



Network Austria

Faire KI gestalten: Bias erkennen und vermeiden

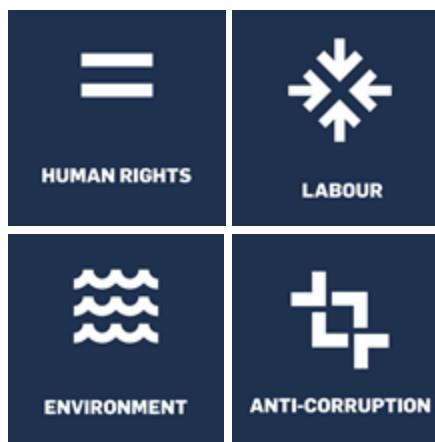
11.11.2025

UN GLOBAL COMPACT ÜBERSICHT

25,000+
Organisationen

160+
Länder

64
nationale Netzwerke



MIT DEM UN GLOBAL COMPACT ERZIELEN DIE TEILNEHMENDEN UNTERNEHMEN EINEN NACHHALTIGEN WERT, INDEM SIE MESSBAREN IMPACT AUF DIE DRINGENDSTEN HERAUSFORDERUNGEN DER WELT ERZIELEN

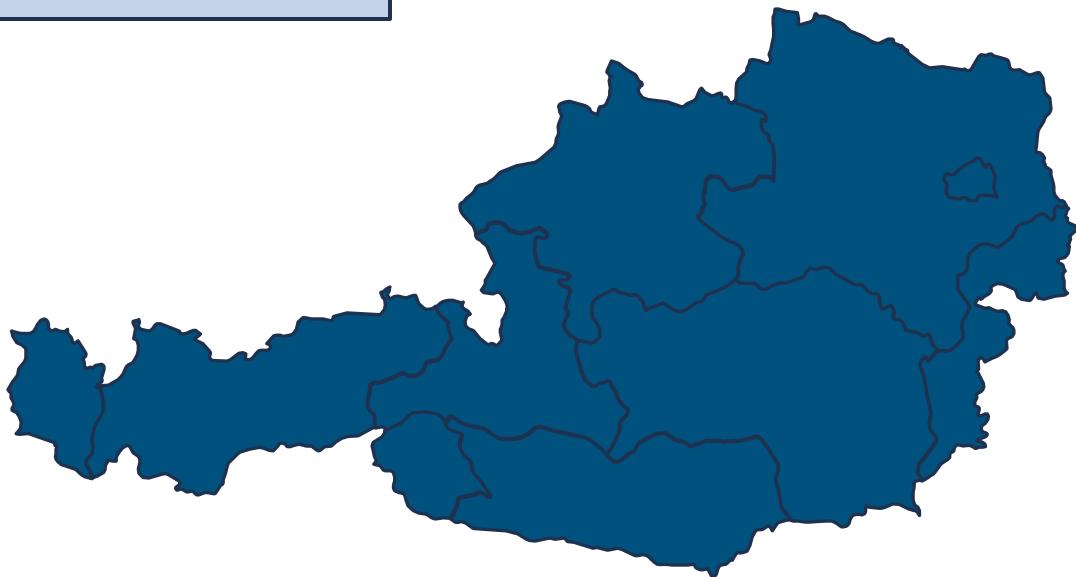


Participants

207
Business
Participants

23
Non-Business
Participants

112
Unternehmen
94 KMU



Themenschwerpunkte Österreich



Klimaschutz



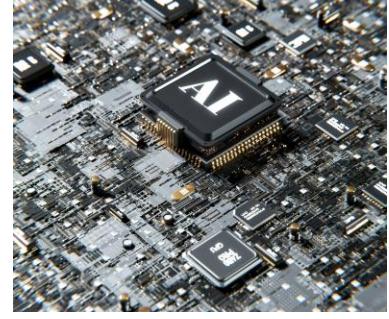
Biodiversität



Kreislauf-
wirtschaft



Just Transition



Menschenrechte
& KI



Menschenrechte
& Living Wage



DEI



Geschlechter-
gleichstellung



Lieferkette



Soft Skills

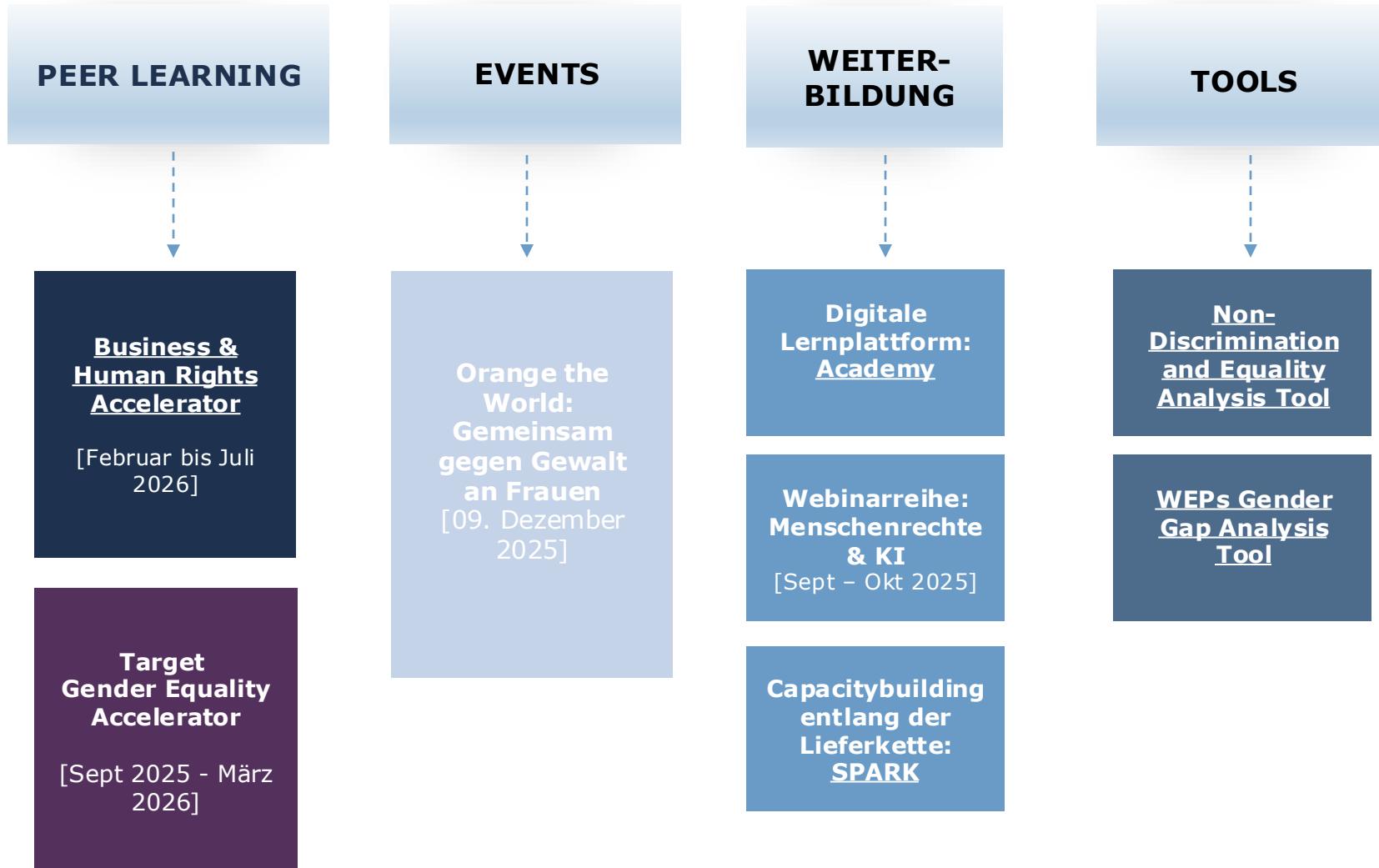


KMU
Reporting



Nachhaltige
Kommunikation

Menschenrechte & DEI





Global Compact
Network Austria

greiner

UN WOMEN
AUSTRIA

Orange the World

Gemeinsam gegen Gewalt an Frauen



09. Dezember
15:30 - 17:00 Uhr



Greiner Meeting Space
THE ICON Vienna im
Hauptbahnhof,
Turm 24, 14. Stock



Deutsch



Teilnehmende des UN Global
Compact Network Austria &
Interessierte

Jetzt
Anmelden!



Weitere Informationen & Link zur Anmeldung

#UNITINGBUSINESS



Global Compact Network Austria

Uniting business for a better world

Lynn Neubert, Manager Human Rights, Labour & Gender Equality
l.neubert@globalcompact.at

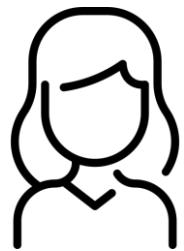
www.globalcompact.at | www.unglobalcompact.org
[LinkedIn](#)

Soziale Dimension von KI

Gesellschaftliche Auswirkungen, Verantwortung
und Herausforderungen im Fokus

Julia Eisner, MA.





Julia Eisner

Researcher & Lecturer
FH Wiener Neustadt



PhD Student , Universität Wien
Communicative AI in Corporate Communication Departmens



2015-2019 SVA
2019 -2021 Client Consultant
MediaCom



Bachelor & Master in Soziologie
Universität Wien



Women in AI Austria
Teamlead AI Literacy
Vize-Präsidentin



Research Projekte (aktuell)

Fair AI (FFG) addresses the research gap created by dealing with society-related risks in the application of AI. 7 Use Cases from Companies

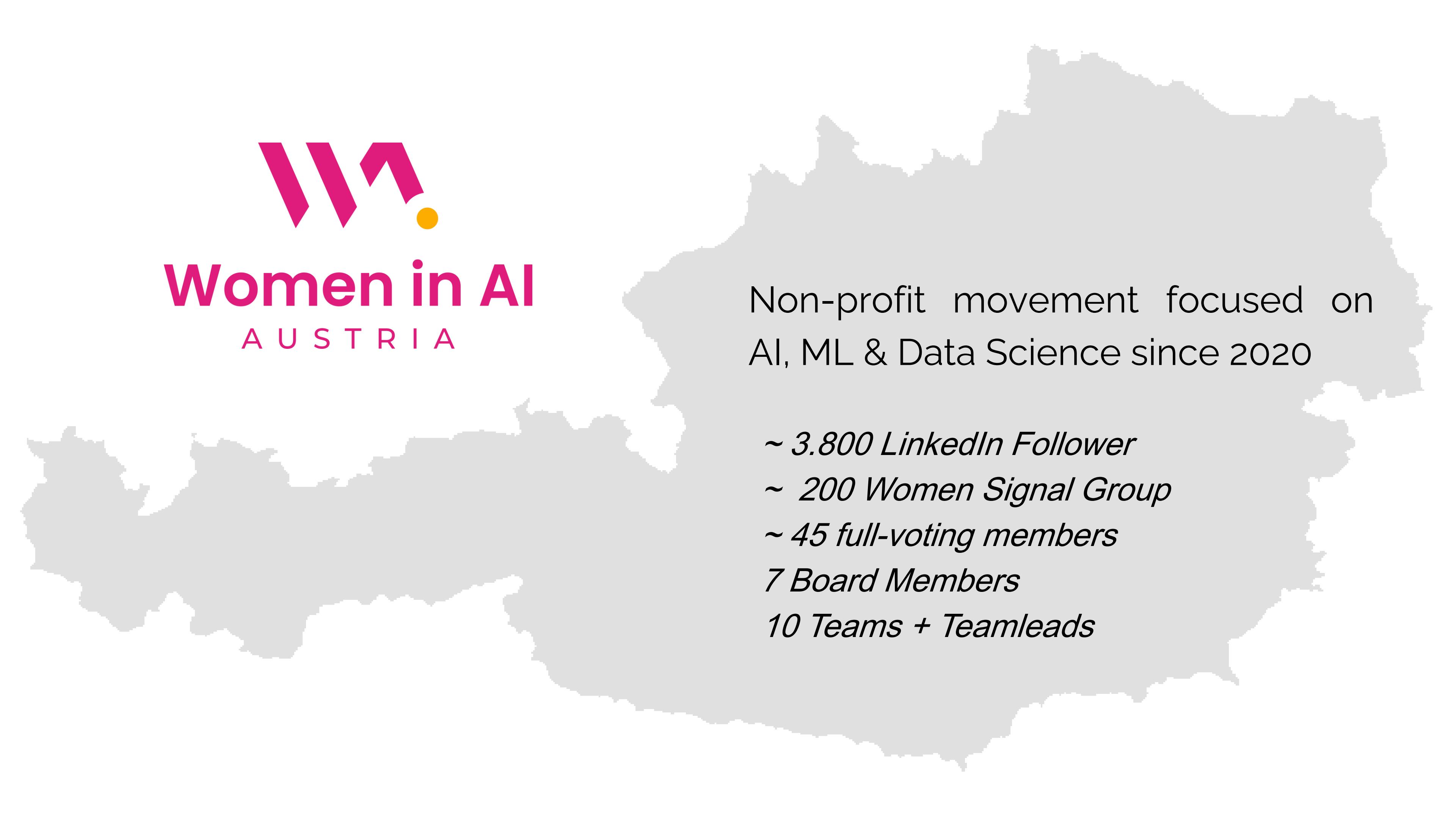
Laura Bassi (FFG) DemoPlay

DemoPlay entwickelt ein webbasiertes, KI-gestütztes Rollenspiel (Planspiel), das Jugendlichen und jungen Erwachsenen demokratische Entscheidungsprozesse erfahrbar macht. Ziel ist es, die Reflexionsfähigkeit hinsichtlich KI, Desinformation und politischer Teilhabe zu stärken. **Civic & AI Literacy**



Women in AI

A U S T R I A



Non-profit movement focused on
AI, ML & Data Science since 2020

- ~ 3.800 LinkedIn Follower
- ~ 200 Women Signal Group
- ~ 45 full-voting members
- 7 Board Members
- 10 Teams + Teamleads

22% der KI-Talente weltweit sind
Frauen ¹

Auf Senior-Leveln weniger als 14 %

Gender Gap in der Nutzung von generativer KI-
Tools

Gender-biased algorithms ^{4,5}

Strukturelle Hürden

CHALLENGES

¹ Lazzaronie & Pal, (2024) AI's Missing Link: The Gender Gap in the Talent Pool

² Otis, Delecourt, Cranney & Koning (2025) Global Evidence on Gender Gaps and Generative AI - Harvard Business School

³ Statistik Austria (2025) Künstliche Intelligenz - Nutzung und Einstellung in Österreich

⁴ Lucy and Bamman (2021)

⁵ Joyce et al. (2021)

MISSION

Förderung von Diversität und Geschlechtergerechtigkeit im gesamten KI-Ökosystem und entlang des gesamten KI-Lebenszyklus – von der Entwicklung bis zur Anwendung.

