



Das Webinar wird aufgezeichnet.

KI aktiv gestalten oder beherrscht werden?

Wie Organisationen KI-Anwendungen beurteilen und sinnvoll einsetzen können

Webinar im Rahmen des Projektes Austrian Lab for AI Trust – ALAIT*

15. September 2026
09:30 bis 12:00 Uhr | zoom

Das Projekt ALAIT wird im Auftrag des Bundesministeriums für Innovation, Mobilität und Infrastruktur umgesetzt.

winnovation TU WIEN leiwand.ai APA

1



Nutzungsbedingungen

Diese Slides stehen unter der Creative Commons Lizenz CC BY-NC-ND 4.0
<https://creativecommons.org/licenses/by-nc-nd/4.0/deed.de>
 Verwendung nur unter:

- **Namensnennung**
- **Nicht kommerzielle Nutzung**
- **Keine Bearbeitungen**

Das Projekt ALAIT wird im Auftrag des Bundesministeriums für Innovation, Mobilität und Infrastruktur umgesetzt.

winnovation TU WIEN leiwand.ai APA

CC BY NC ND

2

Begrüßung

Michael Wiesmüller

Leiter der Abteilung III/5 – Digitale- und Schlüsseltechnologien für industrielle Innovation

Bundesministerium für Innovation, Mobilität und Infrastruktur

Austrian Lab for AI Trust (ALAIT)

3

Thematische Einführung

Brigitte Ömer-Rieder

Moderatorin, winnovation consulting gmbh

Austrian Lab for AI Trust (ALAIT)

4

4

Einige aktuelle KI-Schlagzeilen ...



KI-Firma Anthropic wirft chinesischen Rivalen Schummelei vor

dpa, 11.09.2026

OpenAI behauptet, ein 90 Jahre altes Mathematik-Problem in 88 Stunden gelöst zu haben

The Download from MIT Technology Review, 09.09.2026

Nvidia will Hugging Face für knapp 13 Milliarden Dollar übernehmen

Der Spiegel, 03.09.2026

Rechenzentren boomen, sind aber verstärkt von Naturkatastrophen bedroht

Handelsblatt, 12.08.2026

Austrian Lab for AI Trust (ALAIT)

5

5

**Wer baut
die größten KI-Modelle,
die größten Rechenzentren,
die größten Plattformen?**



Austrian Lab for AI Trust (ALAIT)

6

6

KI aktiv gestalten, statt von ihr beherrscht werden



Foto credit: Canva

Austrian Lab for AI Trust (ALAIT)

7

7



Foto credit: Canva

Auf der Ebene von Organisationen geht es um Innovation mit Qualität:

Wie bringen wir Künstliche Intelligenz

- nutzenstiftend,
 - sicher und
 - zuverlässig
- in die Anwendung?

Welche KI-Systeme sind vertrauenswürdig und erfüllen die jeweiligen Anforderungen?

Und wie wird eine Organisation „ready“ für KI?

Austrian Lab for AI Trust (ALAIT)

8

8

Genau das ist der Beitrag, den wir mit ALAIT leisten



**Vertrauen durch Wissen:
Gemeinsam KI verstehen
und gestalten**

**ALAIT
Austrian Lab
for AI Trust**

Forschungs- und Entwicklungsprojekt ALAIT - Austrian Lab for Artificial Intelligence Trust

Das F&E-Projekt ALAIT entwickelt ein innovatives ALAIT Technology Impact Assessment für Künstliche Intelligenz und fördert durch das ALAIT Labor und ein Train-the-Trainer-Netzwerk den öffentlichen Diskurs über KI-Vertrauen. Ergebnisse werden in AI-Trust-Dossiers veröffentlicht und breit gestreut. Unser Ziel ist es, Vertrauen in KI-Technologie durch Transparenz und Information zu stärken.

Projektlaufzeit: September 2024 – Jänner 2027

Die Projektergebnisse – ALAIT-Trust-Dossiers & Branchen-Reports – stehen zum Download bereit.

Website: <https://science.apa.at/project/alait/>

9

9

Austrian Lab for AI Trust

Konsortium:



im Auftrag von:



Vertrauen in KI

basiert auf 3 Säulen

Transparenz
Zugänglichkeit
von Information

Governance
rechtliche und
soziale Normen

Literacy
Wissen und
Kompetenz

Neue Methoden und Werkzeuge
für die Risikobeurteilung von
KI-Systemen in der Praxis

1. **ALAIT Technology Impact Assessment:**
 - KI-Risikoradar
 - Ex-ante Risikobewertung von KI-Systemen
2. **ALAIT Dossiers: Bewertung von 8 verschiedenen KI-Technologien**
3. **ALAIT Labore mit Branchen**

10

10

Warum steht Impact- und Risikobewertung im Zentrum?



- 1 **Risikobewertung prüft, konkretisiert und verbessert KI Use Cases und die KI Governance** in Organisationen



- 2 **Risikobewertung ist Teil der Erfüllung von Gesetzen und Normen**
 - Regulatorische Anforderungen durch EU AI Act, Haftungsregeln, DORA, DSGVO u.a.
 - Freiwillige Umsetzung von KI-Qualitätsstandards und KI-Zertifizierung

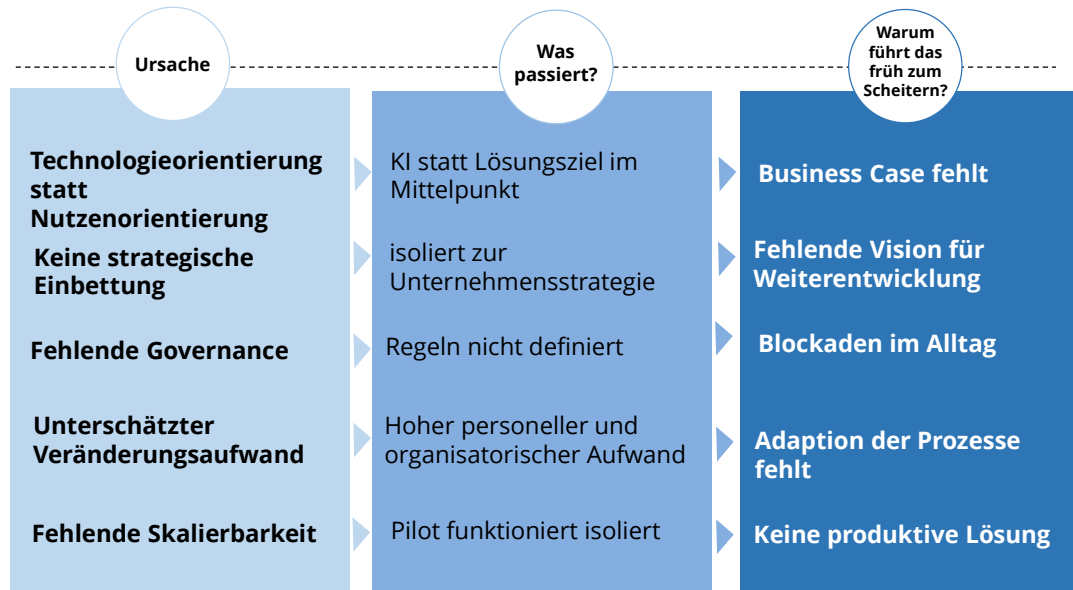
Mehr Qualität und Erfolg in der KI-Implementierung

von Low Risk bis zu High Risk KI-Werkzeugen

In 95% der Unternehmen generieren GenAI-Piloten keinen Wert

Quelle: MIT: The GenAI Divide, State of AI in Business 2025

Woran KI-Projekte in Organisationen scheitern



Ergebnisse aus Umfragen und wissenschaftlichen Studien 2025-2026

13

13

Zielsetzung für das heutige Webinar

- **Neue Methoden für die KI-Praxis vorstellen**
- **Anwendungsbeispiele geben**
- **Überblick und Inspiration geben: Umgang mit KI-Risiken und Möglichkeiten der Governance**

