Künstliche Intelligenz im öffentlichen Sektor: Wie gewährleisten wir Transparenz und Accountability?

Künstliche Intelligenz und Digitale Transformation des Staates: Revolution oder Evolution?
Dreiländertagung SGVW, Bern, 27. März 2026
Breakout-Session

Tobias Polzer, Wirtschaftsuniversität Wien (tobias.polzer@wu.ac.at)
Jakob Kühler, Universität Potsdam und Universität Kassel (jakob.kuehler@uni-kassel.de)

Herausforderungen

NEWSLETTER KÜNSTLICHE INTELLIGENZ

Anthropic, das Pentagon und der Streit um KI im Krieg

Das US-Verteidigungsministerium bestraft Anthropic dafür, dass es dem Militär Grenzen setzt. Das könnte der ganzen Branche schaden – und am Ende Werbung für Claude sein

Philip Pramer
9. März 2026, 10:24

DERSTANDARD

(Anwendungsmöglichkeiten von KI inkludieren auch Massenüberwachung und automatische Waffensysteme)

Word embeddings quantify 100 years of gender and ethnic stereotypes

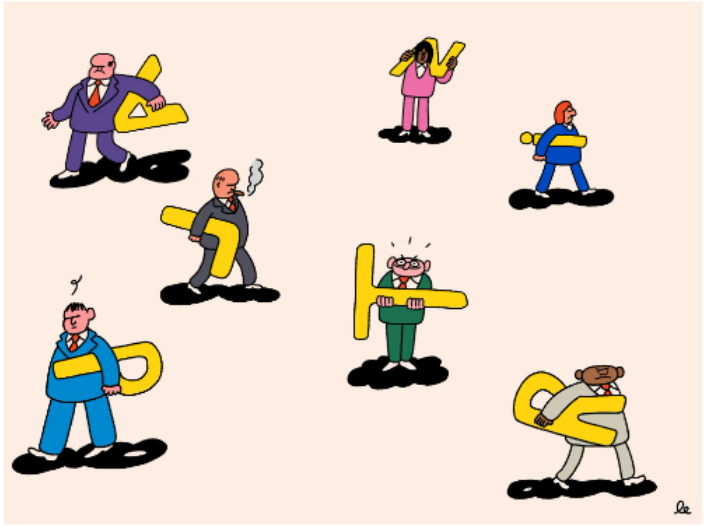
Nikhil Garg[a,1], Londa Schiebinger[b], Dan Jurafsky[c,d], and James Zou[e,f,1]

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

REPUBLIK

Magazin Feed Dialog



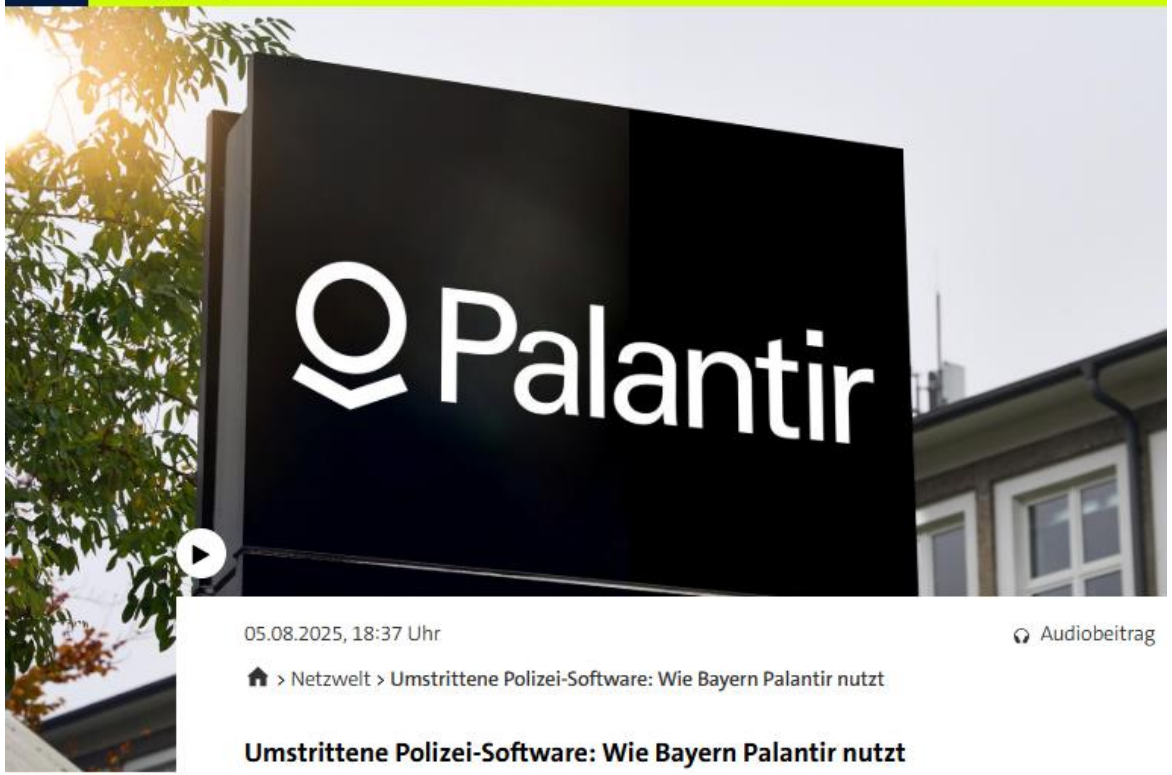
Schweizer Konzerne finanzieren Palantirs Überwachungs-Software

Die europäische Finanzbranche hat Milliarden in den US-Tech-Konzern Palantir investiert, dessen Produkte auch für völkerrechtswidrige Kriege eingesetzt werden. Nun ist klar: Zu den Aktionären gehören auch die UBS, die ZKB und die Schweizerische Nationalbank.

Von Adrienne Fichter, Lorenz Naegeli, Yves Wegelin (Text) und Leillo (Illustration), 19.03.2026

BR24 Bayern Wir vor Ort Kommunalwahl Wirtschaft Sport #Faktenfuchs Dein Argument Kultur

