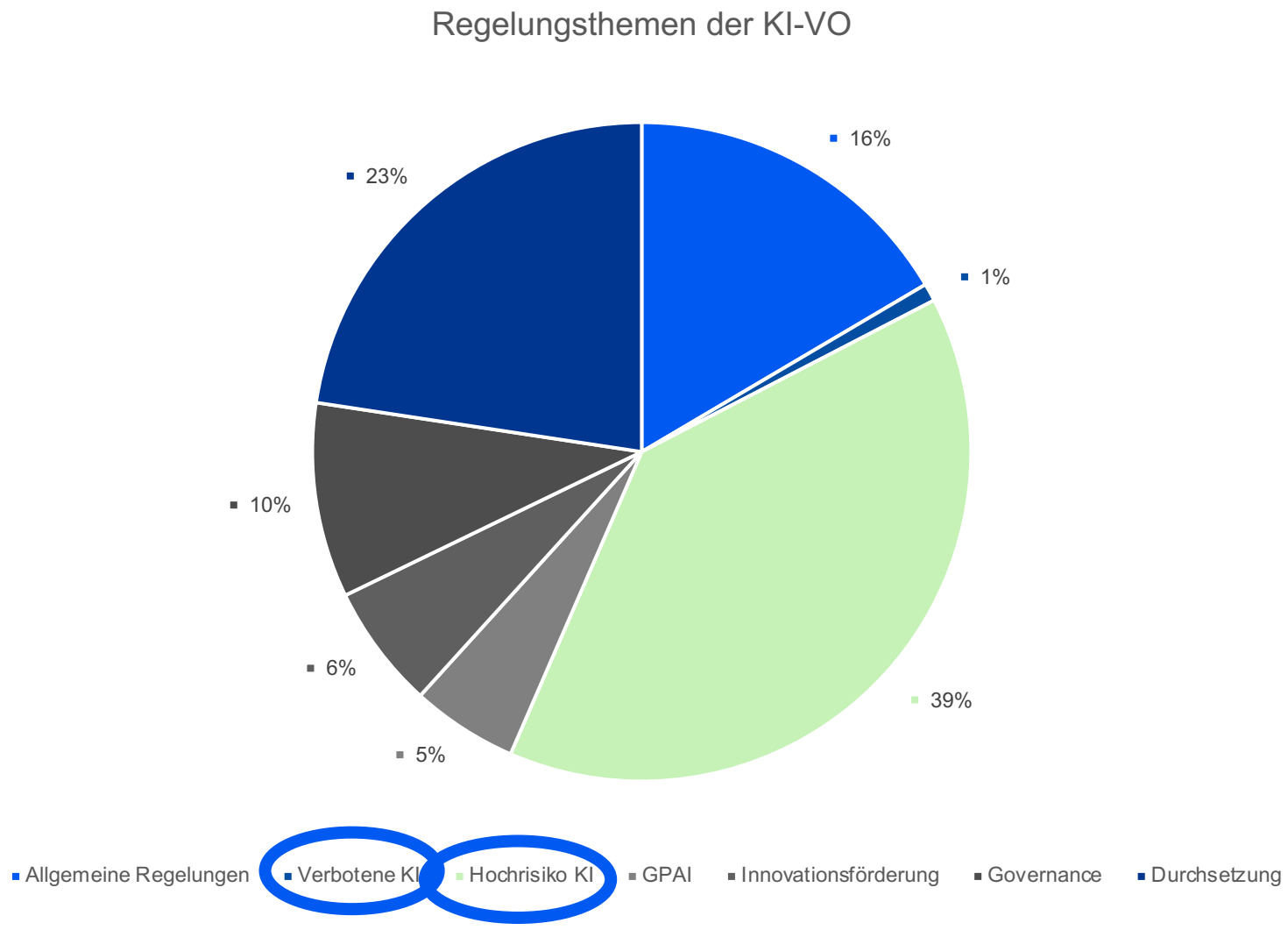
…in Diskussion!