Zielgruppen:

- **Multiplikator:innen, Mitarbeiter:innen von Unternehmen und öffentlicher Verwaltung**

Austrian Lab for AI Trust (ALAIT)

14

14

Agenda

- 09.30 Beginn
 - Begrüßung**
Michael Wiesmüller, BMIMI
 - Einführung**
Brigitte Ömer-Rieder, winnovation – Moderation
- 10:00 Keynote: **KI-Risiken frühzeitig identifizieren und entschärfen: Die ALAIT ex-ante KI-Risikobewertung**
Gertraud Leimüller, leiwand.ai
- 11:00 Kurzimpuls: **KI-Risiken im Kontext**
Sabine Köszegi, TU Wien
- 11:20 Kurzimpuls: **Überblick über zentrale technologische Risiken bei KI**
Rania Wazir, leiwand.ai
- 11:35 Good Practice: **KI-Governance in der APA**
Verena Krawarik, Austria Presse Agentur
- 12.00 Abschluss und Ausblick

Austrian Lab for AI Trust (ALAIT)

15

15

Keynote

KI-Risiken frühzeitig identifizieren und entschärfen: Die ALAIT ex-ante KI-Risikobewertung für Praktiker:innen

Gertraud Leimüller, CEO leiwand.ai

Austrian Lab for AI Trust (ALAIT)

16

16



17



18

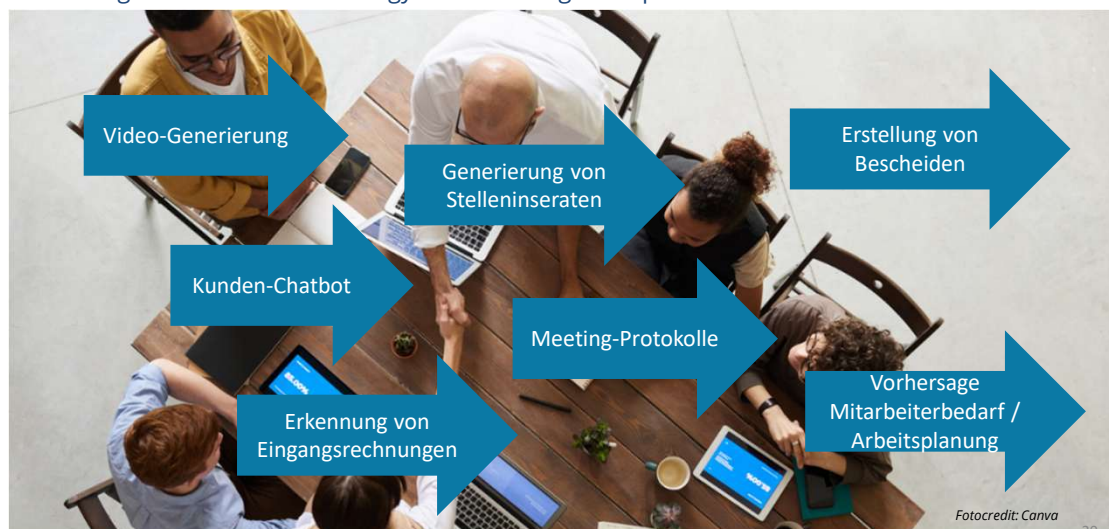


19

19

Wie kommen KI-Qualität und KI-Sicherheit in die Organisation?

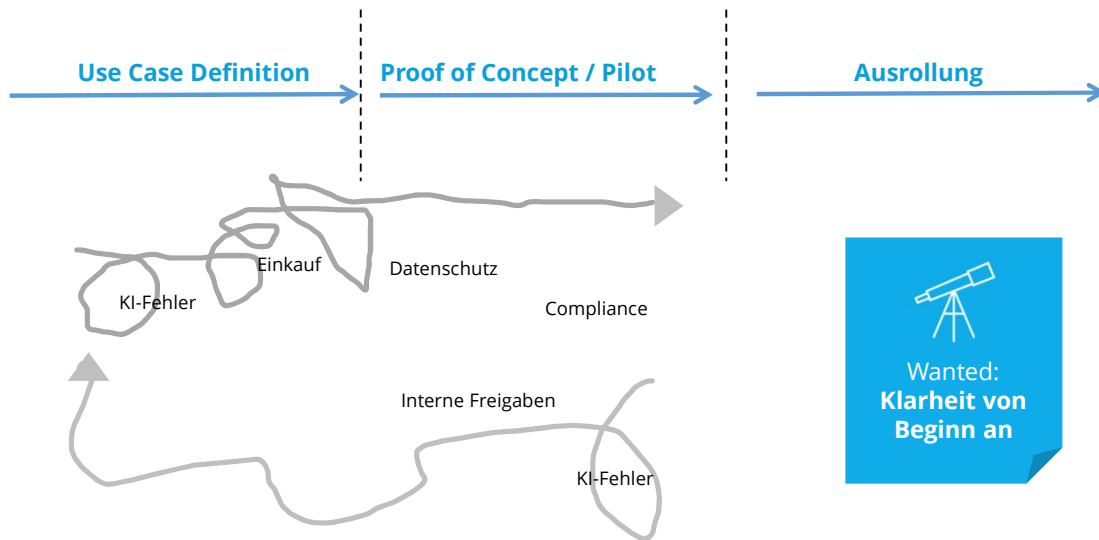
Erfahrung in der Praxis: Technology Push – KI-Angebote prasseln auf Unternehmen nieder



20

20

KI Use Cases: Der lange Weg von der Idee zur Umsetzung

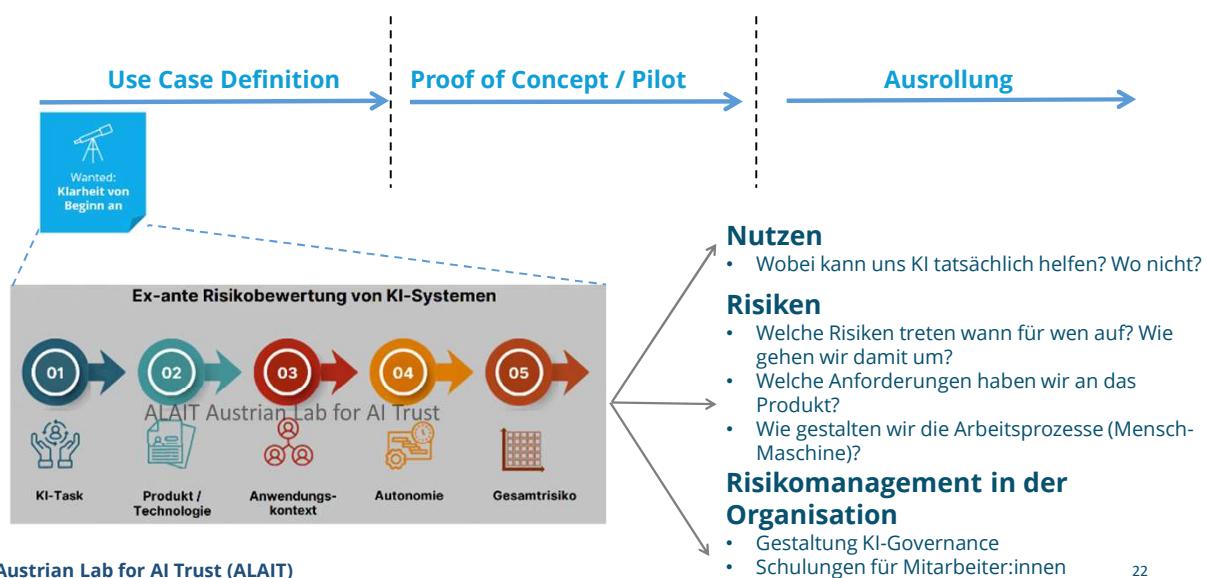


Austrian Lab for AI Trust (ALAIT)