Netzwelt



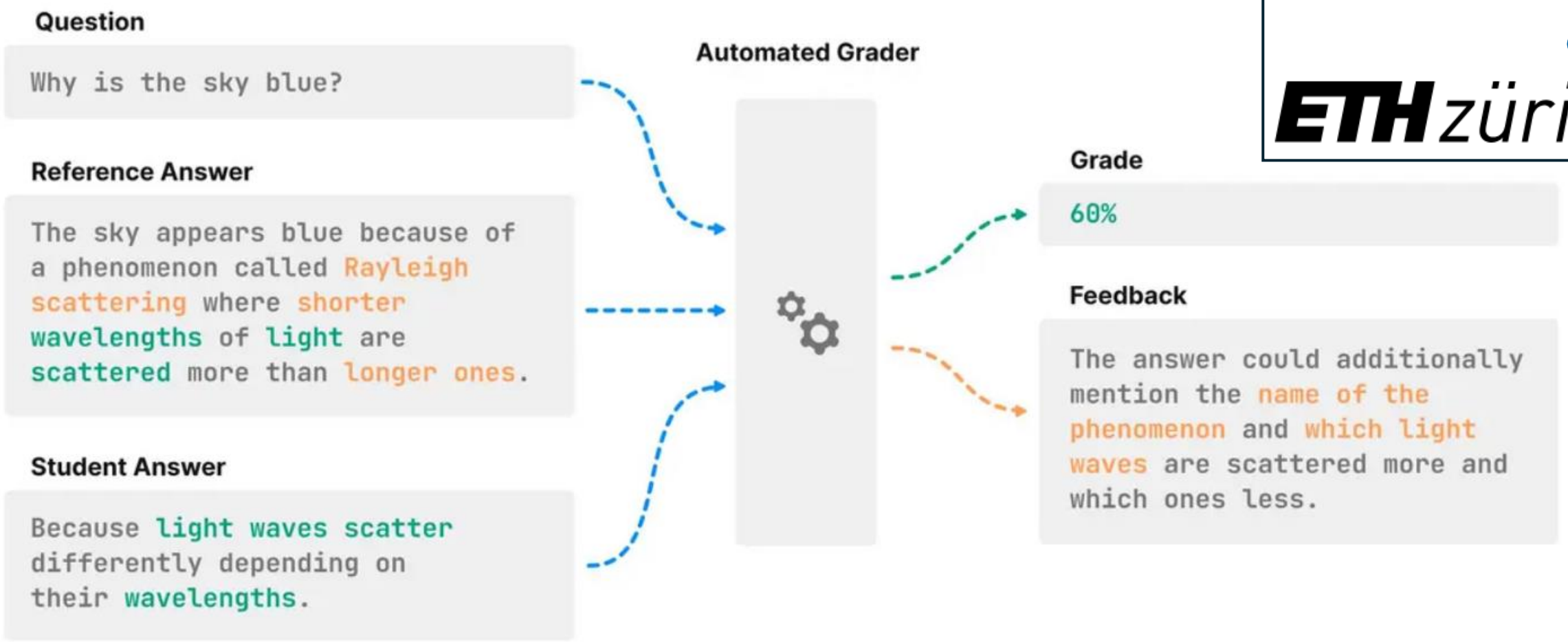
05.08.2025, 18:37 Uhr Audiobeitrag

Netzwelt > Umstrittene Polizei-Software: Wie Bayern Palantir nutzt

Umstrittene Polizei-Software: Wie Bayern Palantir nutzt

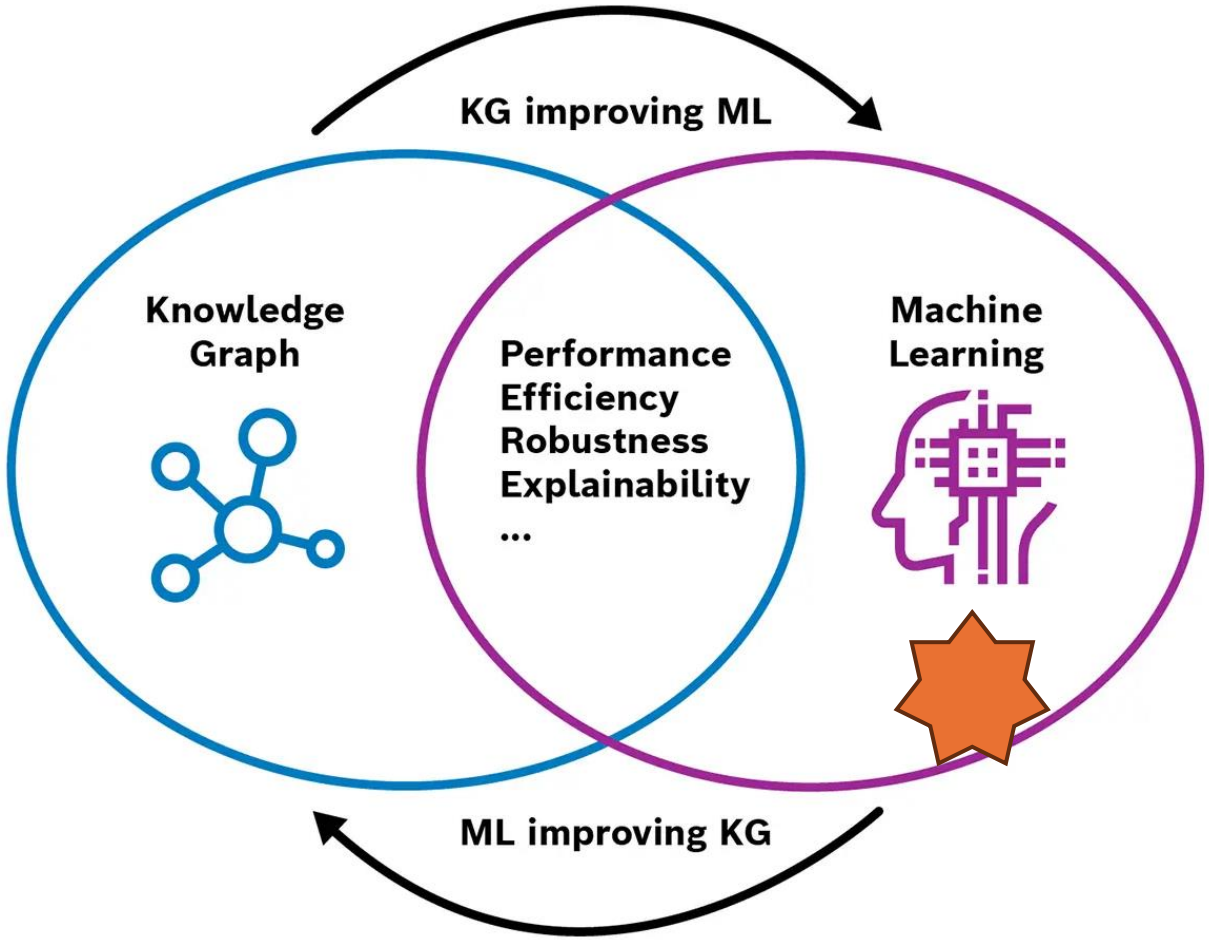
Bundesinnenminister Dobrindt liebäugelt mit dem bundesweiten Einsatz von Analysesoftware der umstrittenen US-Firma Palantir. In Bayern läuft diese bereits seit Monaten – zwar mit strengen Auflagen, aber von Datenschützern skeptisch beäugt.