- Verschiebung der Geltung der Regelungen zu:
  - Einstufung von KI-Systemen als Hochrisiko-KI-Systeme
  - Anforderungen an Hochrisiko-KI-Systeme und
  - Pflichten der Anbieter und Betreiber von Hochrisiko-KI-Systemen und anderer Beteiligter

- Diese Regeln sollen (zeitlich gestaffelt) erst gelten, wenn die Kommission bestätigt, dass angemessene Unterstützungsmaßnahmen zur Einhaltung der Regelungen vorhanden sind.

- **Spätestens** sollen die Regeln gelten:
  - für Hochrisiko-KI-Systeme nach Anhang III: 12/2027
  - für Hochrisiko-KI-Systeme nach Anhang I: 08/2028

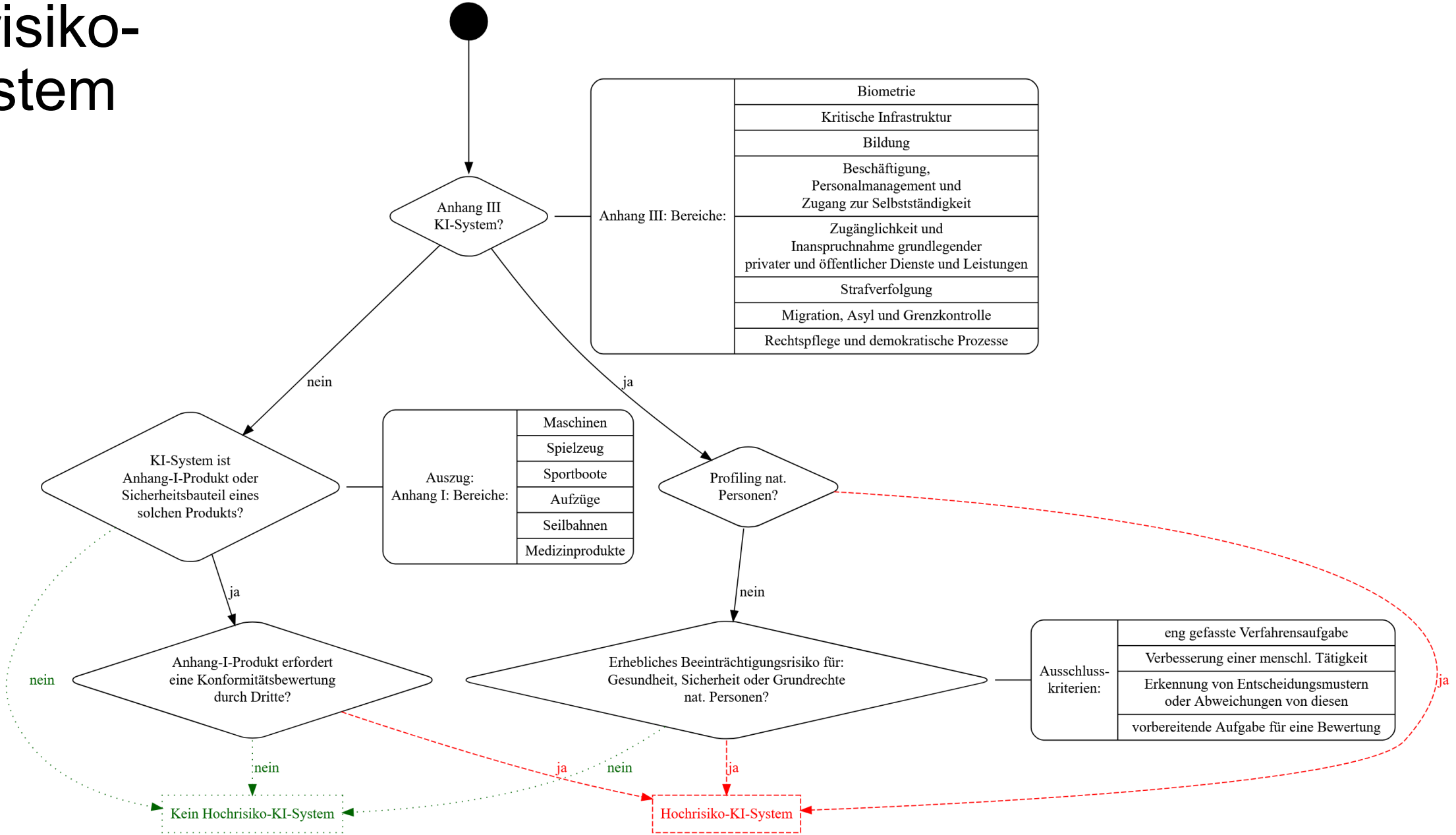
Was regelt die KI-VO?