21

21

KI Use Cases: Der lange Weg von der Idee zur Umsetzung



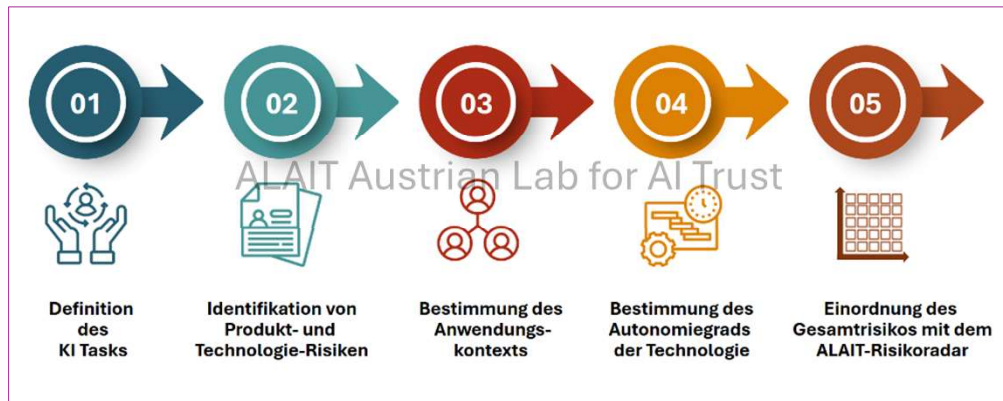
Austrian Lab for AI Trust (ALAIT)

22

22

Übersicht:

Die 5 Schritte der ALAIT ex-ante Risikobewertung von KI-Systemen



Austrian Lab for AI Trust (ALA IT)

23

23

Beispiel: KI-Unterstützte Erstellung von Meeting-Protokollen



24

24



Beispiel:
KI-unterstützte Erstellung von Meeting-Protokollen

Schritt 1: Definition der KI-Aufgabe (Task)

„Wir möchten jetzt auch KI einsetzen“

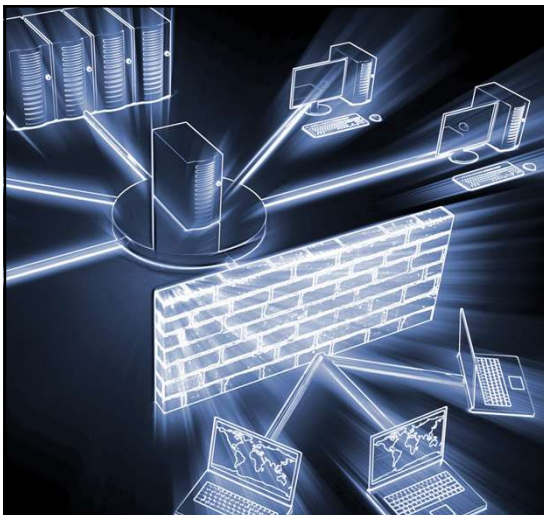
„Wir wollen den Zeitaufwand der Mitarbeiter:innen für das Schreiben von Meeting-Protokollen reduzieren“



- Aufgabe(n) eng und konkret definieren
- Aktuelle Herausforderungen der Nutzer:innen erkunden (zB Workshops, Befragungen)

25

25



Beispiel:
KI-unterstützte Erstellung von Meeting-Protokollen

Schritt 2: Identifikation von Produkt- und Technologierisiken

- Was tut ein KI-Produkt, das Meeting-Protokolle erstellt? Identifikation der **wesentlichen technologischen Teilschritte**
- Welche Risiken sind mit diesen Teilschritten jeweils verbunden? Benennung der inhärenten Risiken



- Desktop-Recherche
- Auswertung von Produktbeschreibungen und Model Cards
- Literatur-Auswertung

26

26

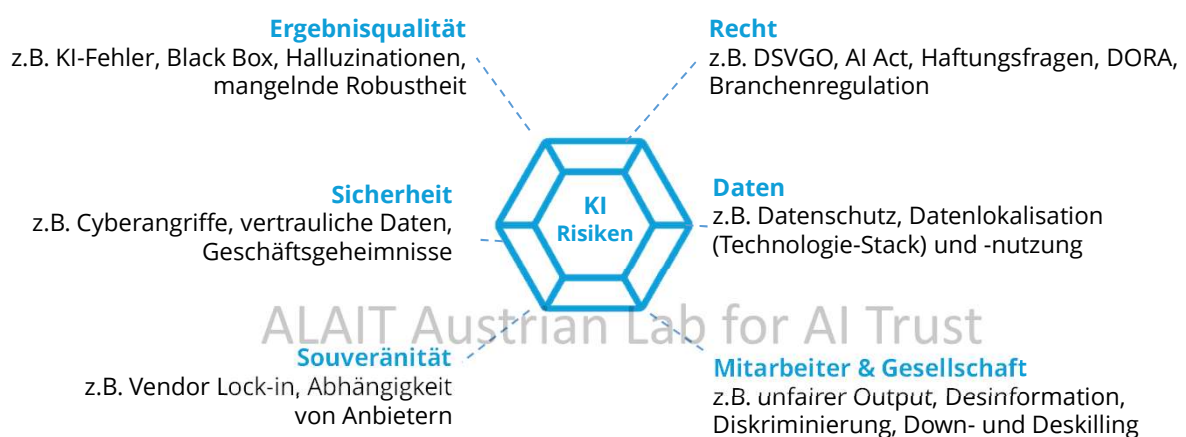
Was tut ein KI-System, das Meeting-Protokolle erstellt? Welche Risiken sind mit diesen Teilschritten verbunden?



27

27

Dimensionen des KI-Risikos für Organisationen



Austrian Lab for AI Trust (ALAIT)

28

28



Beispiel:
KI-unterstützte Erstellung von Meeting-Protokollen

Schritt 3: Bestimmung des Anwendungskontexts

1. Wieviel Sicherheit für Personen erfordert das Einsatz-gebiet und wie hoch ist das Risiko laut EU AI Act?
2. Sind sensible Daten von Personen und der Organisation betroffen?
3. Werden externe Betroffene negativ beeinflusst, z.B. Kund:innen, die Öffentlichkeit oder ges. Gruppen?
4. Welche Schäden können in der eigenen Organisation oder bei Dritten entstehen, z.B. rechtliche, finanzielle oder reputationsbezogene Schäden?