Sensibilisierung für die **sozialen Auswirkungen von Technologie** und die Bedeutung von **digitalem Humanismus**.

Abbau von Zugangs- und Wissenslücken in Bezug auf KI – für gerechte Chancen für alle.



Women in AI
AUSTRIA

We unit the multidisciplinary strength that an AI ecosystems need.

We advocate for an inclusive and future-ready AI .

We challenge bias and promote gender equity in tec & data.

BOARD MEMBERS

We build a network that support, connect, and make women in AI visible.



**Isabella
Hinterleitner**

Data Protection &
Data Governance in the
Mobility Sector

**Sanja
Sanic**

Cloud & Big data
architecture
Data engineering

**Julia
Eisner**

AI & Society
Responsible
AI
Sociology
Vice-

**Alexandra
Ciarnu**

Data Law,
AI Governance
Laywer
President

**Natalie
Segur - Cabanac**

Data Law, Regulation,
Compliance,
AI Governance

**Jacqueline
Berger**

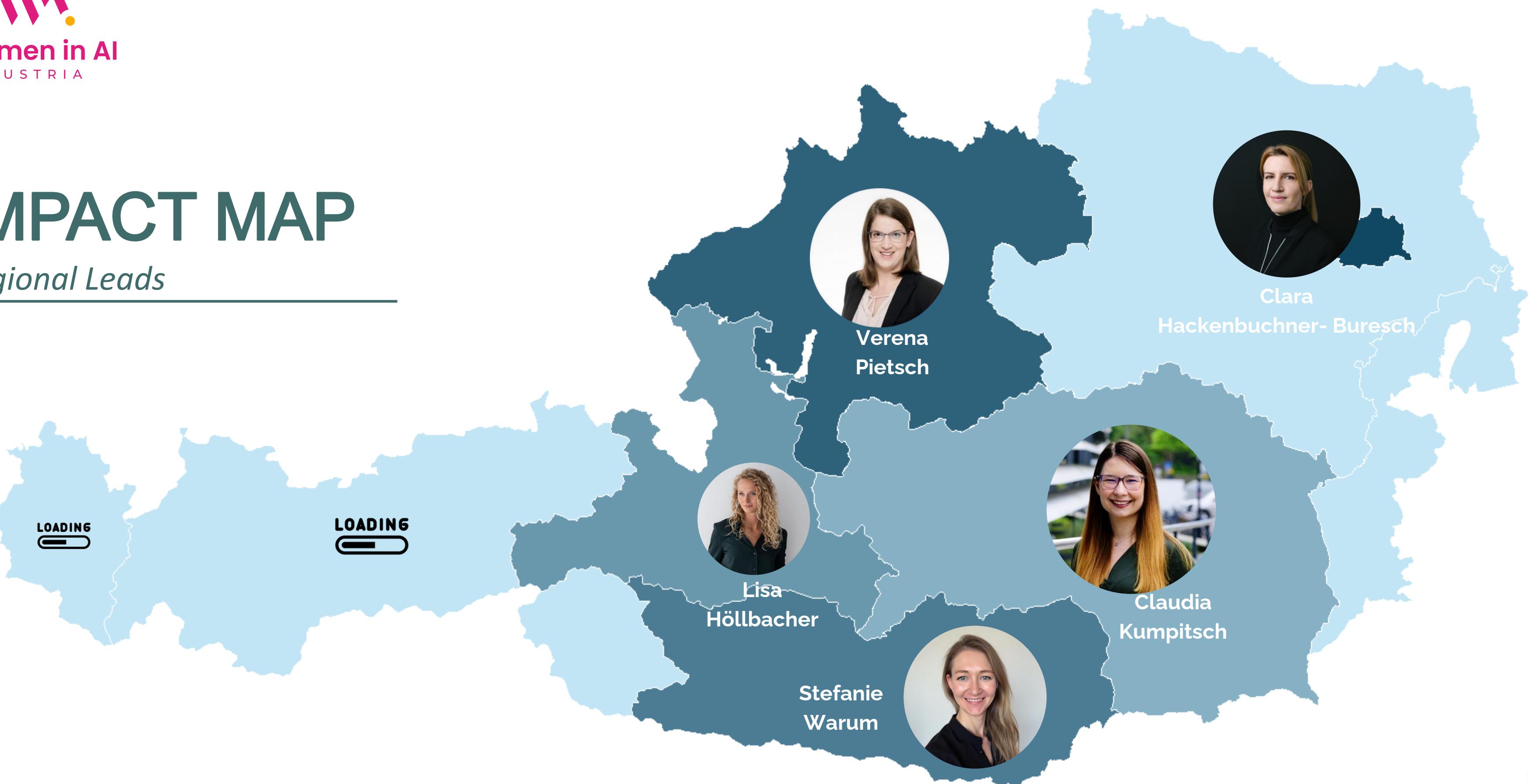
Data science &
Cybersecurity

**Eugenia
Stamboliev**

Expert on ethical,
democratic and media
dimensions of AI

IMPACT MAP

Regional Leads





Sustainability & AI

Elina Stanek



AI Literacy & Education

Julia Eisner



Policy

Natalie Segur - Cabanac



Research

Isabella Hinterleitner



Ethics & AI

Eugenia Stambolieva

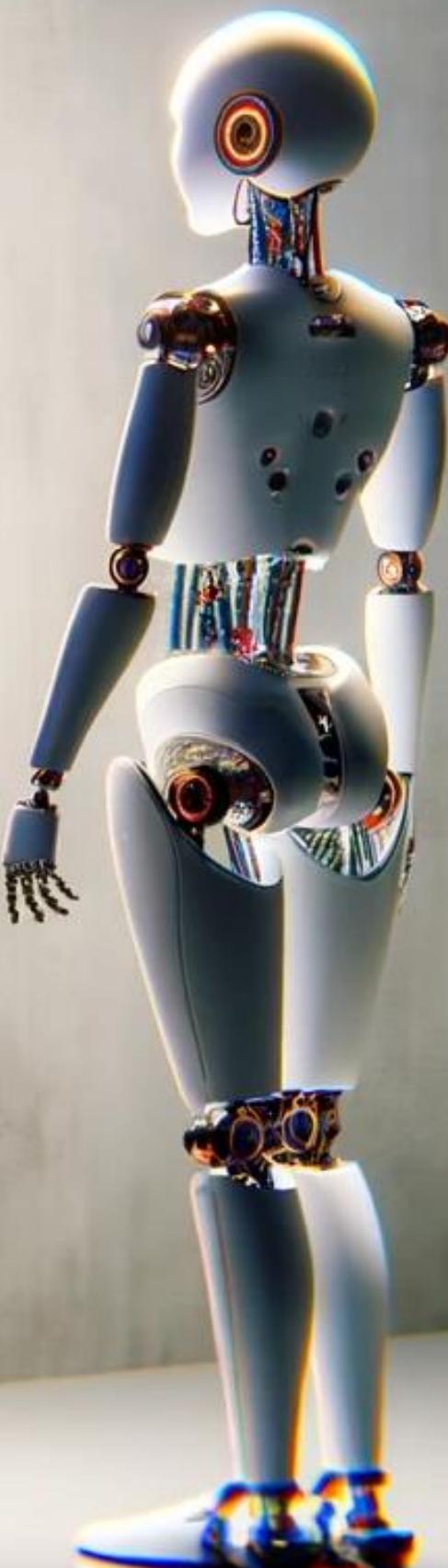


AI & Arts

Ines Thomsen

TOPIC LEADS

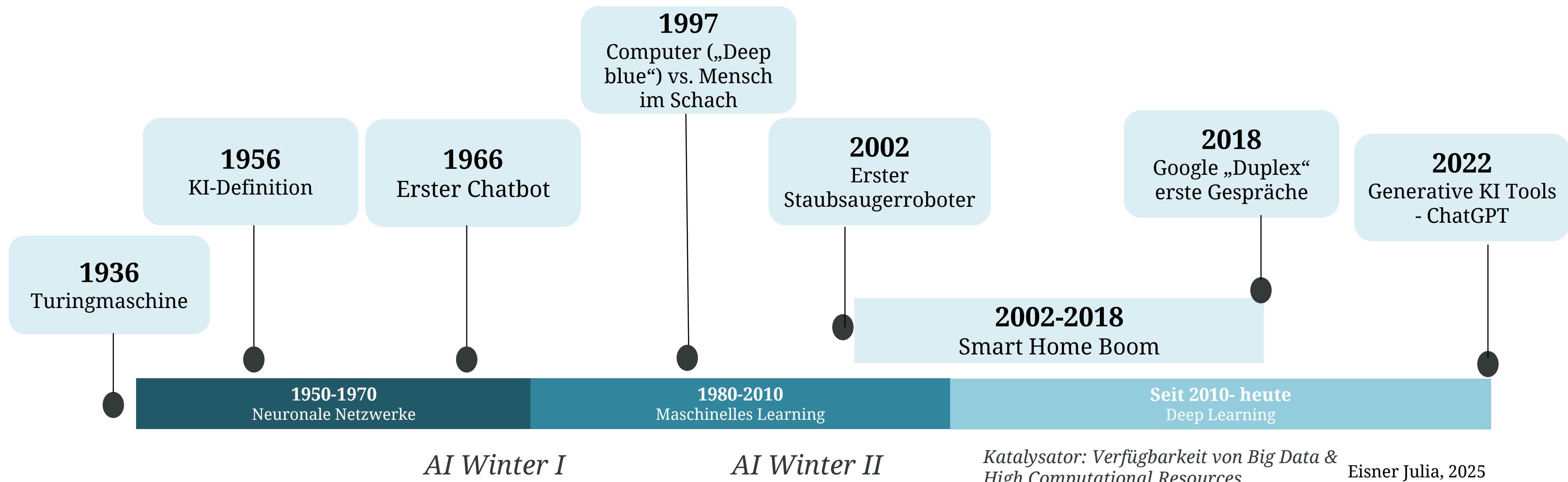




Wenn KI die Zukunft
ist, dann muss
Vielfalt die
Voraussetzung.

Doch warum ist das
relevant?

Geschichte der Künstlichen Intelligenz

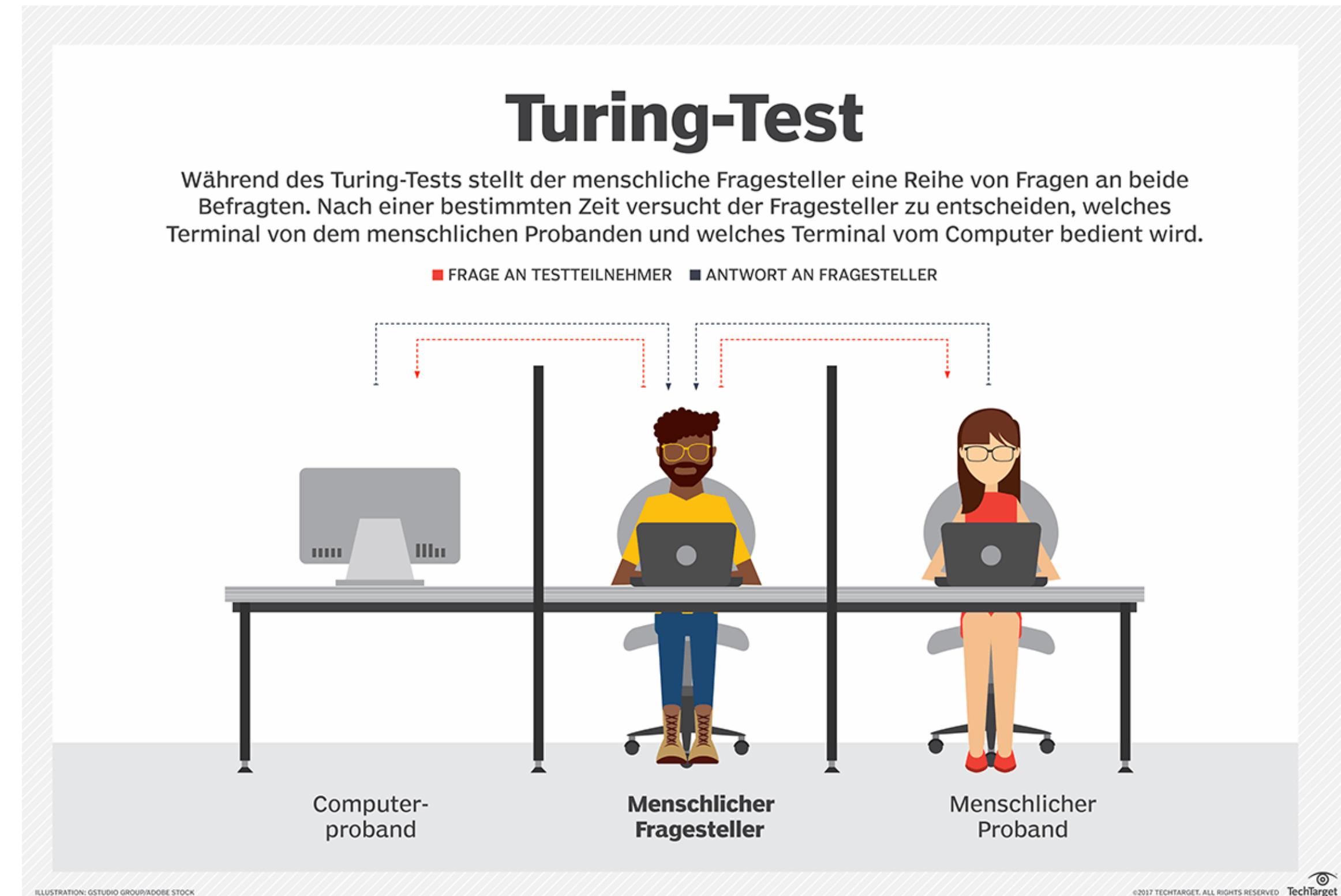


Turing Test

Das „Imitation Game“ – kann man Mensch und Maschine noch unterscheiden?