Verbotene Praktiken

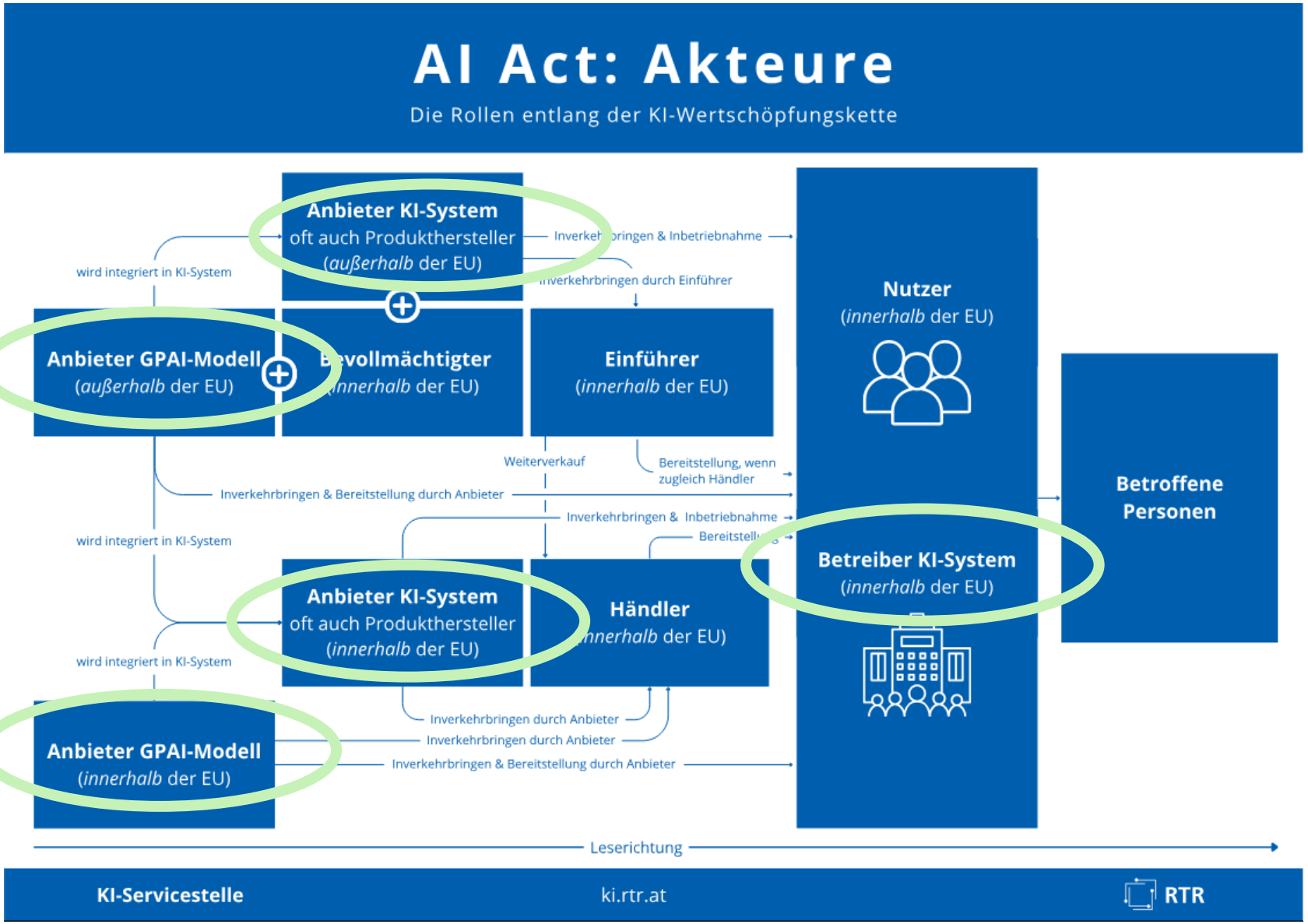
- **Unterschwellige Manipulation** mit **erheblichem Schaden**
- **Ausnutzen** vulnerabler Personengruppen
- Schlechterstellung durch **Social-Scoring**
- **Profiling** zur Beurteilung, ob eine Straftat begangen werden wird
- Erstellung von Datenbanken zur **Gesichtserkennung** aus Bildern aus dem Internet
- **Emotionserkennung** am Arbeitsplatz oder in Bildungseinrichtungen
- **Biometrische Kategorisierung** zur Gewinnung sensibler Daten
- **Biometrische Fernidentifikation** (mit strengen Ausnahmen für Strafverfolgungszwecke)