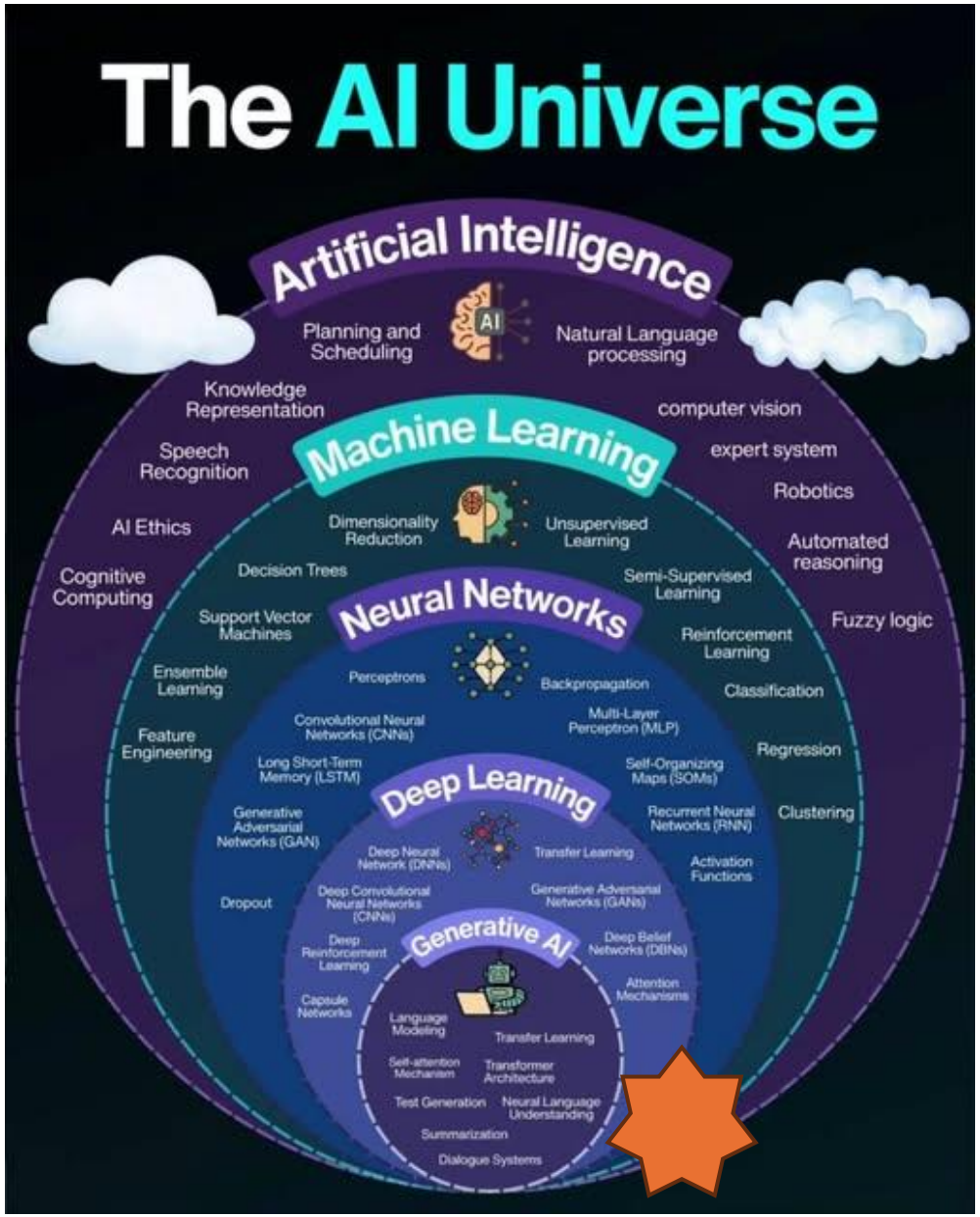
Von Fritz Espenlaub

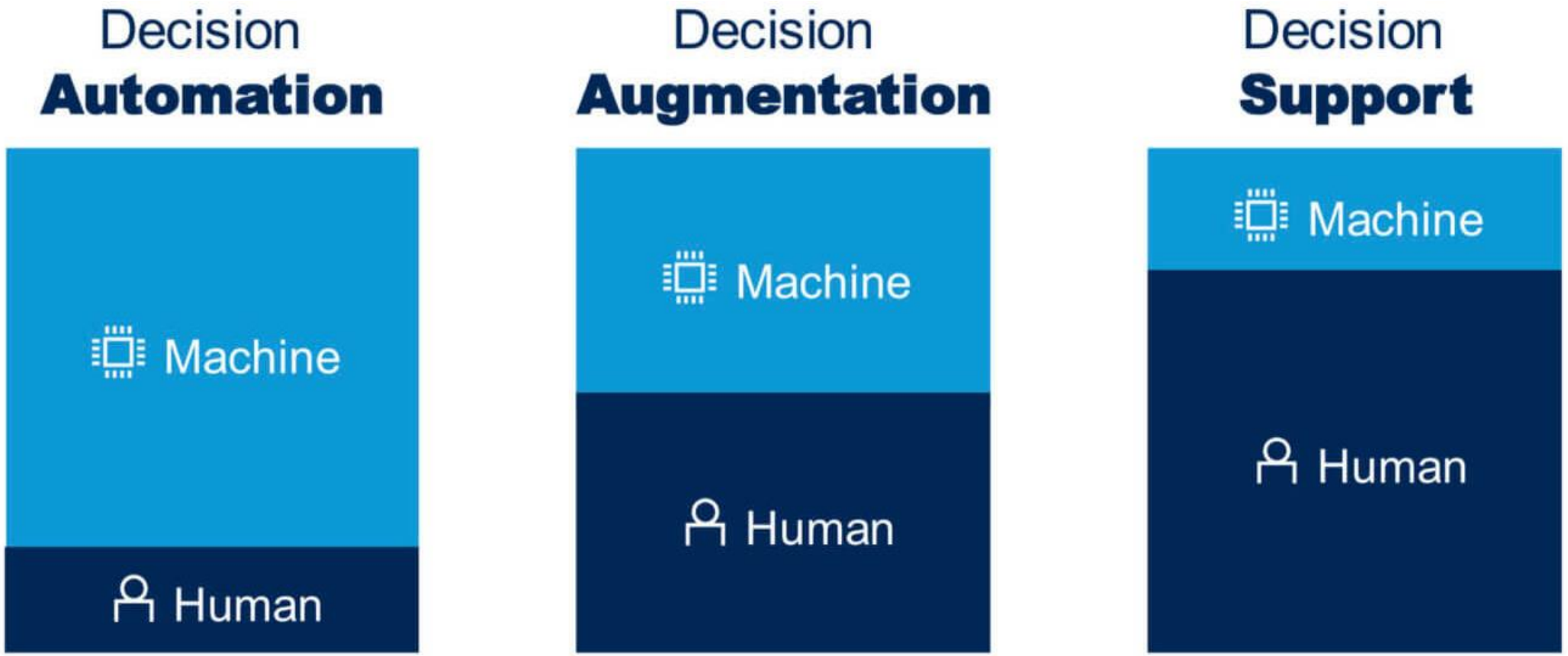


Pilotprojekt von ZHAW und ETH:
Automatisierte Benotungen von Prüfungen (mit Möglichkeit/Recht, Einspruch zu erheben)

KI-‚Spielarten‘







Argumente:

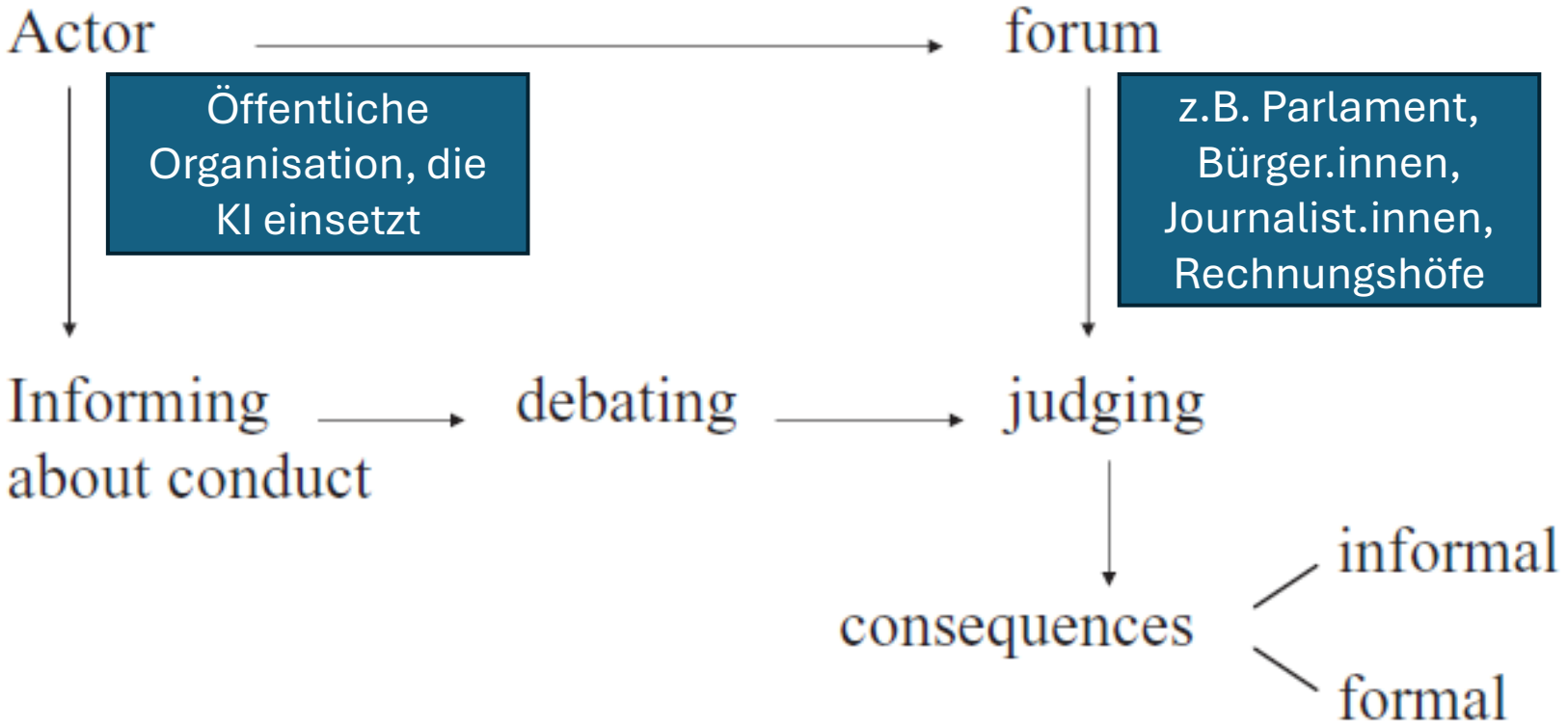
- ‚Automatisierung führt zu Jobverlust' (Miller 2018) vs.
- ‚Automatisierung kann helfen, Bias zu überwinden' (Miller/Keiser, 2021)

Gartner

Setting the Scence – Accountability im öffentlichen Sektor

Definition von Accountability für die Nutzung von KI-Systemen:

„the relationship between actors and forums, where the actors are required to explain and justify the decisions they made using AI algorithms and face consequences for such decisions“ (Chen et al. 2026: 2).