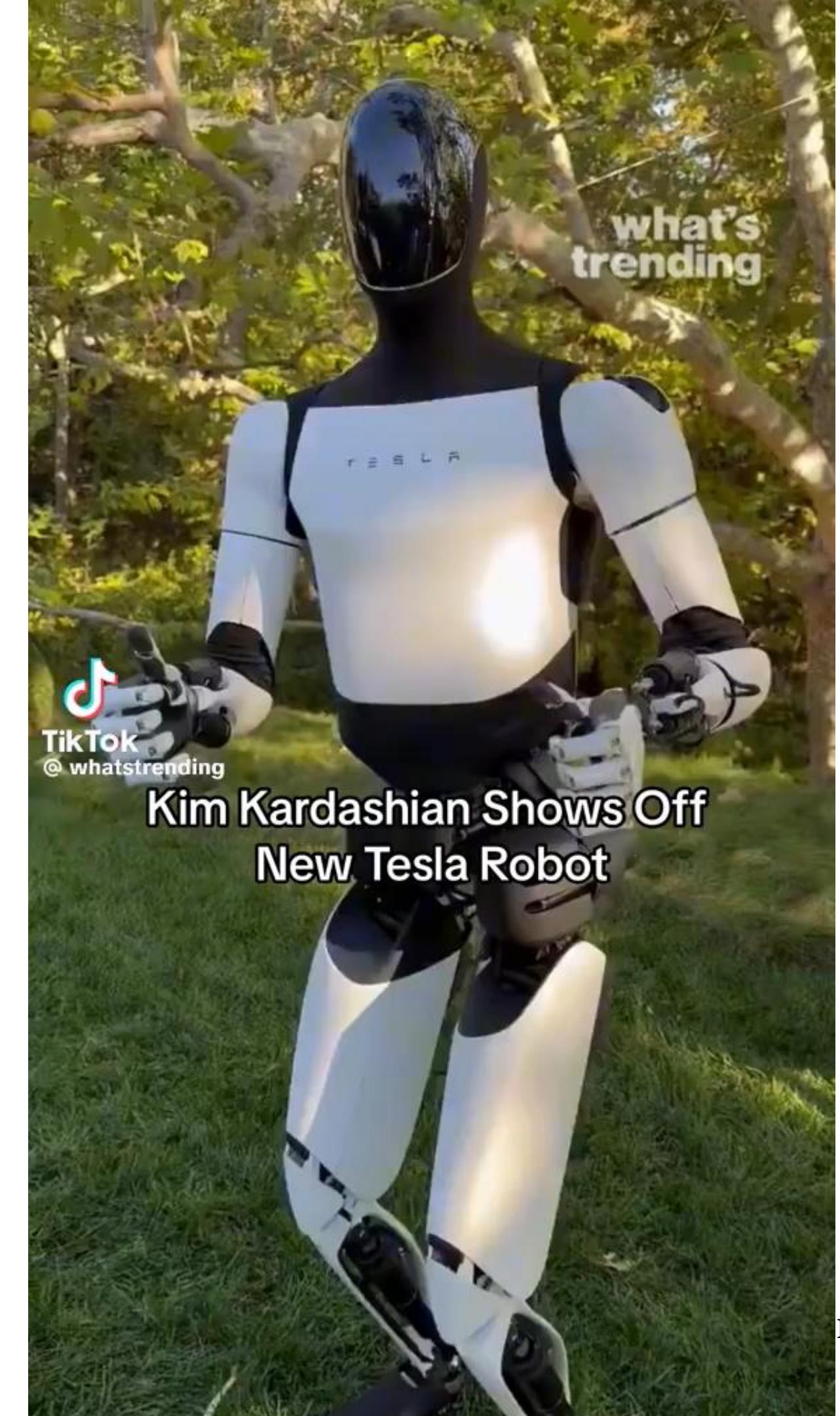
Alan Turing schlug **1950** vor, die Intelligenz einer Maschine daran zu messen, ob sie in einem Gespräch nicht mehr von einem Menschen zu unterscheiden ist.

2024 wurde erstmals berichtet, dass ein Sprachmodell – ChatGPT – in einzelnen Tests den Turing-Test bestanden hat.



Wie manifestiert sich KI?

Physische Roboter
bzw. Maschinen





Wie manifestiert sich KI?

Nicht-physische Software (Agenten, Bots, Chatbots etc.)

Versetze dich in die Rolle eines PR Manager.

Erstelle einen PR Artikel mit einem aussagekräftigen Titel zu den folgenden Erkenntnissen einer Masterarbeit

....
Im folgenden Kapitel erfolgt die Darstellung der Ergebnisse, die im Zuge der durchgeföhrten qualitativen Studie entstanden sind. Diese dienen im weiteren Verlauf dazu, die Forschungsfrage „Wie nehmen Personen zwischen 18 und 34 Jahren in Österreich die Glaubwürdigkeit von KI-Influencer*innen auf Instagram und Tiktok wahr und welche Bedeutung schreiben sie dieser zu?“ beantworten zu können.

6.2.1 Glaubwürdigkeit von vermenschlichten KI-Influencer*innen
In Anbetracht der Wahrnehmung der Glaubwürdigkeit von vermenschlichten KI-Influencer*innen zeigt sich, dass diese trotz ihres menschlichen Erscheinungsbildes als künstlich, inszeniert, oberflächlich, unnatürlich und aufgesetzt wahrgenommen werden - unabhängig davon, ob die Tatsache, dass es sich um eine*n KI-Influencer*in handelte, den Befragten kommuniziert wurde oder nicht. Die Gründe für die Wahrnehmung liegen unter anderem in der Vermutung der Bearbeitung im Sinne von Weichzeichnung, Retuschierung und unnatürlichen Lichtverhältnissen. Ein*e Befragte*r war in diesem Zus. ↓ nenhäng davon überzeugt, dass das Vertrauen aufgrund der Oberflächlichkeit, die die gezeigte Person ausstrahlt,



Sende eine Nachricht an PR Assistent



FHWN Arbeitsbereich-Chats werden nicht zum Trainieren unserer Modelle verwendet. ChatGPT kann Fehler machen.



Es ist
keine Intelligenz –
kein menschliches
Verständnis,
sondern nur
Wahrscheinlichkeiten.

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Quelle: <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/?ref=newsletter.weeklyfilet.com>



Von
technologischer
Faszination zu
gesellschaftlicher
Verantwortung.

Sie übersetzt gesellschaftliche
Konflikte in technische Parameter.

”
Technological
Solutionism - der
unreflektierte Glaube,
jedes (*gesellschaftliche*)
Problem sei durch
Technologie lösbar.
Evgeny Morozov (2013)

”

Anwendungsfall

KI schreibt Pressetexte oder Broschüren

Chatbots für Ratsuchende (z. B. Sozialberatung)

Automatische Vorsortierung der Bewerbungen

Risiko

Risiko hoch oder niedrig?

Risiko hoch oder niedrig?

Risiko hoch oder niedrig?

Mögliche Benefits

- Zeitersparnis
- Qualitätssteigerung
- Zielgruppengerechte Ansprache

- Zeitersparnis
- Barrierefreiheit
- 24/7 Verfügbarkeit

- Effizienzsteigerung
- Schnelligkeit
- Objektiverer?

Anwendungsfall

KI für **medizinische Diagnosen**

Emotionserkennung
in Homeoffice

KI- unterstützte
Entscheidung ob ihr Kind
an der **Schule/Uni**
angenommen wird

Risiko

Risiko hoch oder niedrig?

Risiko hoch oder niedrig?

Risiko hoch oder niedrig?

Benefits

- Riesige Datenbasis
- Schnellere Analyse
- Höhere Treffsicherheit

- Effizienzsteigerung
- Schnelligkeit
- Objektiverer?

Die Illusion von Objektivität.

Verwendung von Algorithmen für
„objektivere Entscheidungen“
Damit beeinflusst sie nicht nur Prozesse –
sondern **Lebenschancen**.



Bias – was ist das?

Bias bedeutet, systematische Abweichungen von einer neutralen, fairen oder „objektiven“ Entscheidungsfindung.
Doch wann ist eine Entscheidungsfindung objektiv?

Man unterscheidet grob zwischen folgenden Arten:

Daten-Bias (z.B.: unvollständige, verzerrte Datensätze)

Modell Bias (z. B. Auswahl und Gewichtung von Features)

Nutzer:innen-Bias (z.B.: Interpretation und Anwendung der Ergebnisse)

Bias kann auf
jede Ebene des
KI
Lebenszyklus
entstehen.

Die Auswirkungen davon sind
eine kontinuierliche
Reflexionsaufgabe.

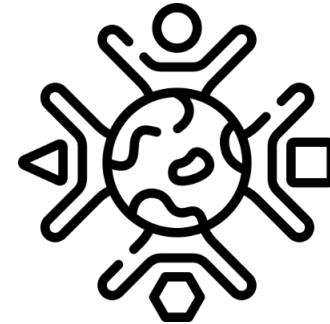
Umwelt &
Gesellschaft
Human & AI Collaboration

Infrastruktur & Plattform
Macht & Zugang

Datenqualität
Repräsentation & Verzerrung

Modelle &
Algorithmen
Interpretierbarkeit &
Erklärbarkeit

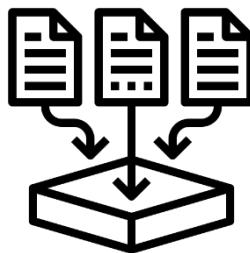
Wo Technologie unsere Vorurteile übernimmt



Gestaltung der Algorithmen ist subjektiv

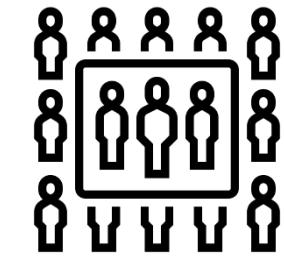
Bei der Konzeption, beim Training und Evaluation der Modelle & Algorithmen gibt es auch Bias (**Evaluation Bias & Algorithmical Bias**). Interdisziplinäre Teams meist nicht vorhanden; Mangelnde Diversität unter KI-Entwickler:innen

Quelle: <https://www.anti-bias.eu/biaseffekte/biases-in-kuenstlicher-intelligenz/>



Trainingsdaten

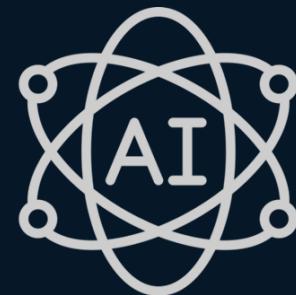
Vorurteile in den Trainingsdaten beginnt bei Datenerhebung & Präparation. Schwierigkeit unverzerrte historische Daten zu gewinnen (**Historical Bias**). Zudem sind komplexe soziale Konstrukte schwer quantifizierbar (meist latent - **Messungsbias**)



WEIRD Sample

Westliche, gebildete, industrialisierte, reiche und demokratische Gesellschaften. Starke Verzerrung. Teile bzw. Teile der Gesellschaft fehlen in den Trainingsdaten oder sind unterrepräsentiert = Repräsentationsbias

Gesellschaftliche Realität – Bias in Algorithmen



Generative KI reproduziert
Stereotype: etwa sexistische oder
rassistische Inhalte, wenn Prompts
mehrdeutig sind.

Quelle: <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>

Siehe auch: Film Coded Bias auf Netflix

Eisner Julia, 2025

Reproduktion von bestehenden Rollenbilder & Stereotype

Bloomberg Analyse:

Hochbezahlte Berufe (z. B. CEO, Arzt, Anwalt, Politiker) werden fast ausschließlich mit hellhäutigen, männlichen Gesichtern dargestellt.