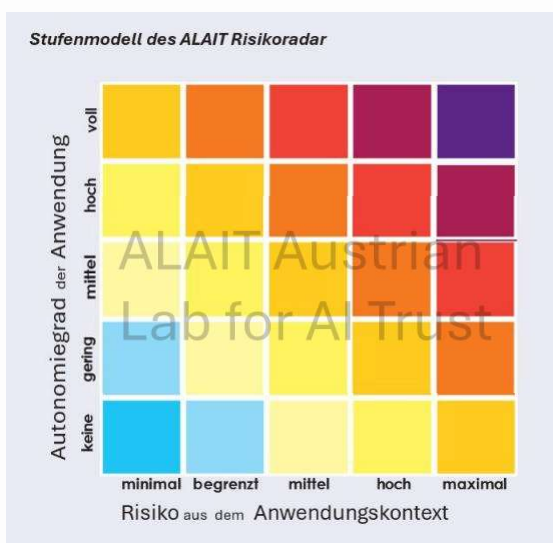


29

29

ALAIT Risikoradar

Stufenmodell des ALAIT Risikoradar



Austrian Lab for AI Trust (ALAIT)

Risiko aus dem Anwendungskontext

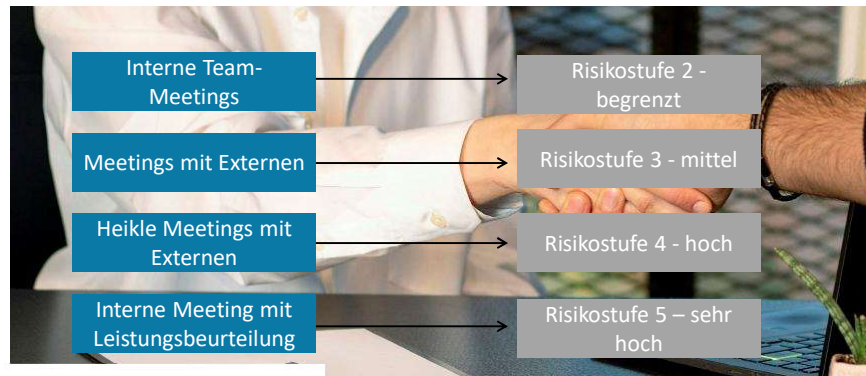
- **1 - Minimal:** KI-System hat keine Auswirkung auf nutzende Personen oder deren Entscheidungsfindung
- **2 - Begrenzt:** KI-Systeme interagieren mit Menschen, treffen aber keine Entscheidungen
- **3 - Mittel:** KI-Systeme beeinflussen kollektive, gesellschaftliche Ebene
- **4 - Hoch:** Hochrisiko-KI-Systeme laut EU AI Act oder mit direkten Auswirkungen auf Leib und Leben von Personen
- **5 - Maximal:** Hochrisiko-KI-Systeme laut EU AI Act, wenn Regeln für Risikomanagement nicht eingehalten werden

30

30

Unterschiedliche Ausprägungen des Anwendungskontexts

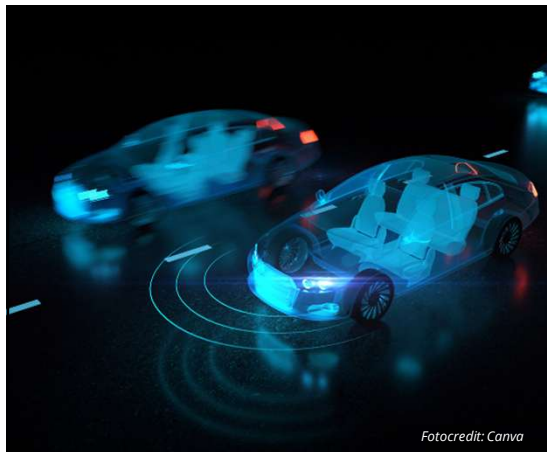
Beispiel: KI-unterstützte Erstellung von Meeting-Protokollen



- Branche, Organisation, Nutzer:innen und betroffene Personen berücksichtigen!
- Risiken aus Schritt 2 mitnehmen!
- Risikoklassen AI Act als erste Annäherung

31

31



Fotocredit: Canva

Beispiel:
KI-unterstützte Erstellung von Meeting-Protokollen

Schritt 4: Faktische Autonomie der Technologie

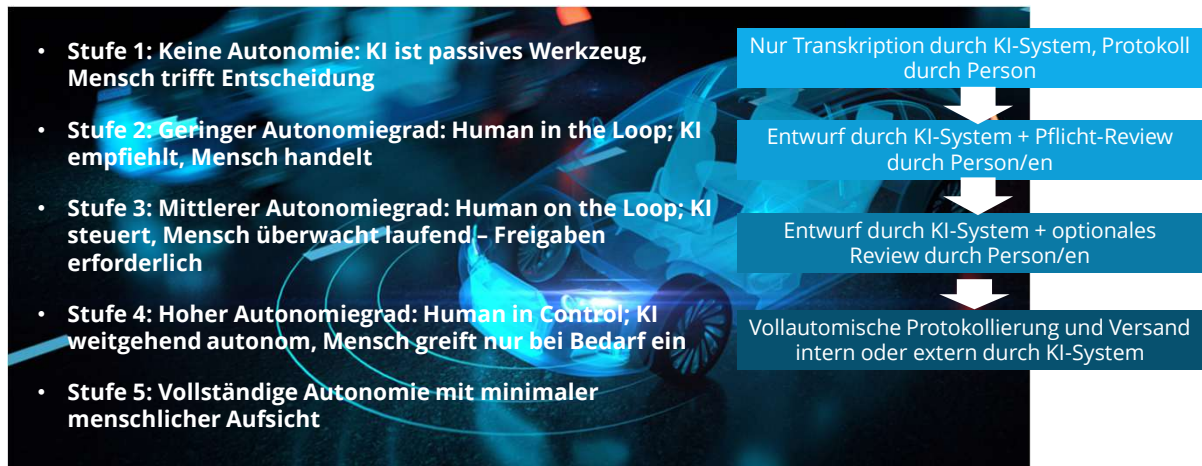
1. Gestaltung des Arbeitsprozesses: Welche Arbeitsschritte und Entscheidungen übernimmt KI-System, was bleibt beim Menschen?
2. Welche Eigenschaften soll das zu entwickelnde oder zu kaufende KI-System (samt der dahinterliegenden Infrastruktur) folglich haben?
3. Wie aktiv und unabhängig führen Menschen ihre Entscheidungs- und Aufsichtsrollen aus? Wie werden sie dabei durch ihre Vorgesetzten, Zeitressourcen und Schulungen unterstützt?



32

32

Unterschiedliche Autonomiegrade



33

33

Stufenmodell des ALAIT Risikoradar



Beispiel:
KI-unterstützte Erstellung von Meeting-Protokollen

Schritt 5: Einordnung in das Risikoradar

Farbskala zeigt das Gesamtrisiko:

- Violett und dunkelrot - sehr hoch
- rot und dunkelorange – hoch
- hellorange und Gelbtöne – mittel
- Blautöne - geringes Gesamtrisiko

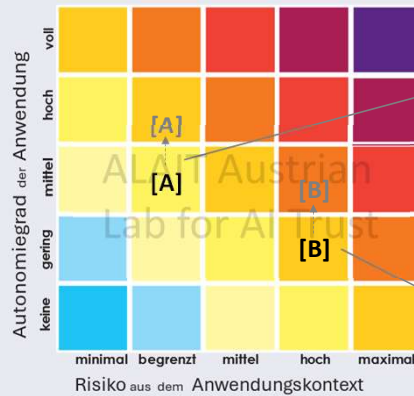


34

34

Ergebnis für: KI-unterstützte Erstellung von Meeting-Protokollen

Stufenmodell des ALAIT Risikoradar



Schritt 5: Einordnung des Gesamtrisikos

A: Interne Meeting-Protokolle

- **A (schwarz):** Automatische Versendung des KI-generierten Protokolls nach 24 Stunden, in denen zuständige Person Korrekturen vornehmen kann – und gemäß Vorgaben in der Organisation auch soll.