Hochrisiko-KI-System



Prüfschema Hochrisiko-KI-System (Art 6 KI-VO)

Wichtige Akteure



Rundfunk und Telekom Regulierungs-GmbH (RTR-GmbH), Die Akteure im AI Act, https://www.rtr.at/rtr/service/ki-servicestelle/ai-act/akteure.de.html CC BY 4.0.

KI ≠ KI

AI Act: Verpflichtungen von Betreibern

Der Umfang der Verpflichtungen nimmt entsprechend der Risikoklassifizierung des KI-Systems ab

	Hochrisiko KI-System	**KI-System** begrenzt. Risiko	**KI-System** minimal. Risiko
KI-Kompetenz	Art. 4	Art. 4	Art. 4
Transparenz gegenüber nachgelagerten Akteuren	Art. 26 (11)	Art. 50 (3), (4)	
Verwendung des KI-Systems laut Betriebsanleitung	Art. 26 (1), (3), (4)		
Menschliche Aufsicht	Art. 26 (2)		
Überwachung des KI-Systems	Art. 26 (5)		
Meldung von schwerwiegenden Vorfällen	Art. 26 (5), 73		
Aufbewahrung von erzeugten Protokollen	Art. 26 (6)		
Sofern relevant, Datenschutz-Folgenabschätzung	Art. 26 (9)		
Zusammenarbeit mit zuständigen nationalen Behörden	Art. 26 (12)		
Recht auf Erläuterung der Entscheidungsfindung im Einzelfall	Art. 86 (1)		
Informationspflichten gegenüber der Arbeitnehmer:innen-Vertretung *sofern Arbeitgeber:in Hochrisiko-KI-Systeme am Arbeitsplatz einsetzt*	Art. 26 (7)		
Registrierungspflicht *sofern EU-Organe, EU-Einrichtungen und sonstige EU-Stellen*	Art. 26 (8), 49		
Genehmigungspflicht einer Justiz- oder Verwaltungsbehörde *sofern Einsatz zur nachträglichen biometrischen Fernidentifizierung*	Art. 26 (10)		
Erstellung einer Grundrechte-Folgenabschätzung *sofern u. a. öffentl. oder private Einrichtungen öffentliche Dienste erbringen*	Art. 27		

KI-Servicestelle ki.rtr.at **RTR**

KI ≠ KI

AI Act: Verpflichtungen von Anbietern

Der Umfang der Verpflichtungen nimmt entsprechend der Risikoklassifizierung des KI-Systems/KI-Modells ab

	Hochrisiko KI-System	**GPAI-Modell** system. Risiko	**GPAI-Modell**	**KI-System** begrenzt. Risiko	**KI-System** minimal. Risiko
KI-Kompetenz	Art. 4	Art. 4	Art. 4	Art. 4	Art. 4
Transparenz gegenüber nachgelagerten Akteuren	Art. 13	Art. 55 (1)	Art. 53 (1) b	Art. 50 (1), (2)	
Anforderungen an Daten	Art. 10	Art. 55 (1)	Art. 53 (1) c, d		
Technische Dokumentation	Art. 11	Art. 55 (1)	Art. 53 (1) a		
Zusammenarbeit mit Behörden	Art. 21	Art. 55 (1)	Art. 53 (3)		
Bennenung Bevöllmächtigter (sofern Drittstaat)	Art. 22	Art. 55 (1)	Art. 54		
Risikomanagement	Art. 9	Art. 55 (1) a, b			
Genauigkeit, Robustheit und Cybersicherheit	Art. 15	Art. 55 (1) d			
Registrierungs- bzw. Mitteilungspflichten	Art. 49	Art. 52 (1)			
Meldepflichten gegenüber Behörden	Art. 73	Art. 55 (1) c			
Aufzeichnung von Ereignissen	Art. 12				
Implementierung menschlicher Überwachungstools	Art. 14				
Kennzeichnungspflichten	Art. 16 b				
Sicherstellung der Barrierefreiheitsanforderungen	Art. 16 l				
Qualitätsmanagement	Art. 17				
Aufbewahrungspflichten	Art. 18, 19				
Korrekturmaßnahmen	Art. 20				
Konformitäts-Bewertungsverf., -Erklärung, -Kennzeichnung	Art. 43, 47, 48				