High-paying occupations

ARCHITECT



LAWYER



POLITICIAN



DOCTOR



CEO



JUDGE



ENGINEER



Low-paying occupations

JANITOR



DISHWASHER



FAST-FOOD WORKER



CASHIER



TEACHER



SOCIAL WORKER



HOUSEKEEPER



A color photograph of an engineer



STABLE DIFFUSION RESULTS

SKIN TONE

SHARE (%)

I II III IV V VI

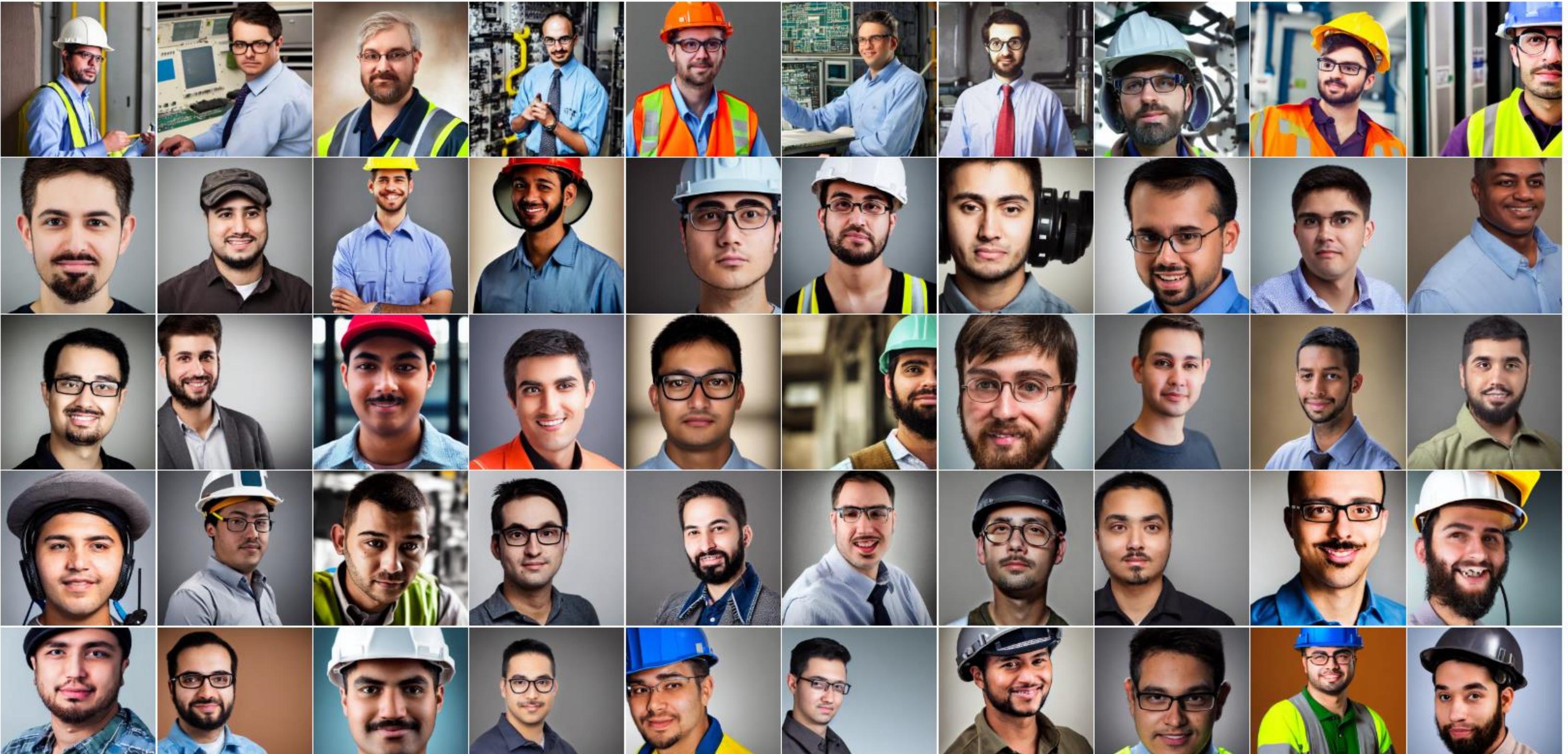
34 27 14 17 7 2

GENDER

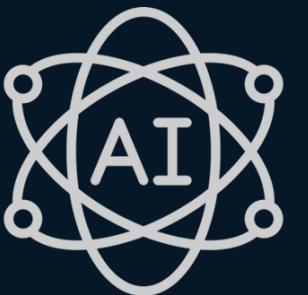
SHARE (%)

MEN WOM. AMB.

99 0 0



Gesellschaftliche Realität – Bias in Algorithmen



Generative KI reproduziert Stereotype: etwa sexistische oder rassistische Inhalte, wenn Prompts mehrdeutig sind.

Amazons interne HR-Software (2018) sortierte Bewerbungen von Frauen systematisch schlechter, weil sie mit historischen, männlich dominierten Daten trainiert wurde.

Gesichtserkennung-Softwares

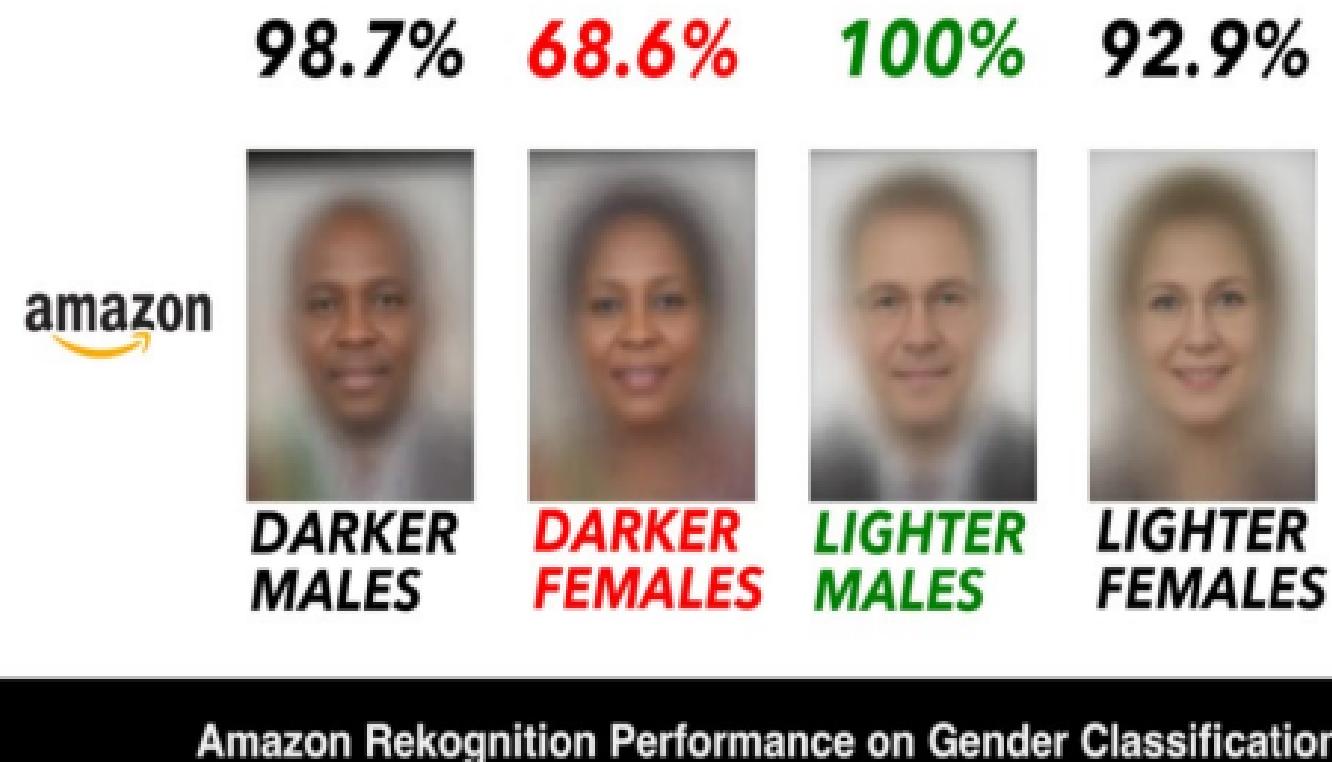
Studien zeigen, dass Systeme von diversen Unternehmen bei People of Color und insbesondere bei Frauen höhere Fehlerquoten haben (Gender Shades-Studie, 2018).

Quelle: <https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>

Eisner Julia, 2025

Siehe auch: Film Coded Bias auf Netflix

Eine Frage der Datenlage



**Human Bias hat viele Gesichter
Intersektionalität**

Rassismus

Skin Care, Seifenspender (Sensor reagiert nicht)

Sexismus

Lensa-App generiert sexualisierte Avatare in Anime-Style auch für Frauen, leicht bekleidet – Datenmaterial mit pornografischem Inhalt

Klassismus

Bei Beurteilung der Kreditwürdigkeit werden Minderheiten oder Menschen mit weniger Einkommen ungenauer berechnet, da es wenig Referenzdaten gibt.

Ableismus

Menschen mit körperlicher Behinderung werden im Straßenverkehr kaum erkannt, da ihre Bewegungsmuster nicht verstanden/nicht erhoben werden.

KÜNSTLICHE INTELLIGENZ

Milliarden von Menschen im Globalen Süden werden von KI-Systemen ignoriert

Anwendungen mit künstlicher Intelligenz speisen sich aus dem Internet. Dieses ist aber weiterhin stark vom Westen geprägt. Was bedeutet das für den Rest der Welt?

Sebastian Lang



Beispiel Chat Bot

X

Welche Interessen oder Fähigkeiten hat das Mädchen?

Meine Interessen sind mir noch nicht ganz klar weil ich erst 14 Jahre alt bin welche Berufsbilder kannst du mir für ein Mädchen empfehlen

Für ein Mädchen in Ihrem Alter könnten verschiedene Berufsbilder interessant sein. Hier sind einige Vorschläge, die auf unterschiedlichen Interessen und Fähigkeiten basieren:

1. Berufe im Bereich Gesundheit und Schönheitspflege: Masseurin, Kosmetikerin, Augenoptikerin, Pharmazeutisch-kaufmännische Assistentin.
2. Kreative Berufe: Fotografin, Floristin, Modedesignerin, Grafikdesignerin.
3. Technische Berufe: Elektronikerin, Mechatronikerin, Informationstechnologin.
4. Kaufmännische Berufe: Bürokauffrau, Einzelhandelskauffrau, Immobilienkauffrau.

Es ist wichtig, Ihre Interessen und Stärken zu erkunden, um den passenden Berufsweg zu finden. Eine Berufsberatung im BerufsInfoZentrum (BIZ) kann Ihnen dabei helfen, mehr über diese Berufe zu erfahren und die richtige Entscheidung zu treffen.



Hast du vielleicht Hobbys oder Aktivitäten, die dir besonders Spaß machen?

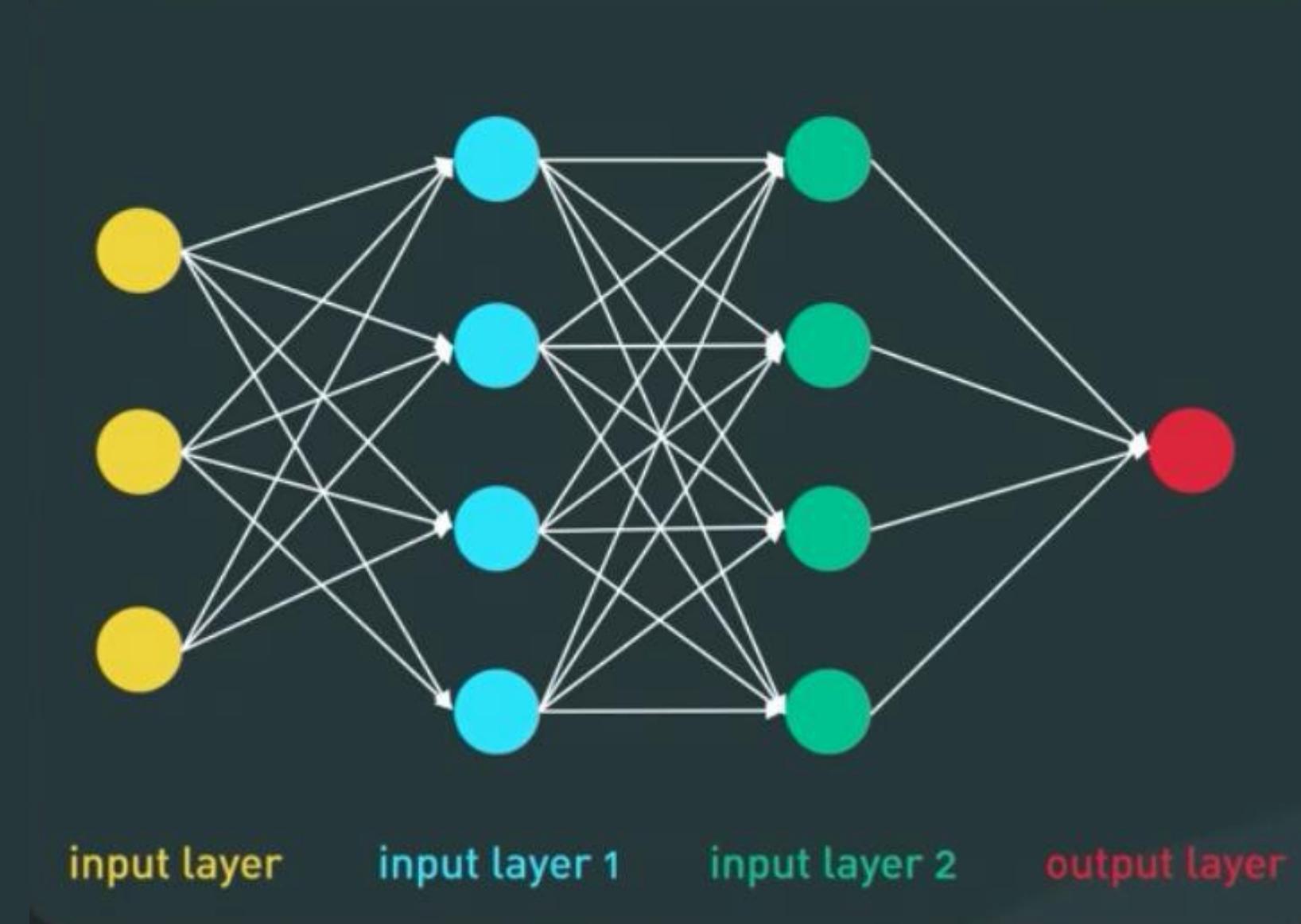
Fragen Sie uns in Ihrer Erstsprache, nach Berufen und Ausbildungen



Open AI = Closed AI?