- **A (grau):** Wäre – in späterer Version - automatische Versendung ohne Überprüfung durch Menschen

Begleitend:

**Erprobung,
Rollen-
zuordnung,
Monitoring,
Schulung**

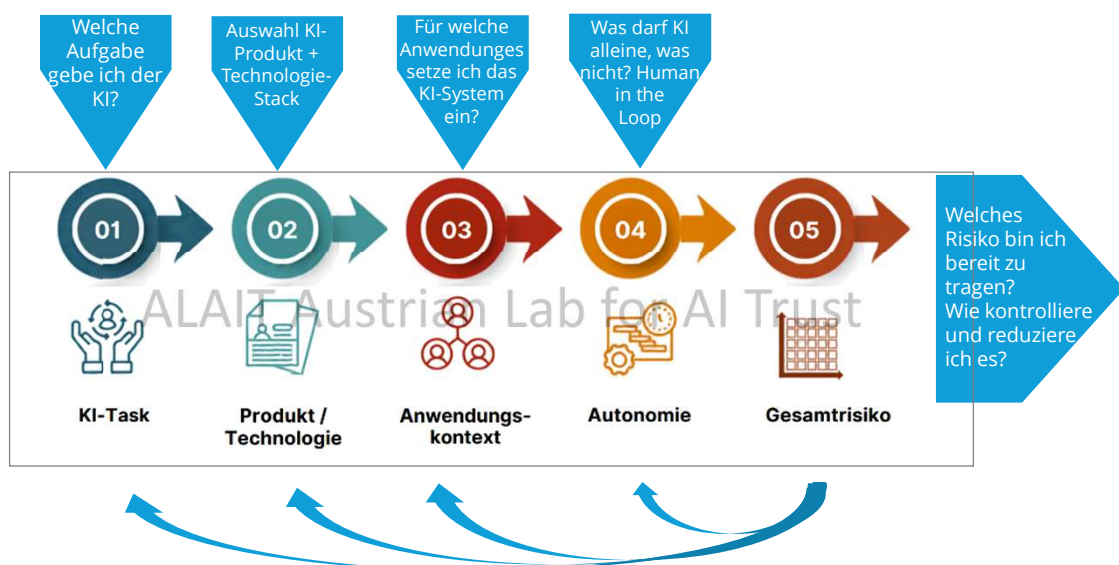
B: Heikle Meeting mit Externen

- **B (schwarz):** zwingende Freigabe durch Protokollverantwortliche Person vor dem Versand.
- **B (grau):** Automatische Versendung des KI-generierten Protokolls nach 24 Stunden, in denen verantwortliche Person Korrekturen vornehmen kann – und gemäß Vorgaben in der Organisation auch soll.

35

35

KI-Risiko ist eine Gestaltungsentscheidung – keine technische Entscheidung



Austrian Lab for AI Trust (ALAIT)

36

36

Maßnahmen zur Risikominimierung

- ✓ **Sorgfältige Auswahl** des KI-Systems mit **interner Testphase**
- ✓ **Ausschluss** des Trainings des KI-Systems durch **Input-Daten aus der Organisation** (Gefahr des Verlusts von Geschäftsgeheimnissen und vertraulichen Daten)
- ✓ **DSGVO-konforme** Architektur des Gesamtsystems durch Hosting in **EU-Data Center** mit europäischem Betreiber, Privacy-Standards wie Verschlüsselung und möglichst keine (oder nur kurzfristige) Speicherung von Audio- und Videodaten beim Anbieter
- ✓ Ausdrückliche **Zustimmung von Mitarbeiter:innen und Kund:innen** zur Datenverarbeitung vor Meetings; alternativer Weg (manuelle Protokollerstellung) wenn keine Zustimmung
- ✓ Einholung des Einverständnisses durch **Betriebsrat**
- ✓ **Kennzeichnung** von KI-generierten Protokollen
- ✓ Wiederkehrende **Schulungen**, damit Mitarbeiter:innen aktive Kontroll- und Aufsichtsfunktion ausüben können
- ✓ **Kontroll- und Widerspruchsmöglichkeiten** klar definieren und breit kommunizieren

Austrian Lab for AI Trust (ALAIT)

37

37

KI-Risiken im Kontext

Querschnittsanalyse der ALAIT-Dossiers 1 bis 8

Was Organisationen aus dem Risikoradar für die eigene KI-Governance mitnehmen sollen



Sabine Köszegi · TU Wien · Austrian Lab for AI Trust

38

Risiko ist kein Zustand, sondern eine Zuschreibung

Wer entscheidet, geht ein Risiko ein.

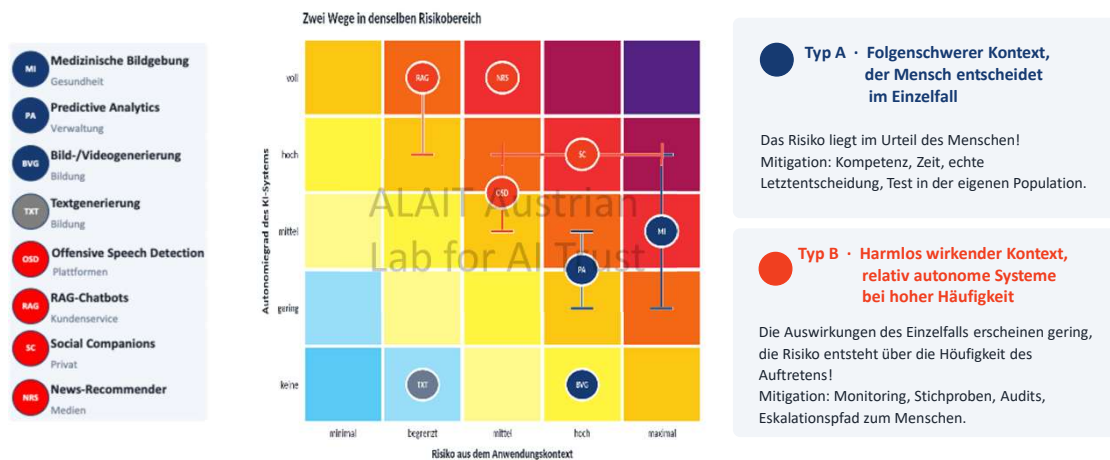
Wer betroffen ist, ohne entschieden zu haben, ist einer Gefahr ausgesetzt.

„Entweder wird der etwaige Schaden als Folge der Entscheidung gesehen, also auf die Entscheidung zugerechnet. Dann sprechen wir von Risiko. Oder der etwaige Schaden wird als extern veranlasst gesehen. Dann sprechen wir von Gefahr.“

Niklas Luhmann, Soziologie des Risikos, 1991, S. 30 f.