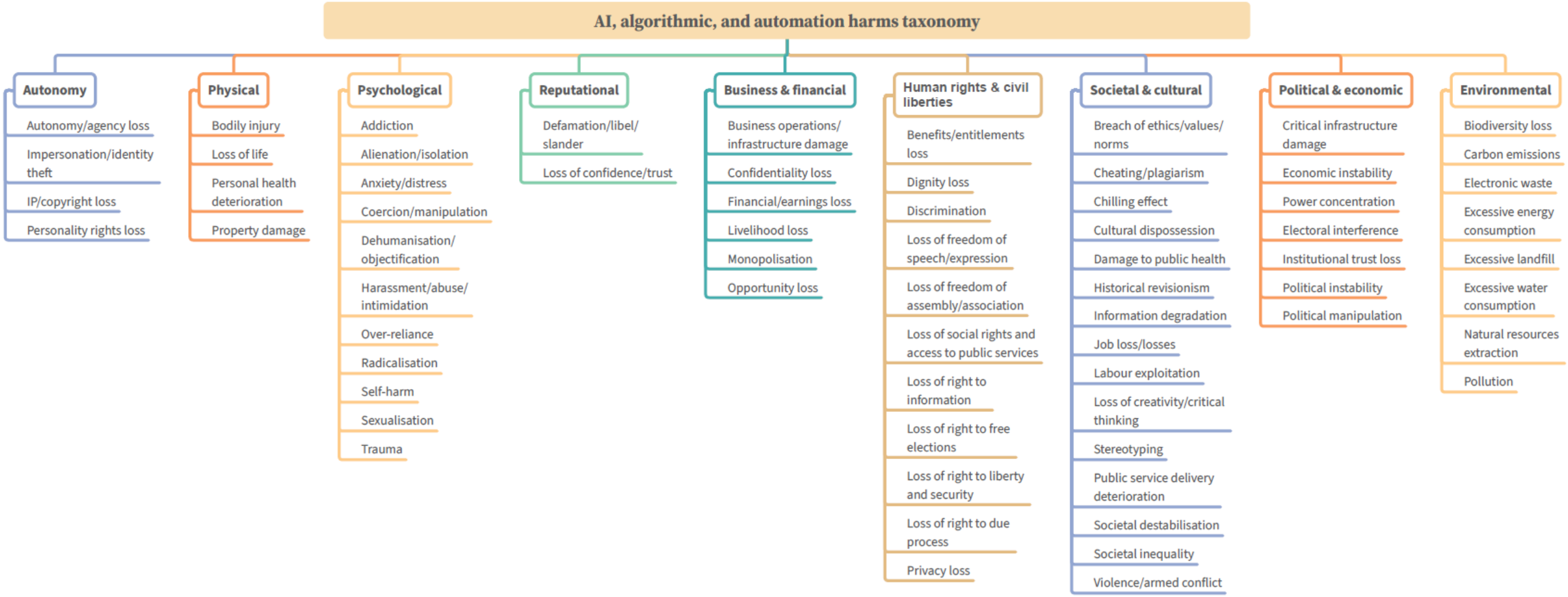
(Bovens 2007: 454)

Diskutieren Sie und schreiben Sie bitte auf (5 Minuten)!

Welche möglichen Arten von Bedenken/ Schäden/negative Auswirkungen bez. Accountability und Transparenz können beim KI-Einsatz auftreten?

Maßnahmen, diese Schäden zu minimieren?

Taxonomie von möglichen Schäden



(Abercrombie et al. 2024: 21)

Herausforderung der „Blackbox“

- **Problem:** Regierungen nutzen KI zunehmend für weitreichende Entscheidungen (z.B. Sozialleistungen, Steuerprüfung, Strafverfolgung)

- **Risiko:** Undurchsichtige Algorithmen können zu „automatisierter Ungerechtigkeit“ führen – voreingenommene Ergebnisse, die schwer anzufechten sind

- **Ziel:** Der Übergang von blindem Vertrauen hin zu **verifizierbarer Sicherheit**

Beispiel: Neuronales Netz als Blackbox

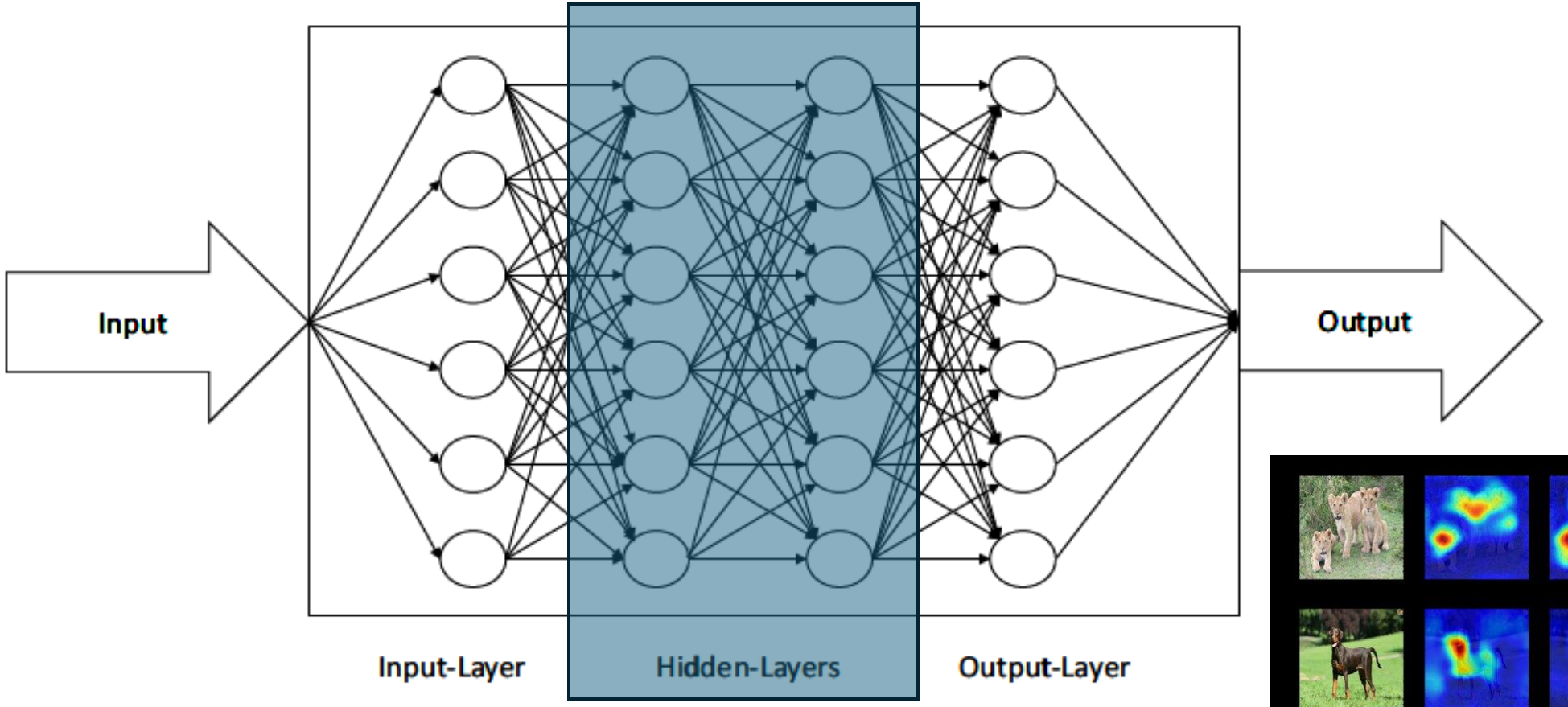
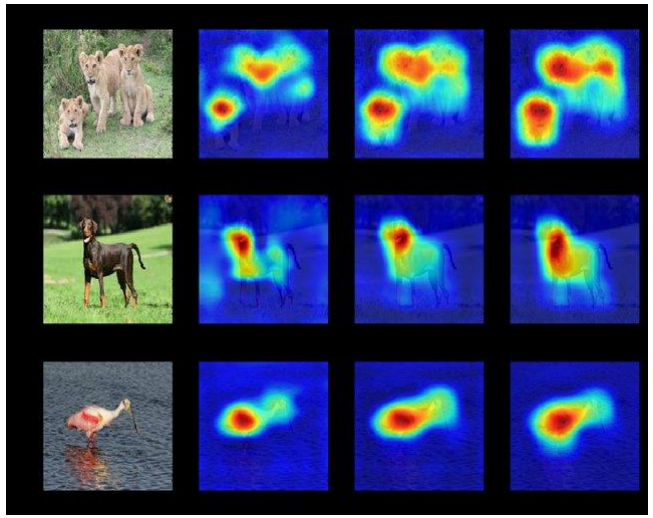