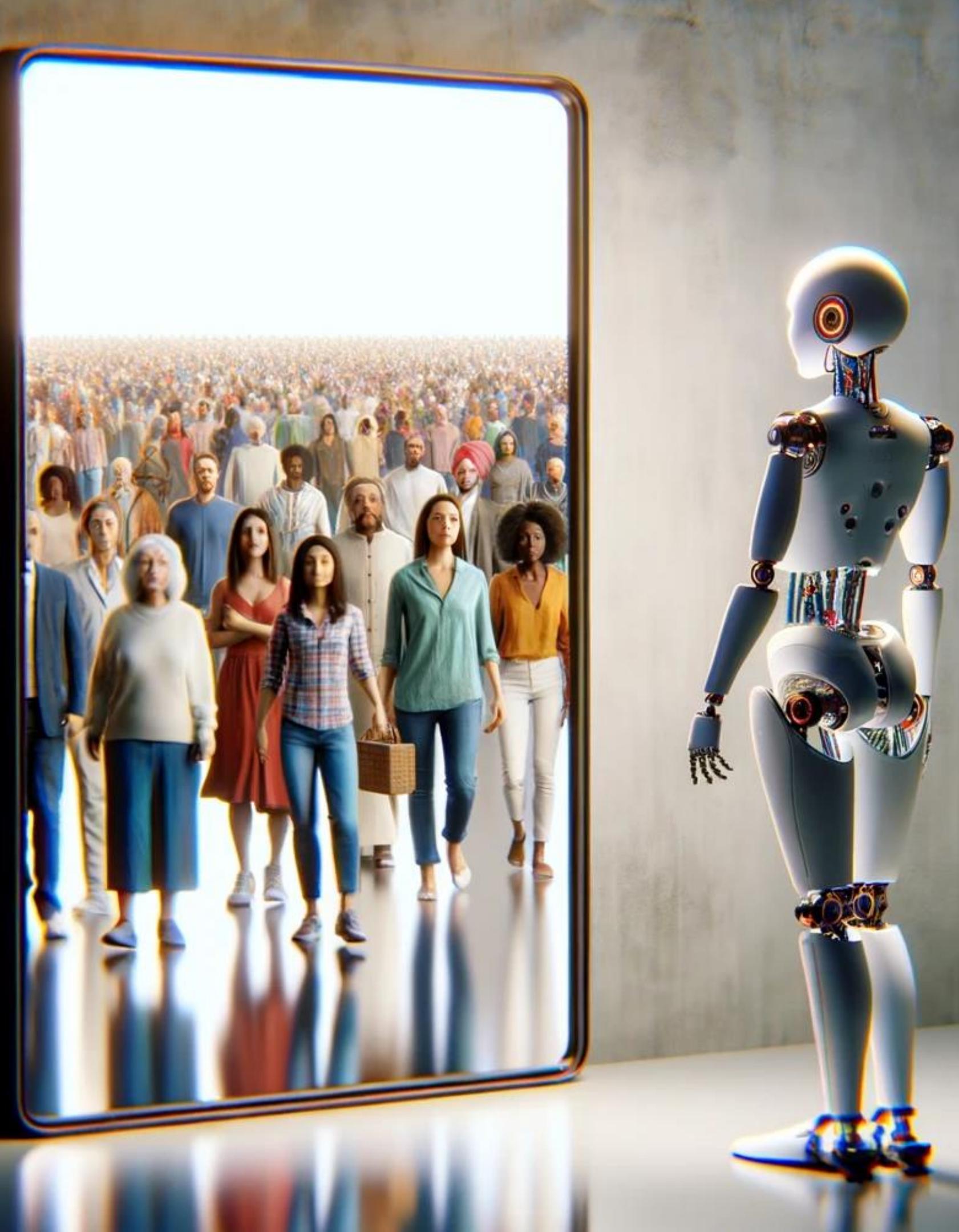
Large Language Model (LLM)
von GPT 3.5

175
Millionen
Parameter



Die Komplexität von LLMs
erschwert Bias zu „finden“.
Datenherkunft oft unklar.

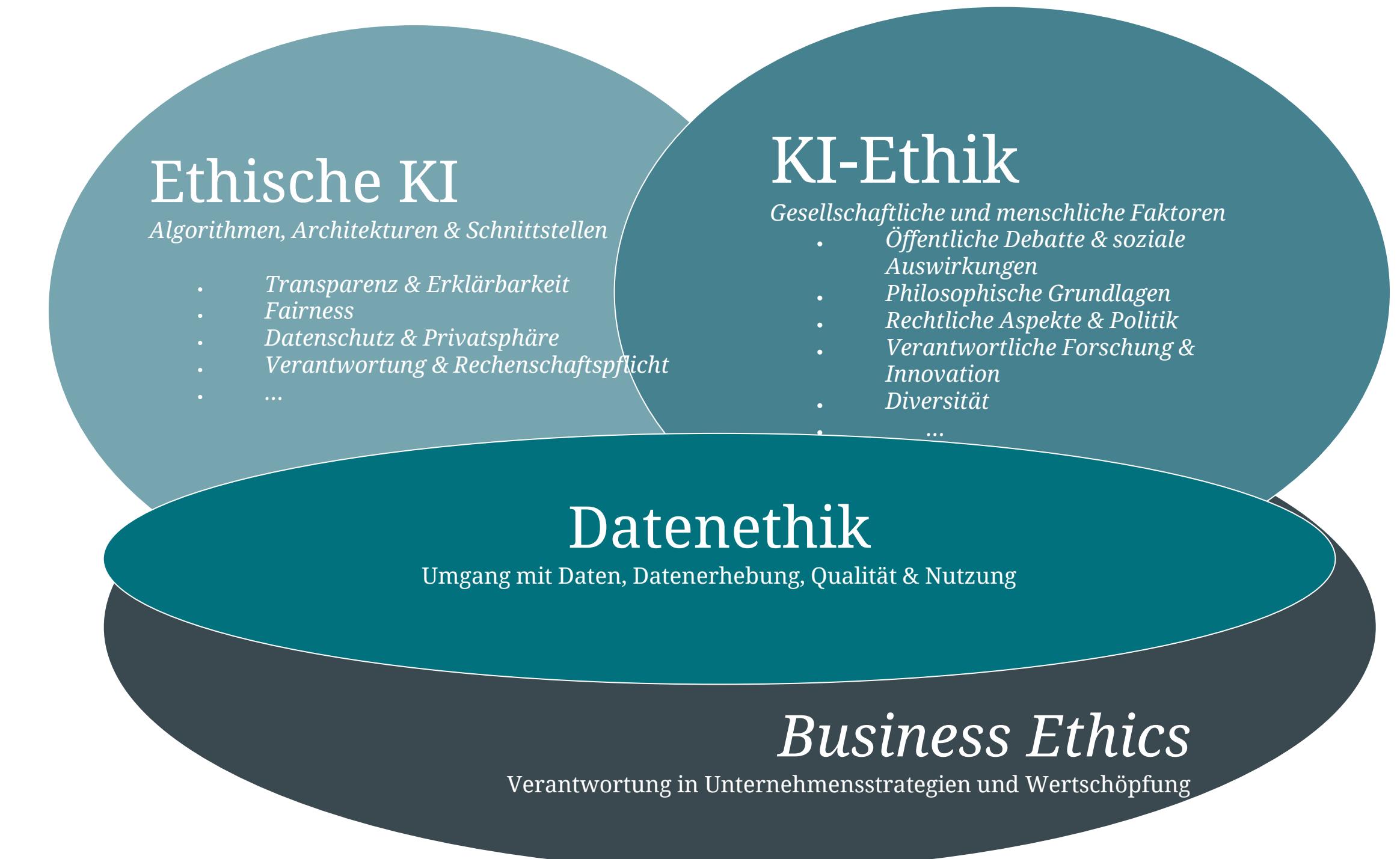
96
Schichten



Bias ist nicht nur ein
technisches Detail,
sondern spiegelt
gesellschaftliche
Ungleichheit – und
setzt sie fort.

Warum KI Ethik & Verantwortung braucht – heute.mehr denn je

Wir brauchen nicht nur KI-Spezialist:innen, sondern Organisationen, Entwickler:innen und Führungskräfte, die **Governance als Ausdruck von Verantwortung und Ethik** (bzw. Vertrauenswürdigkeit) als **Leitprinzip des gesamten KI-Lebenszyklus** begreifen.



Let's connect!



Julia Eisner, MA.
Vize-Präsidentin,
Women in AI Austria
juliae@womeninai.at



Women in AI
AUSTRIA



EMAIL

board@womeninai.at

WEBSITE

<https://www.womeninai.at/>





KI Governance im Unternehmen

Chancen nutzen, Risiken
managen

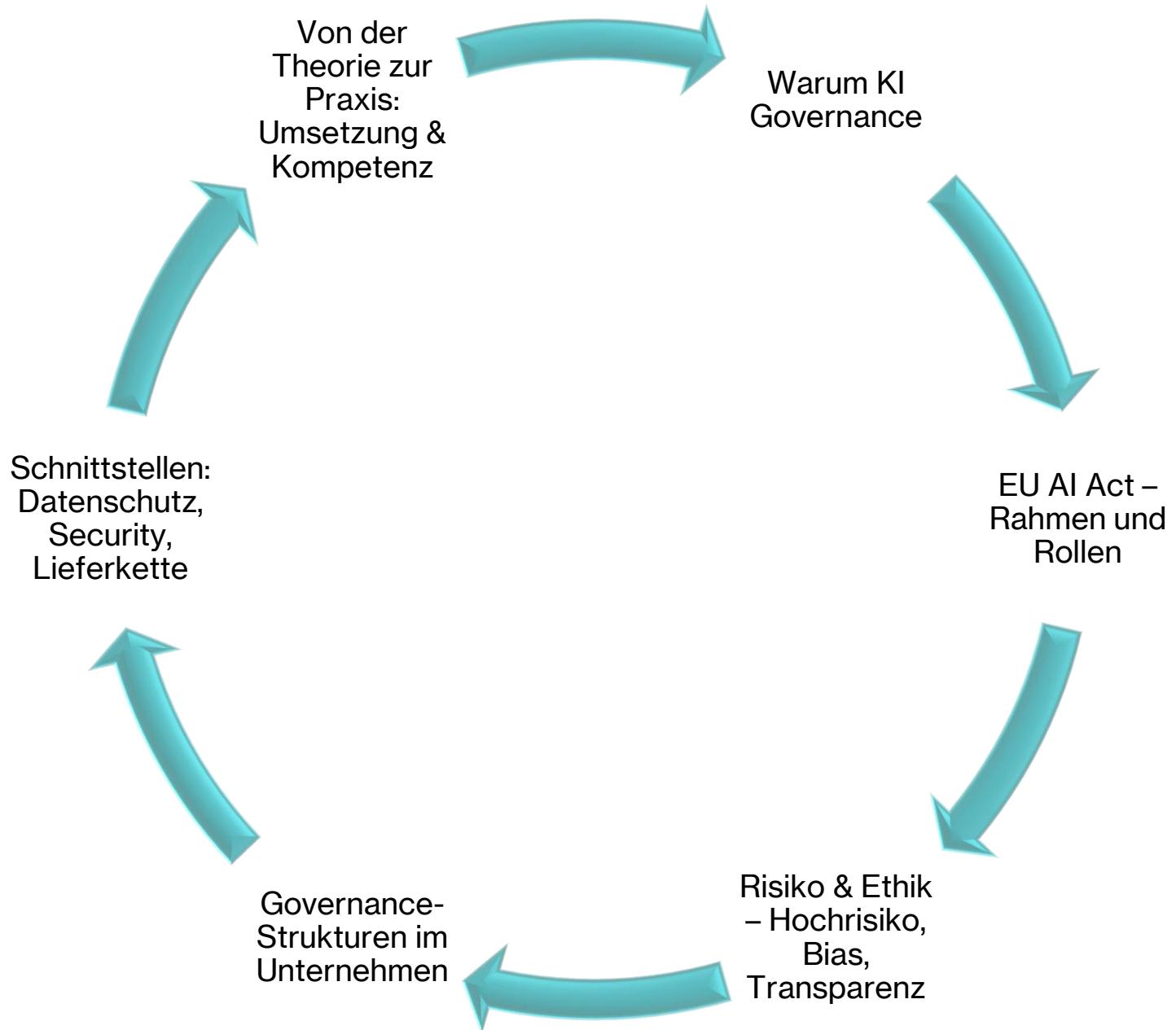
Dr. Natalie Ségur-Cabanac

Dr. Natalie Ségur-Cabanac



- △ Datenschutzbeauftragte von Magenta Telekom
- △ Expertin für Datenrecht, Compliance und Telekommunikationsrecht
- △ KI Governance
- △ Vorständin der ISPA Internet Service Providers Austria
- △ Initiatorin des ISPA Code of Conduct für ISPs
- △ Vorständin Women in AI Austria
- △ Keynote Speakerin
- △ Agiles Datenschutz-, und Compliance Managementsystem