KI-Servicestelle ki.rtr.at RTR

Was ist „künstliche Intelligenz“?

- Ist der AMS-Algorithmus ein KI-System? – Siehe dazu: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus

1:Bildquelle: ITA, ÖAW, Der AMS-Algorithmus (https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus).

Was ist „künstliche Intelligenz“?

- Ist der AMS-Algorithmus ein KI-System? – Siehe dazu: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus

1:Bildquelle: ITA, ÖAW, Der AMS-Algorithmus (https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus).

Was ist „künstliche Intelligenz“?

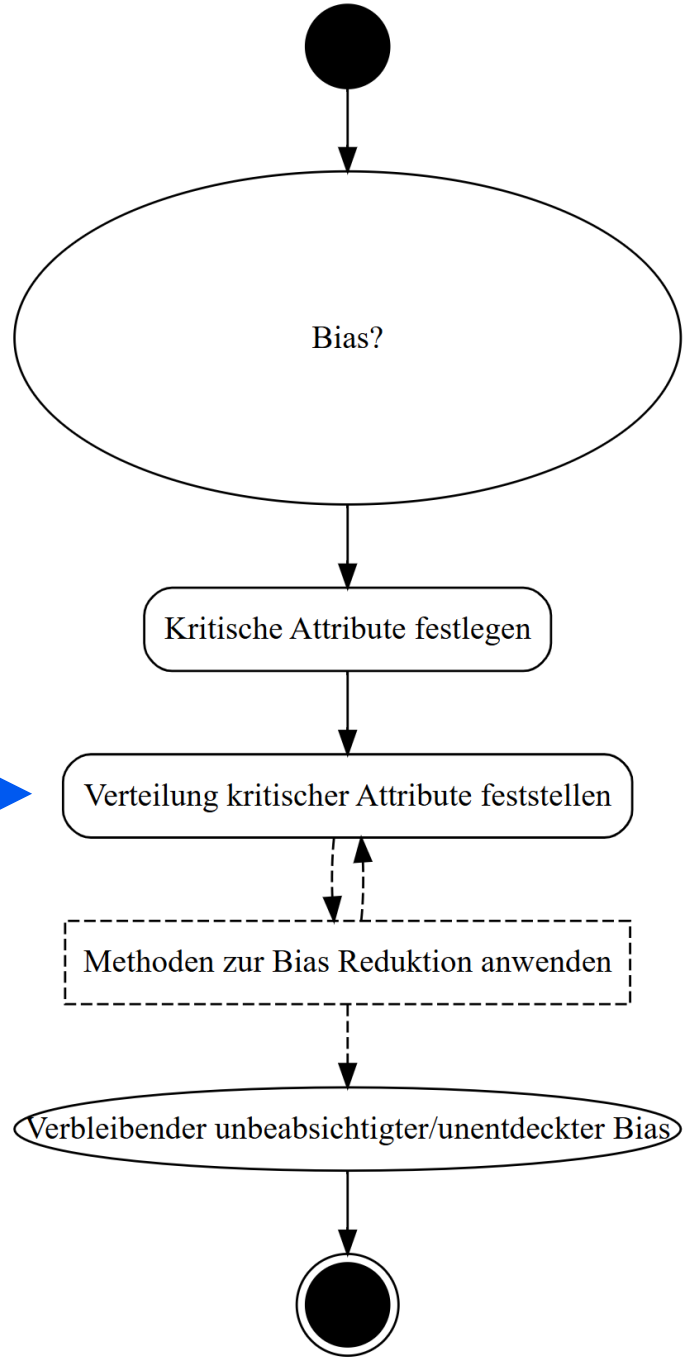
- Ist der AMS-Algorithmus ein KI-System? – Siehe dazu: https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus

1:Bildquelle: ITA, ÖAW, Der AMS-Algorithmus (https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus).