39

Gleiches Risiko, zwei verschiedene Ursachen



40

Technologie- inhärente vs. kontext- abhängige Risikocluster

	A Epistemisch Richtigkeit des Outputs	B Verteilung Bias und Diskriminierung	C Intransparenz Black Box, Rechenschaft	D Daten und Sicherheit DSGVO, Security, IP	E Mensch-Maschine Automation Bias, Deskilling	F Kollektiv- Demokratische Ressourcen
MI Medizinische Bildung	•	•	•	•	○	—
OSD Offensive Speech Detection	•	•	•	○	○	•
PA Predictive Analytics	•	•	•	•	•	○
RAG RAG-Chatbots	•	○	•	•	•	○
NRS News Recommender	•	•	○	•	•	•
SC Social Companions	•	•	○	•	•	•
RVG Bild-/Viduongenerierung	•	•	○	•	•	•
TXT Textgenerierung	•	○	○	○	•	•

• als eigener Risikopunkt ausgeführt ○ im Text oder in den Empfehlungen adressiert — nicht behandelt

Risikocluster A bis E sind technologie-inhärent

Technologie-inhärente Risiken treten unabhängig von Domäne und Autonomiegrad auf.

Risikocluster F ist kontextabhängig

Kollektive Risiken tauchen dort auf, wo KI Öffentlichkeit, Beziehungen oder Lernen vermittelt

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

41

41

Kernrisiken

8/8

Richtigkeit

Halluzination, Fehlklassifikation,
Data Shift zwischen Test und
Einsatz

8/8

Bias

Fehler häufen sich systematisch bei
denselben Gruppen

8/8

Intransparenz

Black Box – auch für die
Fachperson, die sie verantwortet

8/8

Datenschutz

personenbezogene Daten, Security-
Lücken, Urheberrecht

Skalierung als unsichtbare Dimension

Nach EU AI Act nur „begrenztes Risiko“

RAG-Chatbots · Textgenerierung

Stufe 2 von 5. Es gilt im Wesentlichen eine Transparenzpflicht:
Nutzer:innen müssen erfahren, dass sie mit einer KI sprechen.

Genau diese beiden werden ohne Freigabeprozess breit eingeführt –
weil sie als unkritisch gelten.

Ein Fehler mit geringer Einzelfolge, millionenfach wiederholt und von
niemandem bemerkt, ist kein kleines Risiko.

94 %

der 11- bis 17-Jährigen in Österreich nutzen KI-Chatbots

52 %

der 16- bis 54-Jährigen nutzten 2025 generative KI – 2024
waren es 37 %

47 %

der Gespräche Studierender mit KI sind „direkt“: Antwort
ohne Eigenleistung

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

42

42

Wer profitiert?

Wer trägt das Risiko?

MI	Fachpersonal und Verwaltung: bis 62 % Entlastung bei der Bildauswertung	Patient:innen aus unterrepräsentierten Gruppen
PA	Behörde: bis 40 % Arbeitsentlastung, Verfahren bis 31 % schneller	Antragsteller:innen, die Prüflast wandert zu ihnen
OSD	Plattformen: Moderation im Millionenmaßstab	marginalisierte Gruppen; Data-Labeler im Globalen Süden
NRS	Plattformen: über 30 % höhere Klickraten	Nutzer:innen und die demokratische Öffentlichkeit
TXT	Lernende kurzfristig: schneller bessere Ergebnisse	Lernende langfristig: Gedächtnis, Urteilskraft, Eigenständigkeit

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

43

43

Menschliche Aufsicht als Organisationsleistung

Die Rolle des Menschen ist für die Abwendung von Risiken durch KI Systeme zentral. Allerdings steht dem das Phänomen des „Automatisierungsbias“ entgegen!

17–33 %

Dossier 4

der Antworten KI-gestützter Rechtsrecherche-Tools waren ungenau oder erfunden – Nutzer:innen verlassen sich trotzdem darauf.

50 %

Dossier 6

häufiger als Menschen bestätigen gängige Chatbots die Handlungen ihrer Nutzer:innen (Sycophancy).

4 von 5

Dossier 3

Vorhersagen des Kopenhagener Pflegebedarfssystems trafen zu – die Sachbearbeitung sah der fünften nicht an, dass sie falsch war.

Damit menschliche Aufsicht gelingt, müssen Organisationen in Sicherheit investieren:

- **Zeit einplanen**
für die Überprüfung von KI Output (Meta-Arbeit!)
- **Investition in Mitarbeiter*innen-Kompetenz**
AI Literacy aufbauen und De-Skilling verhindern
- **Sichtbare Unsicherheit**
Echte Transparenz über Funktionsweise und Fehlerraten des Systems
- **Widerspruch ohne Nachteil:**
Abweichen vom KI-Output darf nicht bestraft werden

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

44

44

Leitfragen

- 1 Wen trifft ein Fehler und merken es diese Person sofort?**
Wenn die betroffene Person den Fehler nicht direkt bemerken kann, ersetzt kein Beschwerdeweg die eigene Kontrolle.
- 2 Kann die letztverantwortliche Person widersprechen, ohne Nachteil?**
Ohne Zeit, Kompetenz und Rückendeckung ist Aufsicht nominell.
- 3 Wurde das System der eigenen Einsatzumgebung getestet?**
Die Herstellergenauigkeit gilt für dessen Testdaten, nicht für Ihre Fälle.
- 4 Gibt es ein ständiges Monitoring über Fehleranfälligkeit??**
Ohne laufende Messung bleibt jede Risikoeinschätzung eine Momentaufnahme.
- 5 Wer haftet, wenn es schiefgeht?**
Die Organisation. Die Verantwortung ist rechtlich nicht an das System delegierbar.

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

45

45

Empfehlungen zur Umsetzung



© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

46

46

Maßnahmen – konkrete Beispiele

1 Abweichungsquote messen ●● Anteil der Fälle, in denen der Mensch dem Vorschlag widerspricht. Geht sie gegen null, ist die menschliche Aufsicht nur nominell.	2 Unabhängige Erstbeurteilung ● Die Fachperson bildet ihr Urteil, bevor der KI-Output angezeigt wird – erst danach der Abgleich (KI Support).
3 Konfidenzschwelle mit Eskalation ●● Unter einem definierten Schwellwert greift eine verpflichtende Zweitmeinung. Macht „sichtbare Unsicherheit“ zur Regel statt zum Prinzip.	4 Blinde Stichprobe ● Eine feste Quote automatisch verarbeiteter Fälle wird ohne KI nachbeurteilt. Die Differenz ist die gemessene Fehlerquote – nicht die des Herstellers.
5 Meldepflicht bei Modelländerungen ●● Vertragliche Meldung und Re-Validierung mit Herstellern absichern, damit ein zeitgerechtes, eigenes Monitoring bei Systemupdates möglich wird.	6 Anwendungsregister ●● Welche KI-Systeme laufen bei uns, wer verantwortet sie fachlich, sind wir Anbieter oder Betreiber, welche Frist gilt?

● Typ A Folgeschwerer Kontext, Mensch entscheidet

● Typ B Harmloser Kontext, System läuft in Masse

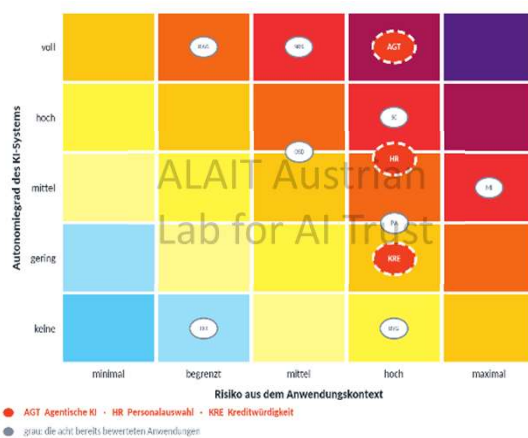
© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

47

47

Weitere Anwendungen (noch nicht) im Radar



- AGT Agentische KI** Systeme, die handeln statt antworten – buchen, bestellen, Code ausführen. Volle Autonomie in einem folgeschweren Kontext.
- HR Personalauswahl** Annex III Nr. 4: Filterung von Bewerbungen, Bewertung von Kandidat:innen, Leistungsbeurteilung.
- KRE Kreditwürdigkeit** Annex III Nr. 5 lit. b, dazu Risikobewertung und Preisbildung in der Lebens- und Krankenversicherung.