Agenda



KI Governance - Balance zwischen Innovation und Verantwortung

Chancen

- Innovation
- Effizienz
- Neue Produkte

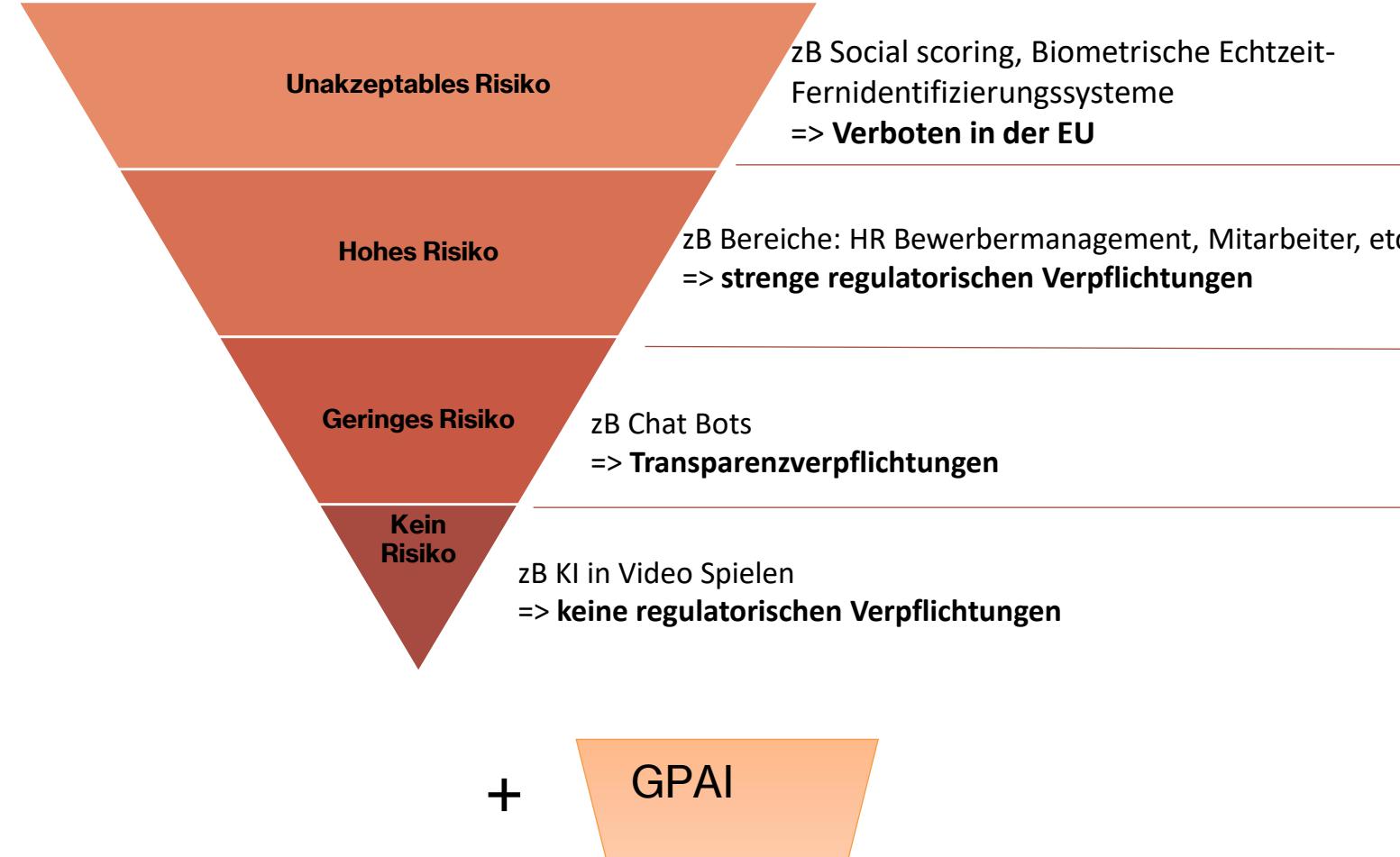
Risiken

- Rechtliche/regulatorische
- reputativ
- operativ
- ethisch



Ziel der KI Governance: Einen klaren Rahmen schaffen, der Innovation sicher ermöglicht und Risiken systematisch steuert.

EU KI Verordnung

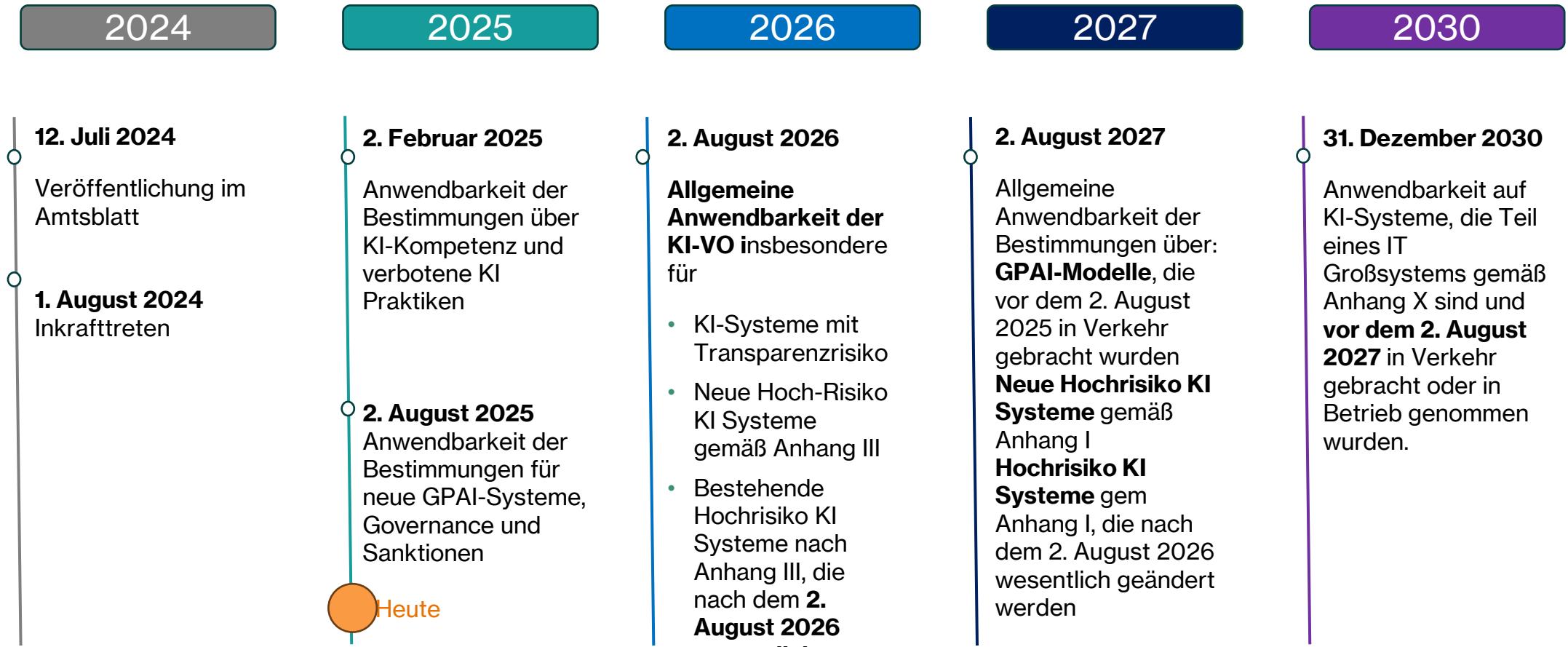


Ziele der KI Verordnung:

Schaffung von Vertrauen, Risikominimierung, Verteilung der Verantwortlichkeiten

- 1. Adressierung bestimmter Risiken** von AI Anwendungen, Risikobasierter Ansatz;
- 2. Rechtssicherheit**, um Entwicklung und Akzeptanz von AI in der EU
- 3. Menschen zentrierter Ansatz.** Setzen der höchsten regulatorischen Standards weltweit

KI Verordnung tritt schrittweise in Kraft

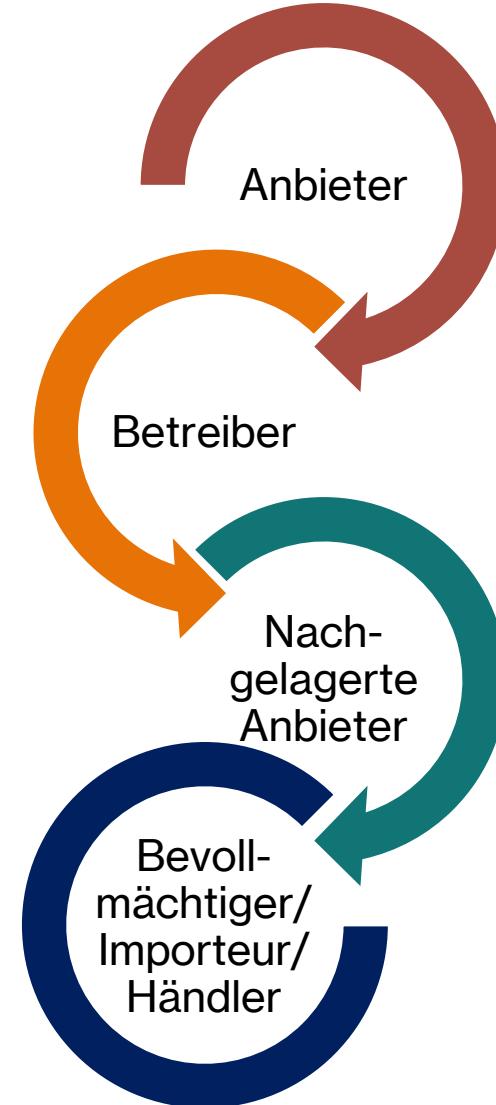


Definition Rollen

eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System in eigener Verantwortung verwendet, es sei denn, das KI-System wird im Rahmen einer persönlichen und nicht beruflichen Tätigkeit verwendet

Verwendet KI im Geschäftsbetrieb
zB Unternehmen für HR Zwecke

Vertriebsrolle in EU
Partnerfirmen



eine natürliche oder juristische Person, Behörde, Einrichtung oder sonstige Stelle, die ein KI-System oder ein KI-Modell mit allgemeinem Verwendungszweck entwickelt oder entwickeln lässt und es unter ihrem eigenen Namen oder ihrer Handelsmarke in Verkehr bringt oder das KI-System unter ihrem eigenen Namen oder ihrer Handelsmarke in Betrieb nimmt, sei es entgeltlich oder unentgeltlich

Entwickelt/ verkauft KI-System
zB Google, Open AI

Anbieter eines KI-Systems, einschließlich eines KI-Systems mit allgemeinem Verwendungszweck, das ein KI-Modell integriert, unabhängig davon, ob das KI-Modell von ihm selbst bereitgestellt und vertikal integriert wird oder von einer anderen Einrichtung auf der Grundlage vertraglicher Beziehungen bereitgestellt wird

Integriert KI eines Dritten
SaaS-Plattform

Definition von KI System



ein **maschinengestütztes System**, das für einen in wechselndem Maße **autonomen** Betrieb ausgelegt sind, das nach seiner Einführung **anpassungsfähig** sein kann und das aus den erhaltenen Eingaben für **explizite oder implizite Ziele ableitet**, wie Ergebnisse wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen hervorgebracht werden, die physische oder virtuelle Umgebungen beeinflussen können

Ziele des AI Act

Verbesserung des Funktionierens
des Binnenmarkts und

Förderung von
menschenzentrierter und
vertrauenswürdiger künstlicher
Intelligenz

Schutz vor schädlichen
Auswirkungen von Systemen der
künstlichen Intelligenz für

Unterstützung von Innovation

Gesundheit,

Sicherheit,

die in der Charta verankerten
Grundrechte,

einschließlich Demokratie,
Rechtsstaatlichkeit und
Umweltschutz, und die



Typen geregelter KI

**Verbotene KI
(Art 5)**

**Hochrisiko KI
Systeme
(Art 6)**

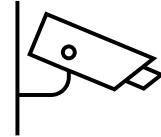
**Sonstige KI
(Art 52)**

**Generative KI
Modelle
Art 52a**

Diese KI-Systeme sind in der EU grundsätzlich verboten.



Emotionserkennung



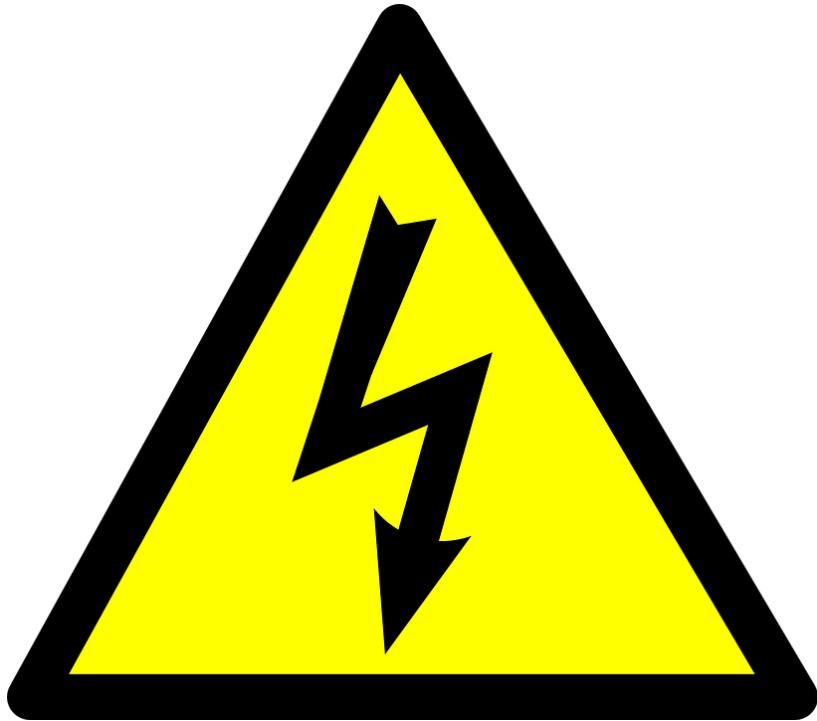
Gesichtserkennung



Social Scoring

- Unterschwellige Techniken, die **unterbewusst, manipulativ, täuschend** Personen **schadhaft beeinflussen** können
- Schadhaftes **Ausnützen** von **Schwachstellen** von **Personen** aufgrund ihres **Alters**, einer **Behinderung** oder einer **besonderen sozialen** oder **wirtschaftlichen** Situation
- die **Verwendung biometrischer Kategorisierungssysteme**, die **einzelne natürliche Personen** auf der Grundlage ihrer **biometrischen Daten** hinsichtlich **Gewerkschaftszugehörigkeit, religiöse** oder **philosophische Überzeugungen, Sexualleben** oder **sexuelle Orientierung** **kategorisieren**
- **Social scoring**
- Verwendung von **biometrischen Echtzeit-Fernidentifizierungssystemen** in öffentlich zugänglichen Räumen zum Zwecke der Strafverfolgung (mit Ausnahmen)
- KI-Systemen, die Datenbanken zur **Gesichtserkennung** durch das **ungezielte Auslesen** von **Gesichtsbildern** aus dem Internet oder aus **Videoüberwachungsaufnahmen** erstellen oder erweitern
- KI-Systemen zur **Ableitung** von **Emotionen** einer natürlichen Person in den Bereichen **Arbeitsplatz** und **Bildungseinrichtungen**, (mit Ausnahmen)

KI-Systeme, die als besonders gefährlich oder ethisch bedenklich erachtet werden.