Wie mit Bias umgehen?

Besondere Herausforderung

„Rest Bias“ ist geringer, umso mehr diverse Expertise zur Verfügung steht!



Feststellung und Reduzierung von Bias

Was ist „künstliche Intelligenz“?

„Systeme zur Verbesserung der mathematischen Optimierung oder zur Beschleunigung und Annäherung traditioneller, gängiger Optimierungsmethoden wie lineare oder logistische Regressionsmethoden fallen nicht in den Anwendungsbereich der Definition eines KI-Systems.“[2,3]

ITA, ÖAW, Der AMS-Algorithmus (https://www.oeaw.ac.at/ita/projekte/der-ams-algorithmus).
2: *EK*, Leitlinien KI-Verordnung, 29.07.2025.
3: *Piska/Seeber*, (K)eine Strafe ohne Gesetz: Der AI Act als rechtliches Pulverfass, ecolex 2025/123.

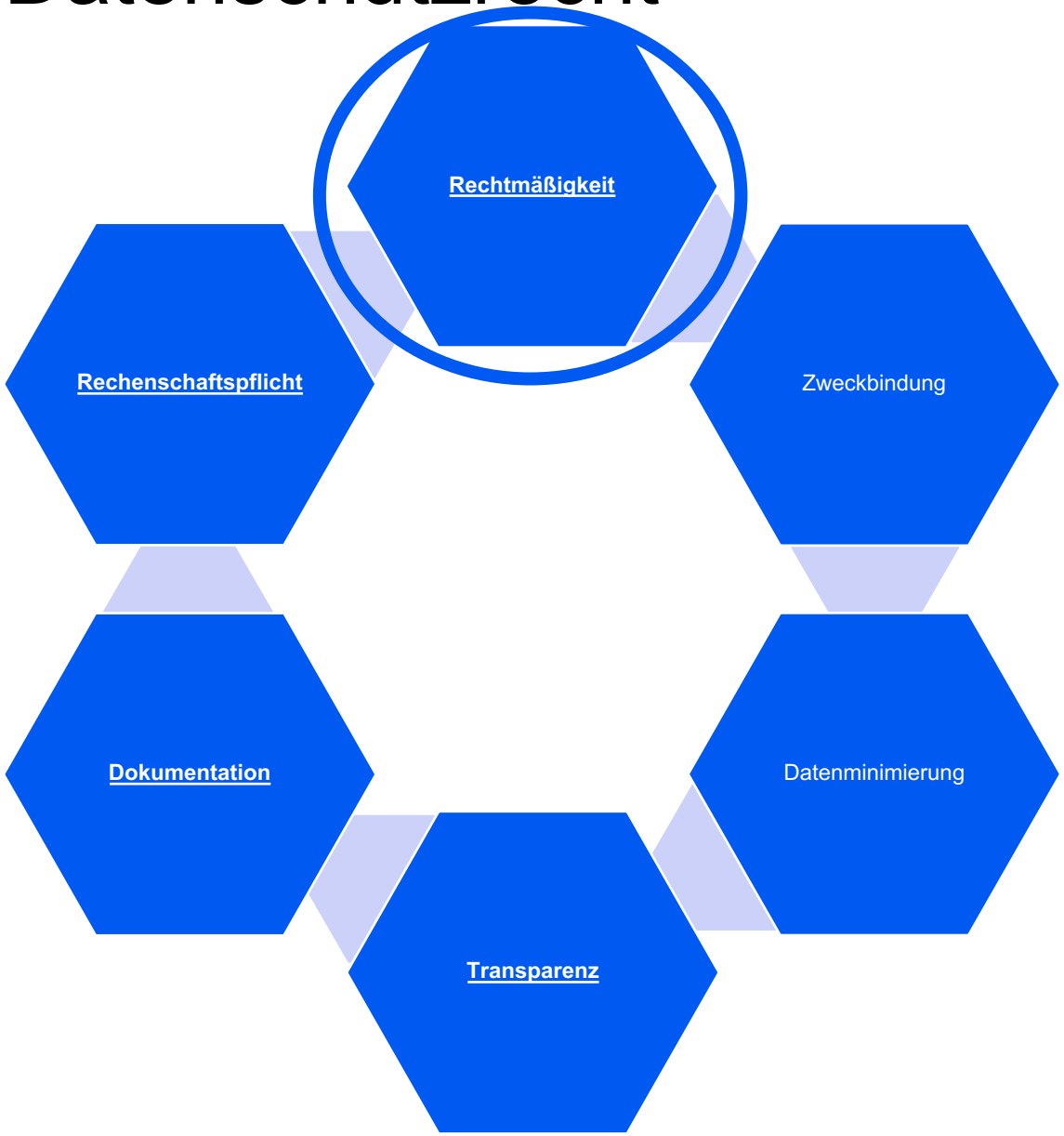
Legal Landscape

Datenschutzrecht

"Data can be either useful or perfectly anonymous but never both."[1]

1:*Ohm*, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review 2010, 1701.

Datenschutzrecht



Datenschutzrecht ist geprägt vom Verbotsprinzip!

– Alles ist verboten, es sei denn, es ist ausnahmsweise erlaubt;

Erleichtertes KI-Training?

…in Diskussion!

Kommissionsvorschlag (Digital Omnibus):

- Erlaubnis besondere Kategorien von Daten (unter gewissen Bedingungen) zur Entwicklung und dem Betrieb von KI-Systemen oder dem Training eines KI-Modells zu nutzen

- Das Entwickeln von KI-Systemen oder das Trainieren von KI-Modellen soll (mit gewissen Einschränkungen) ein berechtigtes Interesse darstellen und gestattet sein.

Urheberrecht

Urheber haben das ausschließliche Recht, ihre Werke auf gesetzlich geregelte Arten zu verwerten - Verwertungsrecht.

1: Vgl. §§ 14, 60 UrhG.