Abb. 1 Vierstufiges Perzeptron als Blackbox. (Eigene Darstellung)



(Piesold und Krause 2025: 103)

Ethische Bedenken (1)

- Training der Algorithmen wirft ethische Fragen auf (Bias in den Trainingsdaten): ‚garbage in – garbage out‘

Prompt in Gemini: „create an image of a normal family“

- Integration ‚unsichtbarer‘ Systemprompts (Moderation von KI-Systemen durch den Hersteller), um unethische Antworten zu vermeiden

- Anpassungsmöglichkeit bei Adaptierung im öffentlichen Sektor → aber: Gefahr des Einbringens von bestimmten politischen Ansichten und (anderem) Bias

Ethische Bedenken (2)

(1) *Inconclusive evidence.* Algorithms are mostly designed to produce outputs through inferential statistics. As statisticians well know, associations between variables are not causation. Besides, an algorithm can compute the level of uncertainty and/or risk associated with a phenomenon, but it does not give certainties. As such, the evidence might not be sufficient to justify action.

(2) *Inscrutable evidence.* In decision-making, there is an assumption that the relationship between data input and conclusions is clear. In machine learning algorithms, this association might not be clearly visible or even possible, depending on the design clarity and the ex-post auditability of the system. This poses questions of the opacity of the algorithmic-driven public service delivery and decision-making.

(3) *Misguided evidence.* The well-known adage "garbage-in, garbage-out" applies also to very sophisticated AI algorithms. If the input data are biased, unreliable or irrelevant, the resulting information will exhibit the same qualities, misguiding the decision makers.

(4) *Unfair outcomes.* AI algorithms can trigger actions that could lead to discriminatory outcomes.

(5) *Transformative effects.* The information produced by algorithms can affect the way people conceptualize society, creating new visibilities and modifying existing ones. It has the ability to re-ontologize societal institutions and identities.

(6) *Traceability.* The above issues exemplify the difficulties in detecting the flaws, effects and responsibility in an algorithmic decision-making process. As such, traceability requires the identification of both the cause and responsibility of an AI output and outcome.

(Bracci 2023: 743)

Risikobasierter Ansatz (gemäß EU AI Act)

Was sind Arten riskanter Anwendung?

- **Unannehmbares Risiko:** (z.B. Social Scoring) – verboten
- **Hohes Risiko:** (z.B. Justiz, Migration) – Erfordert obligatorische **Audits durch Dritte** und strenge Datenqualitätskontrollen
- **Transparenzrisiko:** (z.B. Chatbots) – Müssen eindeutig als KI gekennzeichnet sein, um Täuschung zu vermeiden
- **Minimales Risiko:** (z.B. Spam-Filter) – Erlaubt unter Einhaltung allgemeiner Best Practices

EU: Fokus auch auf Haftungsfragen

Schweiz:

Europaratskonvention
→ Vertrauen aufbauen, mit Transparenz, Menschenrechte, Demokratie als Kern

Strukturierungsvorschlag: Wo können Lösungen ansetzen?

Typen von Accountability

Design – Pilotierung – Skalierung

Einsatz in Verfahren

Typen von Accountability im öffentlichen Sektor

TABLE 2 | Types of accountability of AI-based system use in the public sector.

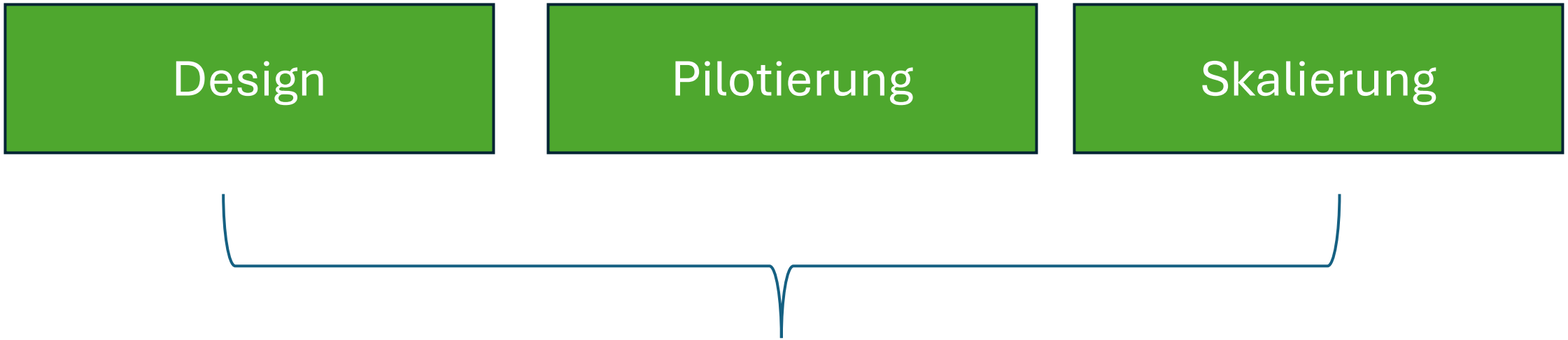
Type of accountability	Who?	To whom?	For what?	Illustrative examples
Legal	Those legally liable for the harm	Claimants, the public	Breach of legal obligations	Legal compliance, due diligence
Bureaucratic	Lower-level bureaucrats	Senior bureaucrats	Not always clearly defined	Rules/standards, reporting, auditing
Professional	Technical experts	Managers (relying on expertise)	Technical decisions	Evidence-based justification, trust in expert judgment
Political	Elected representatives	Constituents	Policy promises and priorities, competent governance, societal benefit	Public opinion, legislative oversight
Markets	Sellers or producers	Consumers	Product quality	Efficiency and transparency to consumers
Information transparency	System operators	Public stakeholders	Information for trust and autonomy	Providing data or information

Source: Modified from Williams et al. (2022).

Legal – Bureaucratic – Professional – Political – Markets – Information transparency

(Chen et al. 2026: 3)

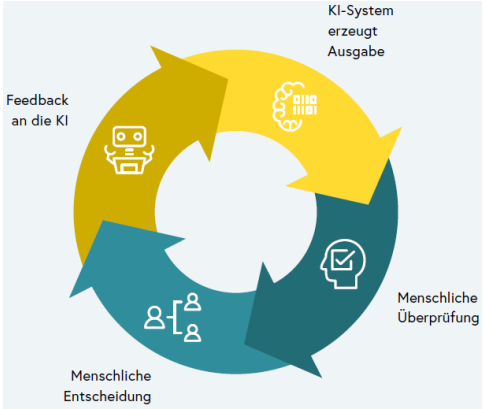
Verschiedene Herausforderungen in Phasen der **KI-Systemeinführung**



- AI-explainability
- Power through design
- AI-literacy/expertise

Verschiedene Herausforderungen in Phasen der **KI-Nutzung**

- Im Verfahren

Input (ex-ante)	Throughput	Output
• Sichtbarmachung, z.B.: ‚Bei der Bearbeitung des Antrags findet ein KI-System Einsatz‘	‚Human in the loop‘ 	• Z.B.: ‚Dieses Schreiben wurde mit Hilfe eines KI-Systems erstellt‘

Zurück zu Ihren Ergebnissen!