Potenzielle Erweiterungen des Radars

- Häufigkeit und Reichweite**
wie oft läuft es, wie viele trifft es
- Arbeitsplatz und Mitbestimmung**
Substitution, Qualifikationswandel
- Anbieterabhängigkeit**
Lock-in, Preise, Souveränität
- Restrisiko nach Maßnahmen**
wo landet der Fall danach

48

Key-Takeaways

1

Es gibt einen universellen Risikokern.

Richtigkeit, Bias, Intransparenz und Datenschutz sind immer zu klären – unabhängig von Domäne und Autonomiegrad.

2

Gleiches Risikoniveau heißt nicht gleiche Maßnahme.

Typ A braucht Urteilskraft beim Menschen.

Typ B braucht Kontrolle im laufenden Betrieb.

3

Wer den Nutzen einstreicht, trägt selten das Risiko.

Diese Frage gehört an den Anfang jeder Einführung – nicht in die Evaluierung danach.

© Koeszegi

Austrian Lab for AI Trust · Dossiers 1–8 · science.apa.at/project/alait

49

49



Überblick über zentrale technologische Risiken bei KI

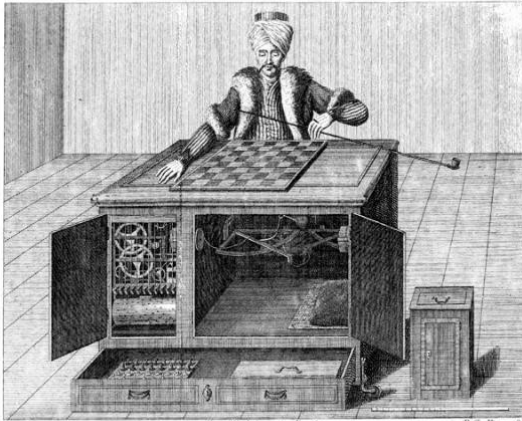
KI aktiv gestalten oder beherrscht werden?
ALAIT Webinar, 15.09.2026

Rania Wazir, Ph.D
Co-founder & CTO, leiwand.ai

50

50

Überblick über zentrale technologische Risiken bei KI



W. v. Kempelen del. Die Schachspieler des Hrn. von Kempelen. P.G. Tietz, sc.
Der Schachspieler des Hrn. von Kempelen, ein künstliches Spielzeug, welches die Kunst des Schachspiels nachzuahmen vermag.

Schachtürke

Erbaut von Wolfgang von Kempelen, 1770

Kupferstich aus: Briefe über den Schachspieler des Hrn. von Kempelen,
 Karl Gottlieb von Windisch [Public Domain](#)

Agenda

1. Verschiedene Arten von KI
2. Cybersecurity & Lock-in Effekte
3. Unverlässlichkeit
4. Wo bleibt die Qualitätskontrolle?
5. KI, aber richtig

51



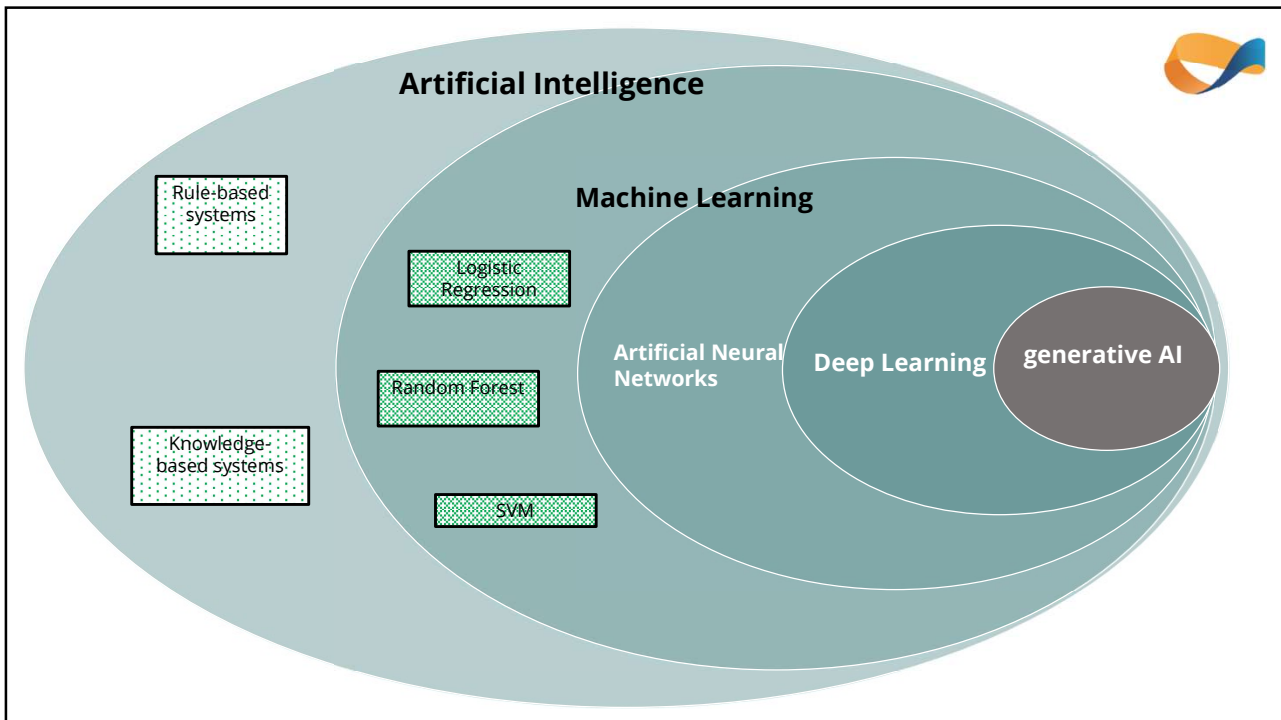
"AI is such an interesting word because it's sort of like the word **transportation** in that you have **bicycles**, you have **gas guzzling trucks**, you have **rocket ships**.

They're all forms of transportation, but they all serve different **purposes** and they have different **cost-benefit trade-offs**."

- Karen Hao, PBS Interview 06.2025

52

52



53



generative AI

- Multimodal AI (Text, Audio, Bild, Video, ...)
- Diffusion models (z.B. Stable Diffusion, SORA, ...)
- LLM (Text – z.B. chatGPT, Claude, DeepSeek, ...)