Urheberrecht(e)

- Vervielfältigungsrecht
- Verbreitungsrecht
- Senderecht
- Zurverfügungstellungsrecht
- etc.

Urheberrecht(e) – Text- und Data-Mining (TDM)

- Vervielfältigung mit dem Ziel der Gewinnung von Informationen über Muster, Trends und Korrelationen;

- Unterscheidung zwischen Forschungseinrichtungen und „Jedermann“
  - Jedermann: Nur, wenn Vervielfältigung nicht ausdrücklich durch Nutzungsvorbehalt verboten wurde.

TDM auf KI-Training anwendbar?

Vgl. *Schippel*, „Urheberrecht first!“ - Die Pflicht zur KI-Policy nach Art. 53 KI-VO, KuR 2025, 547.

Risiken von KI

Beispiel zur Veranschaulichung urheberrechtlicher Herausforderungen bei der Nutzung generativer KI.

Das Urheberrecht an Werken der Literatur, der Tonkunst und der bildenden Künste endet siebzig Jahre nach dem Tod.

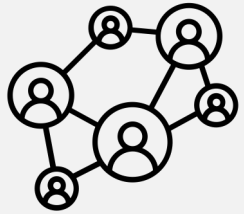
Nutzer haften (idR) für verwendete Ergebnisse!

Vgl. AG Köln: Anwaltliche Verantwortung für KI-generierte Schriftsätze, Computer und Recht 2025, 621–621.

Why do we need AI Factory Austria?

Sovereignty

Ethics and Trustworthiness



Connecting the Ecosystem

Contact

Pia Weinlinger
AI Factory Austria AI:AT

+43 664 78588124
pia.weinlinger@ai-at.eu

Michael Löffler
AI Factory Austria AI:AT

+43 664 88390692
michael.loeffler@ai-at.eu

AI Factory Austria AI:AT
Schwarzenbergplatz 2
1010 Wien, Austria

info@ai-at.eu
training@ai-at.eu
ai-at.eu

@ai-factory-austria

Q & A

Funded by

EuroHPC Joint Undertaking

Funded by the European Union

under discussion with

Federal Ministry
Innovation, Mobility
and Infrastructure
Republic of Austria

FFG
Promoting Innovation.

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).