Hochrisiko KI-Systeme unterliegen strengerem Vorschriften in Bezug auf **Transparenz, Rechenschaftspflicht, Datenschutz und menschliche Aufsicht**. Ziel ist es, ihren **sicheren und verantwortungsvollen Einsatz zu gewährleisten, potenzielle Risiken zu mindern und die Rechte und das Wohlergehen des Einzelnen zu schützen**.

Hochrisiko KI Systeme

1. Kritische Infrastrukturen: KI-Systeme, die in kritischen Infrastrukturbereichen wie Energie, Verkehr und Wasserversorgung eingesetzt werden, gelten als hochriskant. Diese Systeme haben das Potenzial, bei einem Ausfall oder einer Beeinträchtigung erhebliche Schäden oder Störungen zu verursachen.

2. Gesundheitswesen: KI-Systeme, die zu Diagnosezwecken, zur Patientenüberwachung und zur Verabreichung von Medikamenten eingesetzt werden, fallen in die Hochrisikokategorie. Die Genauigkeit und Zuverlässigkeit dieser Systeme sind entscheidend für die Sicherheit und das Wohlbefinden der Patienten.

3. Bildung und Berufsausbildung: KI-Systeme, die zur Bewertung von Schülerleistungen, zur Bereitstellung personalisierter Lernerfahrungen oder zur Entscheidungsfindung über Bildungsangebote eingesetzt werden, gelten als hochriskant. Es ist wichtig, Fairness, Genauigkeit und Transparenz in diesen Systemen zu gewährleisten, um voreingenommene Ergebnisse oder Diskriminierung zu vermeiden.

Hochrisiko KI Systeme Klassifizierung

KI System gilt als risikoreich, wenn:

a) Das AI-System dazu **bestimmt** ist, als **Sicherheitsbauteil** eines **Produkts** verwendet zu werden, **oder** das AI-System ist **selbst** ein Produkt, das unter die in **Anhang I** aufgeführten Harmonisierungsrechtsvorschriften der Union fällt;

und

b) das Produkt, dessen Sicherheitsbauteil gemäß Buchstabe a das **AI-System** ist, oder **das AI-System selbst** als Produkt muss einer **Konformitätsbewertung** durch einen **Dritten** unterzogen werden, damit dieses Produkt gemäß den in **Anhang I** aufgeführten Harmonisierungsrechtsvorschriften der Union **in Verkehr gebracht** oder **in Betrieb genommen** werden kann;

- **Profiling von natürlichen Personen** mittels KI = **immer** Hochrisiko KI
- **Untergeordnete** Unterstützung bei Entscheidungsfindung ist **nicht** hochrisikoreich

Anforderungen an Hochrisiko-KI-Systeme

Für Unternehmen sind die Pflichten für Hochrisiko-Systeme am relevantesten. Dazu gehören unter anderem:

Risikomanagementsystem: Etablierung eines kontinuierlichen Prozesses zur Identifizierung und Minderung von Risiken.

Datenqualität und Data Governance: Hohe Anforderungen an die Trainings-, Validierungs- und Testdaten.

Technische Dokumentation: Umfassende Dokumentation zur Nachvollziehbarkeit.

Protokollierung (Logging): Automatische Aufzeichnung von Ereignissen während des Betriebs.

Transparenz und Information der Nutzer: Klare Information über die Funktionsweise und Risiken.

Menschliche Aufsicht: Sicherstellung, dass Menschen jederzeit eingreifen und das System kontrollieren können.



Generative KI – Transparenz, Sicherheit, IP

GenAI bringt einzigartige Risiken in Output und Interaktion



kontrollierte Einführung mit klaren Sicherheits- und Transparenzstandards

- Dokumentation zum Trainingsmaterial (Kategorien, Quellen, Lizenzstatus); Hinweise zur Nutzung und Grenzen des Modells.
- Urheberrecht & Lizenzen
- Klärung von Nutzungsrechten an Trainingsdaten; Umgang mit urheberrechtlich geschütztem Material; Speicherung/Weitergabe von Nutzerinputs.
- Sicherheit & Red-Teaming
- Verpflichtendes Red-Teaming gegen Halluzinationen, Prompt-Injection, Datenexfiltration; Output-Kennzeichnung/Wasserzeichen.
- **Richtlinien für verantwortungsvolle Nutzung**
- **Richtlinien für sichere Prompts, sensiblen Datenumgang, menschliche Prüfung kritischer Outputs; Logging/Audit.**
- **Bewußtsein über „Schatten“ KI der Mitarbeitenden => Regelungen für Empowerment**

Governance-Strukturen für eine effektive KI-Steuerung

AI Steering Committee / AI Board (Strategische Ebene):

- Besetzung: C-Level (z.B. CDO, CIO, CHRO), Rechtsabteilung, Fachexperten.
- Aufgaben: Festlegung der KI-Strategie, Genehmigung von Richtlinien, Budgetverantwortung.

AI Ethics & Compliance Office (Taktische Ebene):

- Besetzung: Datenschutzbeauftragte/r, Rechtsexperten, KI-Experten.
- Aufgaben: Überwachung der Einhaltung des AI Acts, Durchführung von Risikobewertungen, Beratung der Projektteams.

KI-Projektteams (Operative Ebene):

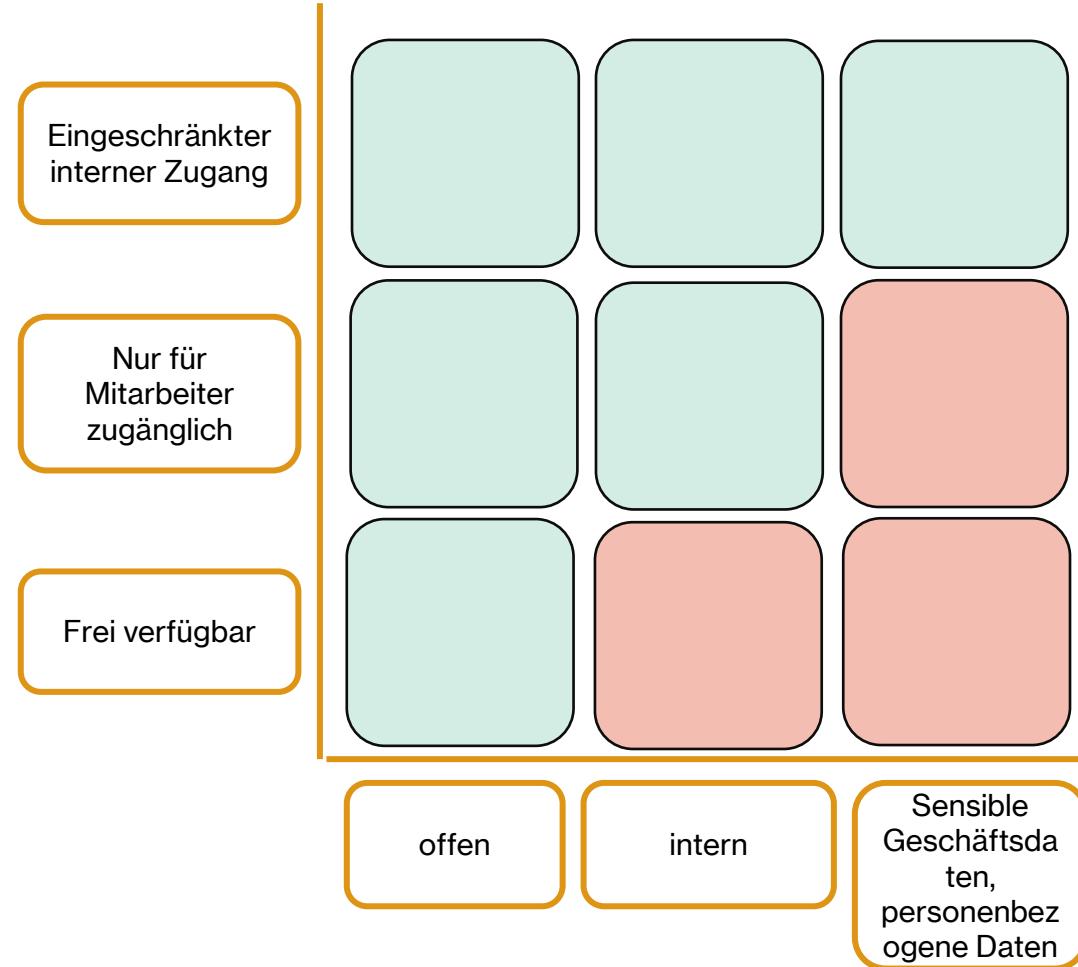
- Besetzung: Product Owner, Data Scientists, Entwickler.
- Aufgaben: Implementierung der Vorgaben in den konkreten KI-Anwendungen.



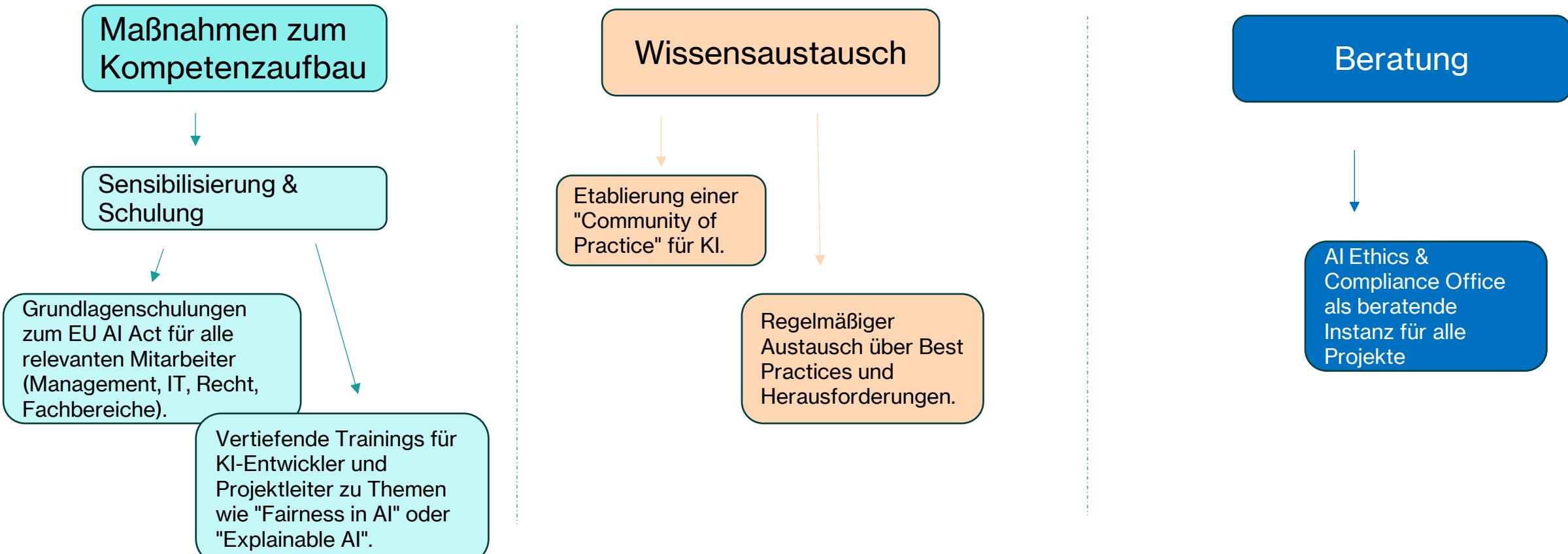
Das Fundament: Datenklassifizierung

- **Grundlage für die Risikobewertung:** Die Art der verwendeten Daten (z.B. personenbezogene Daten, sensible Daten) bestimmt maßgeblich das Risiko eines KI-Systems.
- **Anforderung des AI Acts:** Hohe Datenqualität ist nur mit klassifizierten und verstandenen Daten möglich.
- **Verhinderung von Bias:** Nur wer seine Daten kennt, kann Diskriminierung (Bias) in den Trainingsdaten erkennen und verhindern.

Nur wer seine Daten kennt, kann KI-Risiken steuern.