- Sortieren Sie Ihre Vorschläge bitte auf der Pinnwand zu den folgenden Problembereichen zu, bzw. eröffnen ein neues Feld.
- Zeigen Sie auf, wie Ihre Maßnahmen hier helfen könnten?
- Beziehen sich die Maßnahmen auf:
  - KI-Einführung?
  - KI-Anwendung?

Lücken in der Umsetzung und Anwendung

Typen von Accountability

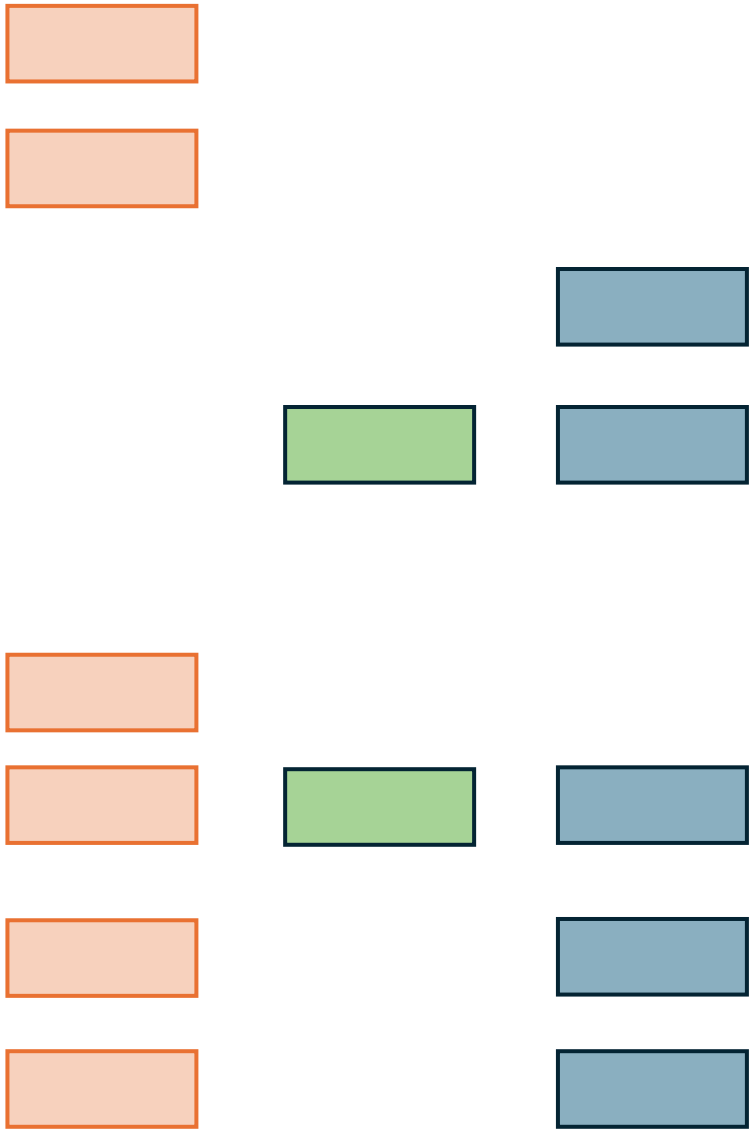
Phasen KI-Einführung

Phasen KI-Einsatz

TABLE 8 | Research-practice gaps and actionable recommendations.

	Research-practice gaps	Recommendations
1	Undefined roles and responsibilities	• Define who makes AI decisions and assign clear responsibility.
2	Uncertainty about legal obligations	• State if the policy is compulsory or advisory and specify legal implications.
3	Absence of AI usage approval procedures	• Identify which AI tools require approval • Establish reporting/approval procedures.
4	Insufficient explainability	• Provide accessible explanations of how AI influences decisions. • Build employees' capacity to interpret decisions informed by AI.
5	Lack of an expert audit	• Require expert audits, especially in high-risk areas.
6	Limited citizen engagement	• Involve citizens before and after deployment. • Ensure feedback channels are in place.
7	Unclear consequences of violation	• Define penalties for non-compliance and outline legal recourse for violations.
8	Missing appeal mechanism	• Guarantee citizens' right to appeal AI-informed decisions through a formal process.

Source: Created by the authors.



(Chen et al. 2026: 14)

Schlaglichter Maßnahmen - Explainable AI (XAI)

„Explainable AI (XAI) ist eine Sammlung von Methoden, die dem Verständnis und der Gewinnung von Erkenntnissen aus Deep-Learning-Modellen dienen. Diese Erkenntnisse können verschiedenen Zwecken dienen, von der Sicherstellung der Modellrobustheit und der Beseitigung von Verzerrungen bis hin zur Förderung wissenschaftlicher Erkenntnisse.“ (FHNW)

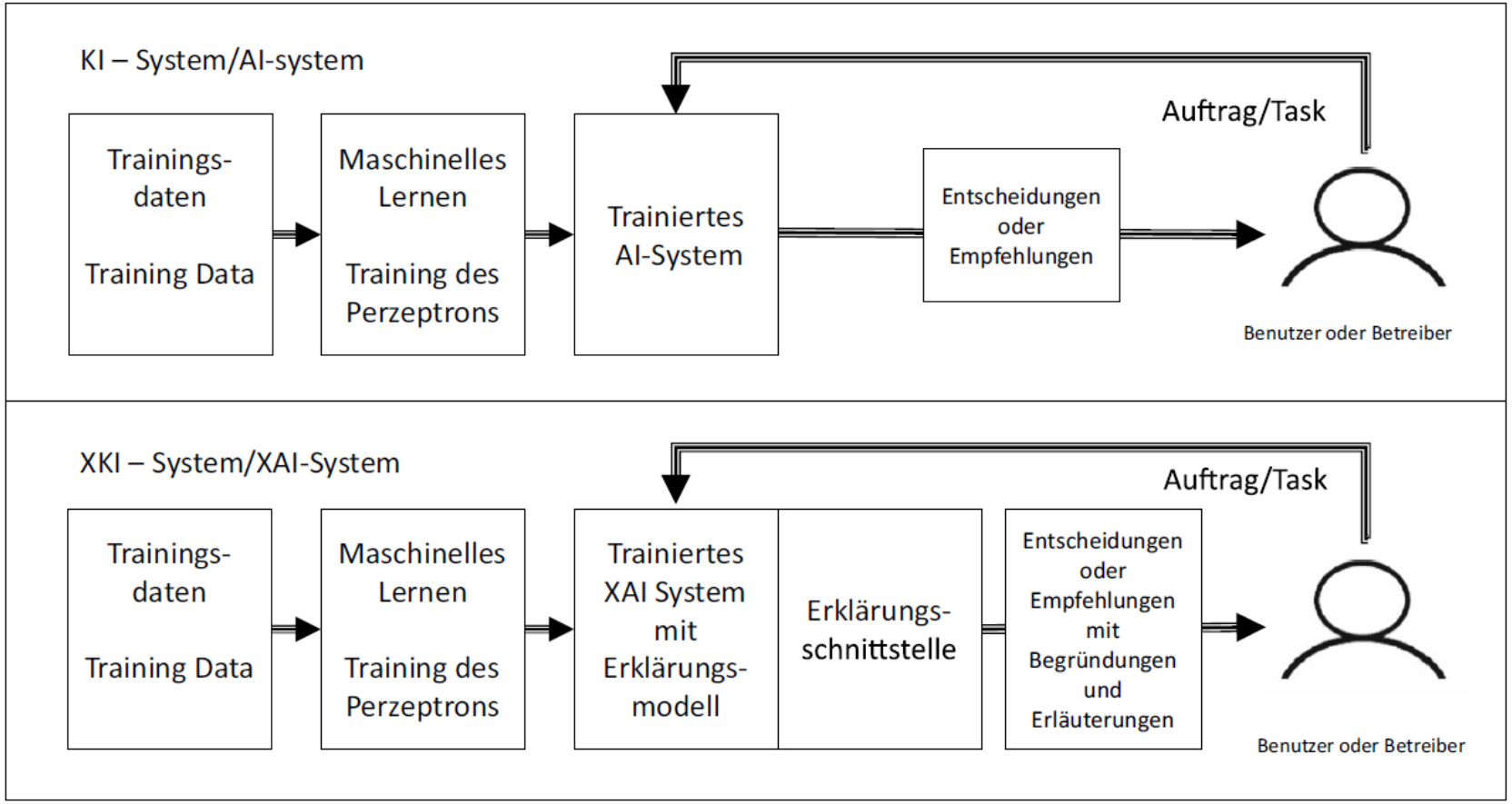


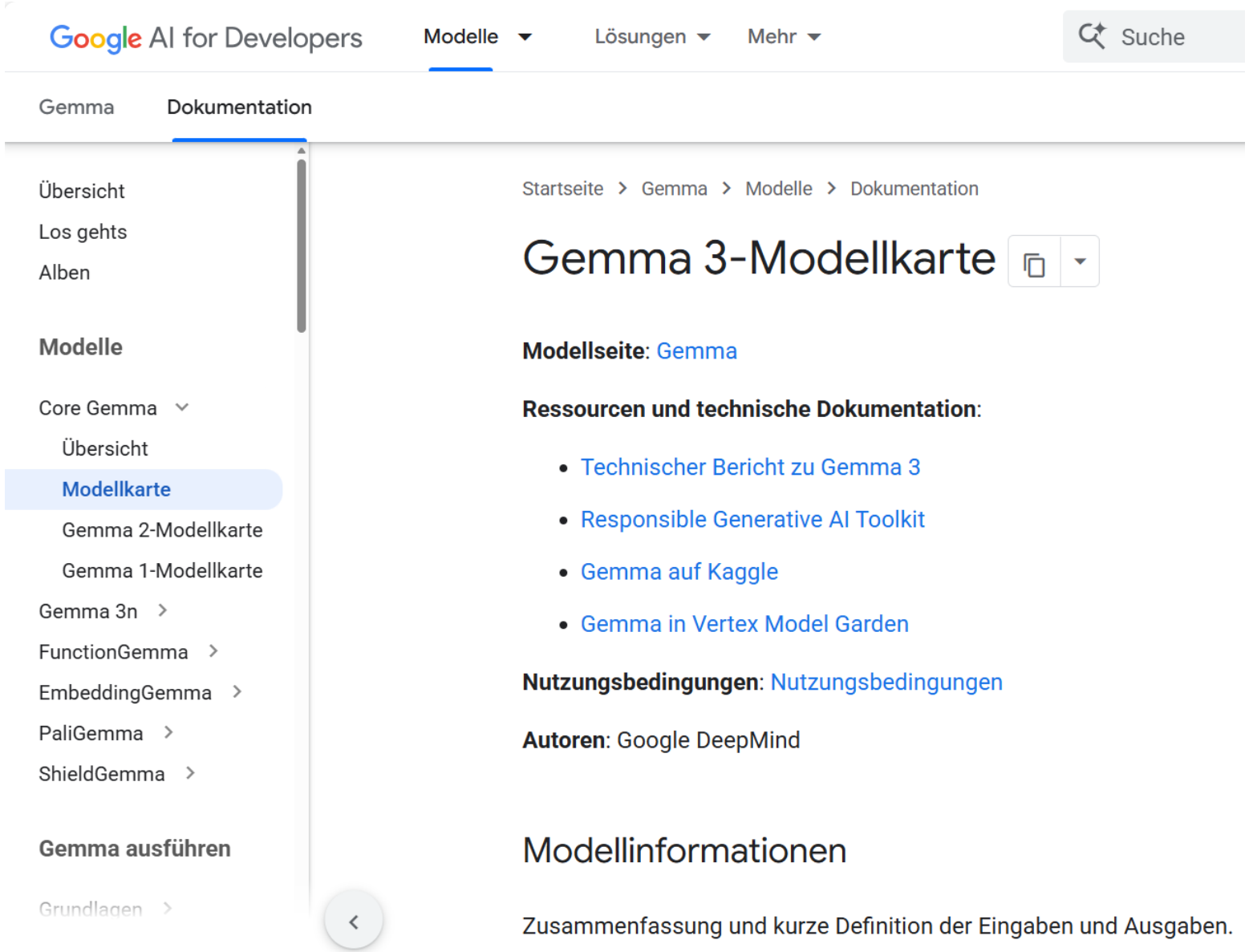
Abb. 6 XAI Konzept. (Eigene Darstellung angelehnt an [DARPA, 2016, 6])

(Piesold und Krause 2025: 112)

Modellkarte – Transparenz

- Kommunikation wichtiger Informationen über ein KI-Modell (z.B. Verwendungszweck, Leistungskennzahlen, Einschränkungen und ethische Aspekte)
- Helfen Entwicklern, Prüfern und Benutzer.innen zu verstehen, wie ein Modell erstellt wurde, worauf es trainiert wurde, wo es gute Ergebnisse liefert und wo es möglicherweise versagt
- Eine gut gestaltete Modellkarte enthält normalerweise:
  - Modellzweck und vorgesehene Anwendungsfälle
  - Datensatzinformationen und Trainingsmethodik
  - Leistungskennzahlen über relevante Untergruppen hinweg
  - Bekannte Einschränkungen und Fehlerszenarien
  - Ethische oder Fairness-Überlegungen
  - Empfehlungen zur sachgemäßen Verwendung und Benutzer.innenschulung

Beispiel Modellkarte



Checkliste – Transparenz

Transparenz

Sind die spezifischen Ziele und Zwecke des Einsatzes der KI-Anwendung identifiziert und dokumentiert?	◯ Ja	◯ Nein
Gibt es eine Dokumentation, die die technische Entwicklung des Modells erläutert?	◯ Ja	◯ Nein
Ist die Funktionsweise der KI-Anwendung nachvollziehbar?	◯ Ja	◯ Nein
Sind die Datensätze, die mit dem KI-System verbunden sind, bekannt? (Siehe 8.2 Datenschutzgrundverordnung)	◯ Ja	◯ Nein
Wird den Nutzer:innen, wann immer möglich, erklärt, wie das KI-System zu seinen Ausgaben, Inhalten, Empfehlungen oder Ergebnissen kommt und welche Logik dahintersteckt?	◯ Ja	◯ Nein
Werden Personen informiert, wann und auf welche Weise sie mit einer KI-Anwendung interagieren?	◯ Ja	◯ Nein

Checkliste – Accountability

Rechenschaftspflicht

Sind klare Verantwortlichkeiten für Entwickler:innen, Betreiber:innen und Nutzer:innen der KI-Anwendung festgelegt?	◯ Ja	◯ Nein
Wurde festgelegt, wer die letztendliche Verantwortung und Rechenschaftspflicht für den KI-Einsatz sowie die Ausgaben des KI-Systems trägt?	◯ Ja	◯ Nein

ABBILDUNG 2 **Übersicht der Potenziale und Herausforderungen eines öffentlichen KI-Registers**

Potenziale
- Mehr Transparenz über KI-Anwendungen im öffentlichen Sektor
- Wissensaustausch innerhalb der Verwaltung

Herausforderungen
- Wenig Anwendungsfälle für ein KI-Register
- Fehlende Informationstiefe in einem KI- Register
- Mehrwert muss Ressourcenaufwand rechtfertigen

Quelle: eigene Darstellung

BertelsmannStiftung

Transparenz – KI-Register

- Organisationsübergreifende Sammlung von KI-Anwendungen

NETZPOLITIK.ORG

Aufsicht und Transparenz

Wie die Niederlande aus KI-Skandalen lernen

Die Niederlande wollen vormachen, wie sich automatisierte Entscheidungssysteme einhegen lassen. Skandale wie die Kindergeldaffäre sollen mit einer neuen Algorithmenaufsicht und Transparenzregistern verhindert werden. Davon könnte sich die EU eine Scheibe abschneiden und den AI Act verbessern.

01.02.2023 um 17:34 Uhr - Tomas Rudl - in Nutzerrechte

Amsterdam sucht unter anderem teil-automatisiert nach falsch Parkenden. Die Gefahren solcher KI-Systeme soll eine Algorithmenaufsicht und ein Transparenzregister entschärfen.