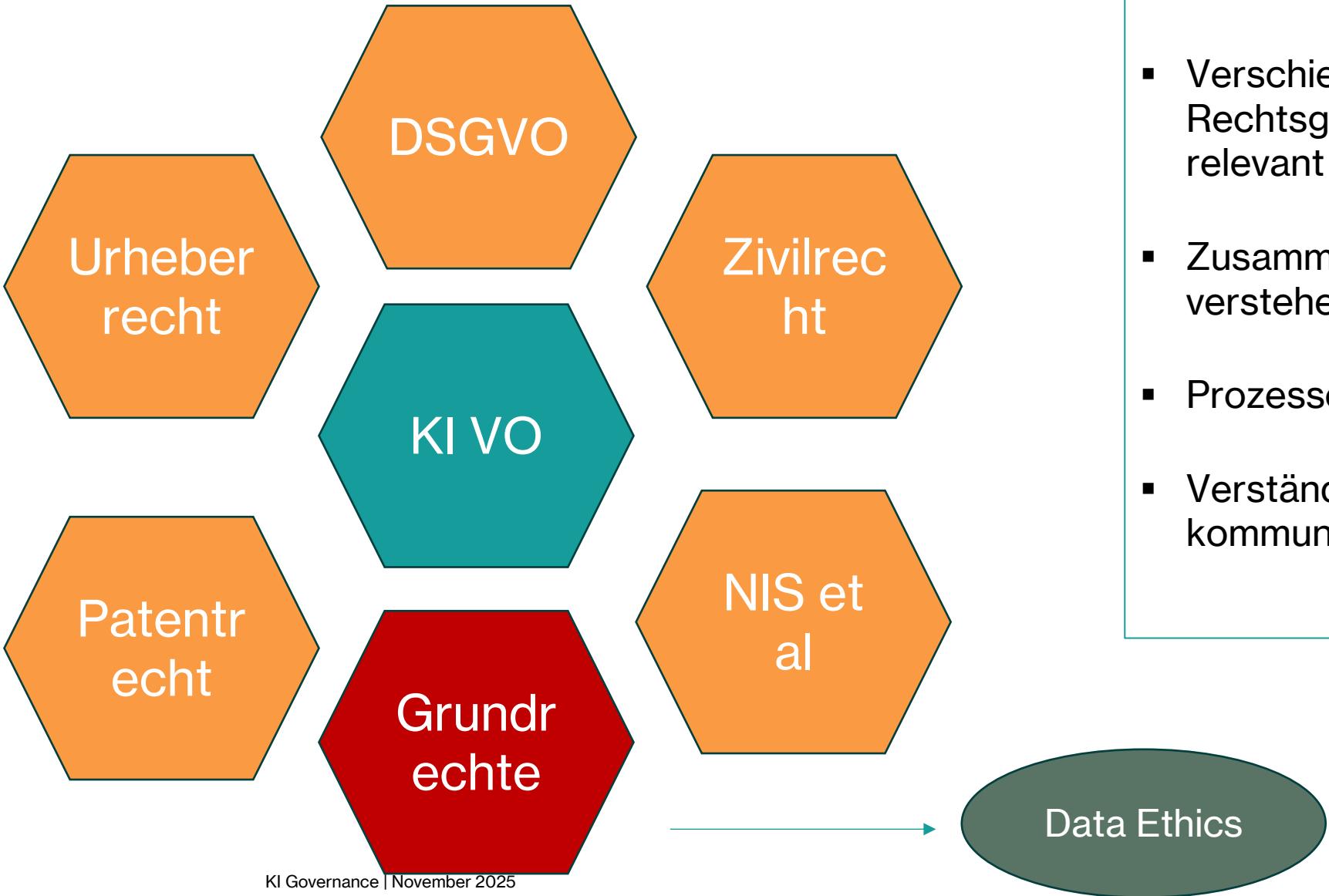


Das Fundament: Aufbau von KI-Kompetenz im Unternehmen



Governance funktioniert nur, wenn die Menschen sie verstehen und umsetzen können.

Über Grenzen gehen



- Verschiedene Rechtsgebiete sind relevant
- Zusammenhänge verstehen
- Prozesse zusammenführen
- Verständlich kommunizieren

Überblick bewahren

Datenschutz, Grundrechte und Schnittstellen zur DSGVO

- **DPIA und AI-Act-Grundrechtsprüfung**
verzähnen; Zweckbindung,
Datensparsamkeit, Rechtsgrundlagen.
- Besondere Kategorien pb Daten
- Zusätzliche **Kontrollen** und **Freigaben**;
Dokumentationspflichten;
Minimierung und **Pseudonymisierung**.
- **Betroffenenrechte**
- **Transparente** Nutzerinformation,
Widerspruchsmöglichkeiten,
Beschwerdekanäle.

AI Security – Schutz von
Daten, Modellen und
Lieferkette

- **Bedrohungen**
 - Adversarial Attacks, Daten-/Modellvergiftung, Model Theft, Prompt-Injection, Datenexfiltration.
- **Kontrollen**
 - Threat Modeling für KI, Guardrails, Input-/Output-Filter, RAG-Härtung, Secret Management, Zugriffskontrollen.
 - SBOM/„Model Bill of Materials“, Signaturen/Integrität, Pen-Tests/Red-Teaming, Backup/Recovery.
- **Betrieb**
 - Security Monitoring für Modelle, Patch-/Update-Prozess, Schlüsselrotation.

Datenschutz- und Grundrechtsprüfungen
sind Gatekeeper für Hochrisiko-Use Case

Security by design in jeder Phase des
Modelllebenszyklus

Third-Party AI – Verträge, Due
Diligence, Verantwortung

- Sicherheits-, Datenschutz- und Bias-Prüfungen; **Transparenz** zu **Trainingsdaten/Modell**.
- Verträge
- **Auditrechte**, Service Levels, Sicherheits- und Incident-Pflichten, IP-/Lizenzgarantien, Datenverwendung.
- Shared Responsibility
- **Klare Abgrenzung** von Anbieter- vs. Deployer-Pflichten; **Notfall-** und **Exit-Regelungen**.

Vertragliche Absicherung und
Sicherstellung der Verantwortungskette

Verantwortung und Vertrauen in der Praxis

Leitprinzipien

- ✓ **Unternehmensrichtlinien über**
 - Fairness
 - Verantwortung
 - Transparenz
 - Rechenschaft
 - Sicherheit
 - Inklusion
 - etc

Mechanismen

- ✓ **Strukturierte Prozesse für**
 - Ethics Assessment
 - erklärbare Modelle und
 - Nutzerverständlichkeit
 - Ev Beschwerde- und Einspruchsprozesse
 - Grundrechtsfolgeabschätzung
- ✓ **BIAS Erkennungssoftware**
(siehe BSI Papier)

Messbarkeit

- ✓ **Governance-Metriken**
 - Bias-Indikatoren
 - Erklärbarkeits-Score
 - Nutzerbeschwerden
 - Korrekturquote

Ethik wird wirksam, wenn sie messbar und entscheidungsrelevant gemacht wird.

KI-Inventar und Klassifizierung – die Basis der Steuerung

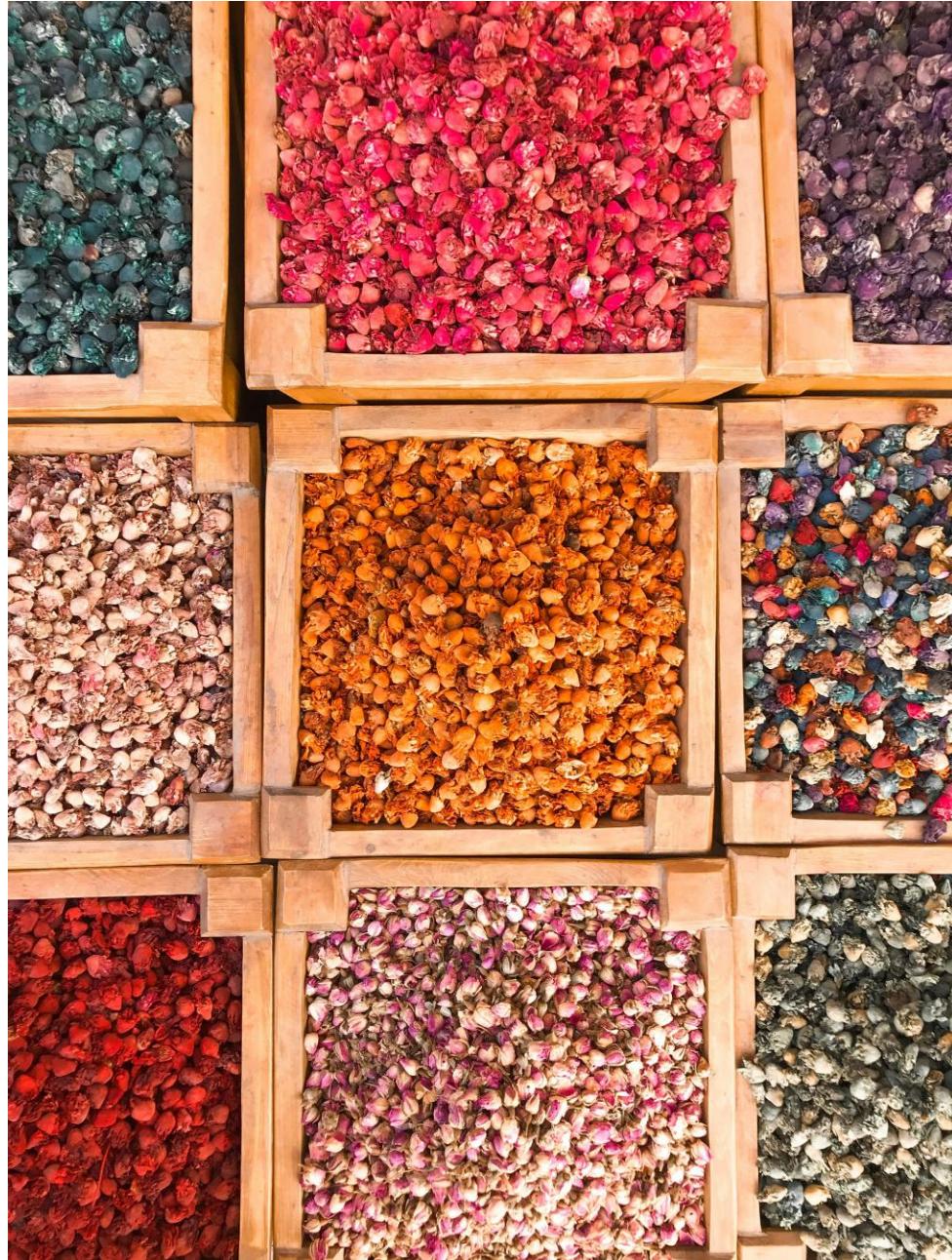
Zentrales Inventar/Register

- ✓ Alle KI-Systeme, Modelle, Drittanbieter-Dienste;
- ✓ Verknüpfung zu Datenkatalog,
- ✓ DPIA,

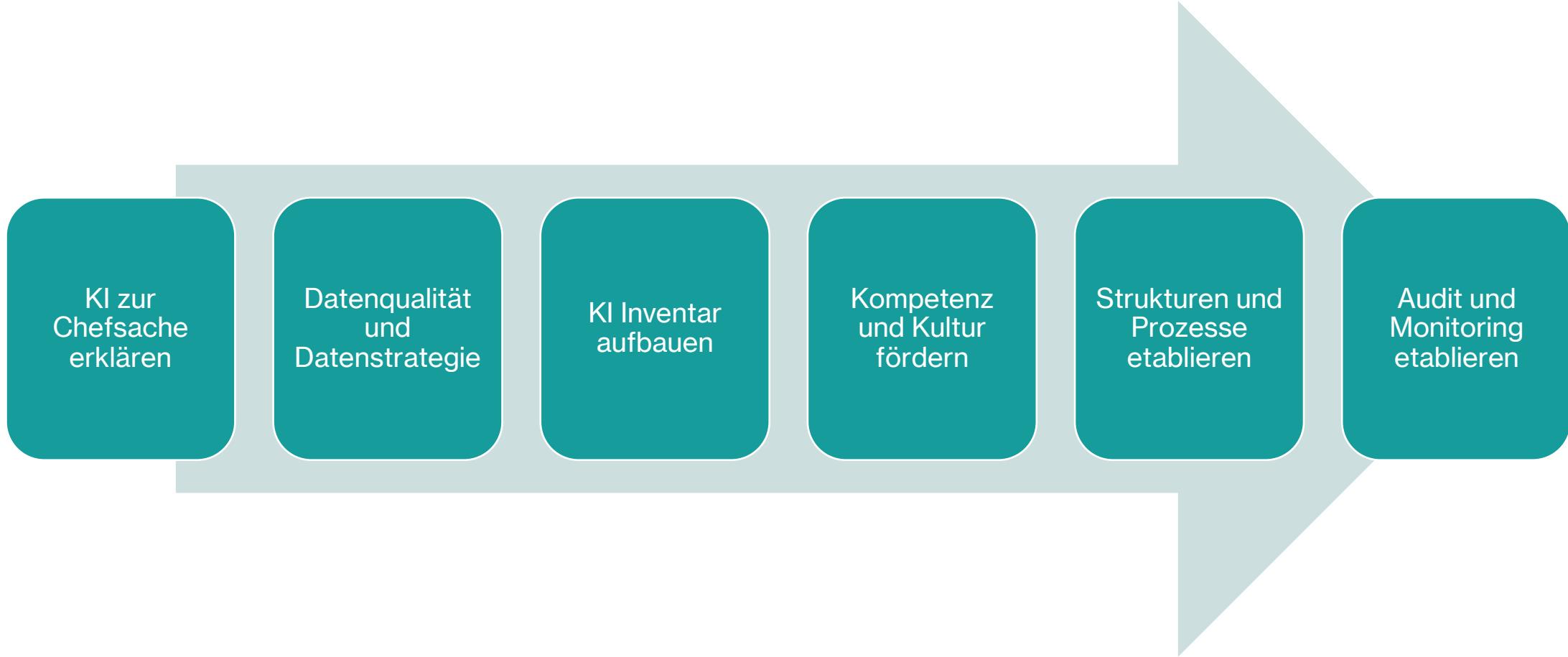
Klassifizierung

- ✓ Risikoklasse je Use Case; Kennzeichnung von Hochrisiko, GPAI-Nutzung, personenbezogene Daten + besonderer Kategorien
- ✓ Versionierung, Change-Historie, Gültigkeitsstatus; Pflicht zur Aktualisierung bei Änderungen

Das Inventar ist die „Single Source of Truth“ für Prüf- und Meldepflichten.



6 Schritte zur KI Governance



KI Governance = Enablement, nicht Bremse

KI zur Chefsache erklären

- KI Strategie
- Datenstrategie
- Data Governance
- Datenqualität





Dr. Natalie Sécur-Cabanac
nataliesegur@gmail.com



Danke für Ihre Aufmerksamkeit!

Fotos: pexels.com