– CC-BY-NC-SA 2.0 Foto: harry_nl / Bearbeitung: netzpolitik.org

Vor dem Einsatz: Durchführung einer Ethik-Folgenabschätzung (EIA)

SCOPING QUESTIONS

1. Project Description
2. Proportionality Screening and Do No Harm
3. Project Governance (Establishing Roles and Responsibilities)
4. Multistakeholder Governance

IMPLEMENTING THE UNESCO PRINCIPLES

5. Safety and Security
6. Fairness, Non-Discrimination, Diversity
7. Sustainability
8. Privacy and Data Protection
9. Human Oversight and Determination
10. Transparency and Explainability; Accountability and Responsibility
11. Awareness and Literacy

(https://doi.org/10.54678/YTSA7796)

‚KI-Red-Teaming‘

- Bürgerservice: Manipulation von Chatbots (Prompt Injection)
- Sozialwesen: Prüfung auf algorithmische Diskriminierung
- Justiz: Validierung von Entscheidungshilfen
- Öffentliche Sicherheit: Umgehung von Gesichtserkennung

(https://www.tuvit.de/de/themen/ki-sicherheit/ki-leistungen/)

Leistungsabfall von KI-Systemen über die Zeit

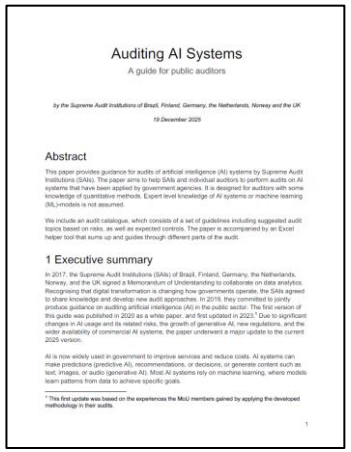
- **Sozialwesen: Betrugserkennung bei Sozialleistungen**
  - Algorithmen, um Profile mit hohem Risiko für Leistungsbetrug zu identifizieren
  - Während einer Wirtschaftskrise oder einer Pandemie ändern sich die Konsummuster und Lebensumstände der Bevölkerung massiv (Stromverbrauch, Adresswechsel)
- **Steuerverwaltung: Erkennung von Steuerhinterziehung**
  - Finanzämter nutzen Modelle, um Anomalien in Steuererklärungen zu finden
  - Neue Steuertricks, Schlupflöcher oder die Einführung von Kryptowährungen verändern, wie Hinterziehung aussieht

Was ist Modelldrift?

Modelldrift bezieht sich auf die Verschlechterung der Leistung eines maschinellen Lernmodells aufgrund von Änderungen in den Daten oder in den Beziehungen zwischen Eingabe- und Ausgabevariablen. Modelldrift, auch als Modellzerfall bekannt, kann sich negativ auf die Modellleistung auswirken und zu fehlerhaften Entscheidungen und schlechten Vorhersagen führen.

(https://www.ibm.com/de-de/think/topics/model-drift)

KI-Audits


Auditing AI Systems

A guide for public auditors

by the Supreme Audit Institutions of Brazil, Finland, Germany, the Netherlands, Norway and the UK

19 December 2025

Abstract

This paper provides guidance for audits of artificial intelligence (AI) systems by Supreme Audit Institutions (SAIs). The paper aims to help SAIs and individual auditors to perform audits on AI systems that have been applied by government agencies. It is designed for auditors with some knowledge of quantitative methods. Expert-level knowledge of AI systems or machine learning (ML)-models is not assumed.

We include an audit catalogue, which consists of a set of guidelines including suggested audit topics based on risks, as well as expected controls. The paper is accompanied by an Excel helper tool that sums up and guides through different parts of the audit.

1 Executive summary

In 2017, the Supreme Audit Institutions (SAIs) of Brazil, Finland, Germany, the Netherlands, Norway, and the UK signed a Memorandum of Understanding to collaborate on data analytics. Recognising that digital transformation is changing how governments operate, the SAIs agreed to share knowledge and develop new audit approaches. In 2019, they committed to jointly produce guidance on auditing artificial intelligence (AI) in the public sector. The first version of this guide was published in 2020 as a white paper, and first updated in 2023.[1] Due to significant changes in AI usage and its related risks, the growth of generative AI, new regulations, and the wider availability of commercial AI systems, the paper underwent a major update to the current 2025 version.

AI is now widely used in government to improve services and reduce costs. AI systems can make predictions (predictive AI), recommendations, or decisions, or generate content such as text, images, or audio (generative AI). Most AI systems rely on machine learning, where models learn patterns from data to achieve specific goals.

[1] This first update was based on the experiences the MoU members gained by applying the developed methodology in their audits.

1

- Richtlinien der Rechnungshöfe von BRA, FIN, GER, NED, NOR und UK
- Herausforderungen:
  - Fokus seitens Entwicklern eher auf technischen Aspekten
  - Kommunikation zwischen Entwicklern und Betreibern
  - Mangelnde KI-Entwicklungsfähigkeiten
  - Nutzung persönlicher Daten in KI-Systemen unklar (Verantwortungskette)
- Expertise von Rechnungsprüfer.innen in folgenden Bereichen: AI-Systemen und Anwendungsfällen, AI/ML-Kenntnisse, Coding/Implementierung, Unterstützung von KI-Systemen durch Cloud-Services

KI-Audits – Zertifizierungen

- Information Systems Audit and Control Association
- Berufsverband für IT-Revisor.innen, Informations-sicherheitsmanager.innen und IT-Governance-Expert.innen

Audit des Quellcodes von KI-Anbietern

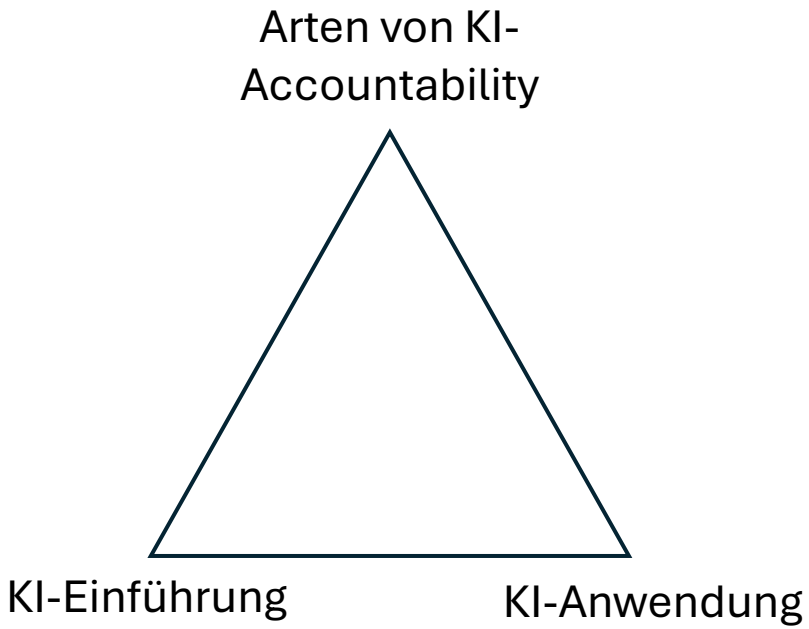
- z.B. BSI AIC4 (Artificial Intelligence Cloud Services Compliance Criteria Catalogue)
- Fokus: Sicherheit von KI-Diensten in der Cloud
- Prüfung des gesamten Lebenszyklus: **Entwicklung (Code)**, Datenqualität, Robustheit gegen Angriffe (z.B. Adversarial Attacks) und Erklärbarkeit.

Bundesamt für Sicherheit in der Informationstechnik

Deutschland Digital•Sicher•BSI•

Takeaways

- Technologie-Anwendungen (zentrale Unterschiede bez. Transparenz in verschiedenen KI-Anwendungen)



(Sinclair 1995)

Danke für die Aufmerksamkeit!

SGVW
DIE WISSENSPLATTFORM
IM ÖFFENTLICHEN SEKTOR

(Danke an Laura Waltersdorfer vom ‚WU Competence Center for Applied AI and Scientific Computing' für Input zu dieser Session!)