LLM

- Daten
- Transformer-Architektur (DL)
- Next-word prediction
- Self-supervised learning

Agentic AI

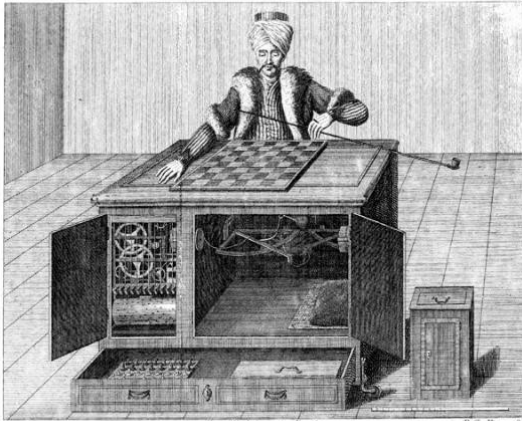
- LLM +
- „Autonomie“ +
- Zugriff auf Tools

Image: Deborah Lupton / <https://betterimagesolai.org/> / <https://creativecommons.org/licenses/by/4.0/>

54

54

Überblick über zentrale technologische Risiken bei KI



W. v. Kempelen del. Ch. a. Meubel van de Schaken P.G. Tritz. sc.
Der Schachspieler, ein künstlicher Mensch, der von einem Mechanismus angetrieben wird, der in einem Kasten verborgen ist.

Schachtürke
 Erbaut von Wolfgang von Kempelen, 1770

Kupferstich aus: Briefe über den Schachspieler des Hrn. von Kempelen,
 Karl Gottlieb von Windisch [Public Domain](#)

Agenda

1. Verschiedene Arten von KI
2. **Cybersecurity & Lock-in Effekte**
3. Unverlässlichkeit
4. Wo bleibt die Qualitätskontrolle?
5. KI, aber richtig

55



Cybersecurity Risiken

- Prompt Injection
- Offenlegung oder Verlust sensibler Informationen
- Supply Chain Risiken (Vulnerabilitäten eingebettet in Modell/Komponente)



<https://genaiowasp.org/resource/owasp-genai-lm-top-10-2026/>
 Image: Anne Fehres and Luke Corroy & Althmedia/
<https://betterimagesofai.org/> <https://creativecommons.org/licenses/by/4.0/>

56

56



Lock-in Effekte

- LLM als Infrastruktur: andere Anwendungen bauen darauf auf
- Plötzliche Unverfügbarkeit
- Keine Kontrolle über Modell-Updates, alte Modelle werden archiviert

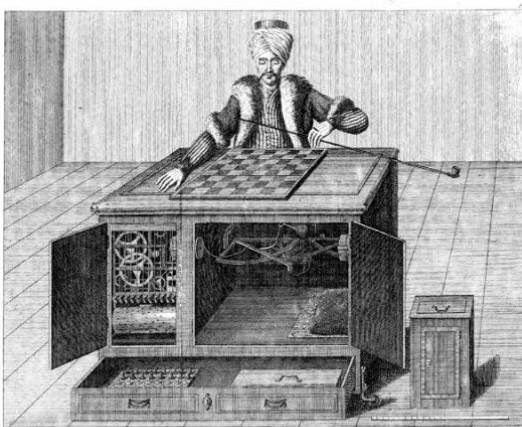


Image: Anne Fehres and Luke Conroy & AIA Media / <https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>

57

57

Überblick über zentrale technologische Risiken bei KI



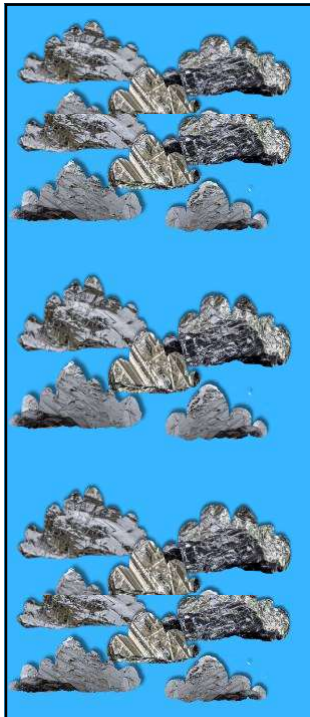
Schachtürke
Erbaut von Wolfgang von Kempelen, 1770

Kupferstich aus: Briefe über den Schachspieler des Hrn. von Kempelen,
Karl Gottlieb von Windisch [Public Domain](#)

Agenda

1. Verschiedene Arten von KI
2. Cybersecurity & Lock-in Effekte
- 3. Unverlässlichkeit**
4. Wo bleibt die Qualitätskontrolle?
5. KI, aber richtig

58



Unverlässlichkeit

- Daten
- Temperatur
- API & Cloud Infrastruktur
- Autonomie

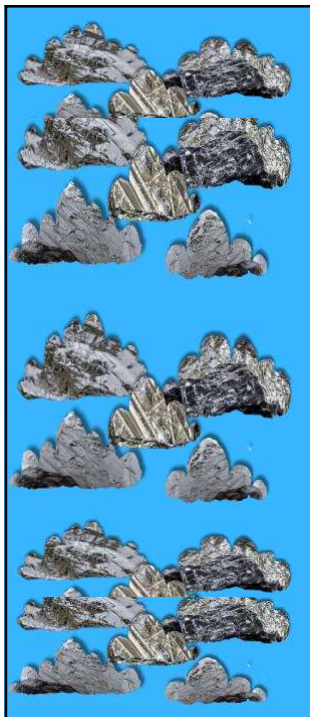
... *ein explosives Cocktail*



Defeating Nondeterminism in LLM Inference, H. He et al.
<https://thinkingsidelines.ai/blog/defeating-nondeterminism-in-llm-inference/>
 Image: Anne Fehres and Luke Conroy & A14Media /
<https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>

59

59



Unverlässlichkeit

- Plausibel ... aber nicht immer richtig
- Gleicher Input ... anderer Output
- Gleicher Input ... anderer Token-verbrauch
- „ähnlicher“ Input ... ?
- Biases und Verzerrungen
- Unbekanntes Fehlverhalten ...

... können Gesundheit und Sicherheit gefährden, bergen Diskriminierungsrisiken und können weitere Menschenrechte verletzen

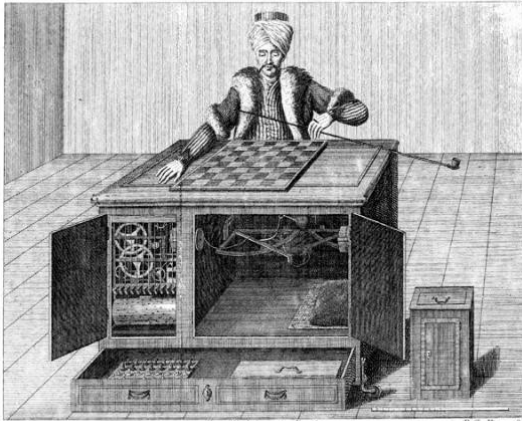


Towards a Science of AI Agent Reliability, Rabanser et al.
<https://arxiv.org/abs/2602.16966>
 Image: Anne Fehres and Luke Conroy & A14Media /
<https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>

60

60

Überblick über zentrale technologische Risiken bei KI



W. v. Kempelen del. C. a. M. del. sculp. P. G. Tietz, fecit.
Der Schachtürke, ein künstlicher Mensch, welcher Schach spielt, und welcher auch die Kunst des Schachspiels zu lehren vermag.

Schachtürke
 Erbaut von Wolfgang von Kempelen, 1770

Kupferstich aus: Briefe über den Schachspieler des Hrn. von Kempelen,
 Karl Gottlieb von Windisch [Public Domain](#)

Agenda

1. Verschiedene Arten von KI
2. Cybersecurity & Lock-in Effekte
3. Unverlässlichkeit
4. Wo bleibt die Qualitätskontrolle?
5. KI, aber richtig

61



Wo bleibt die Qualitätskontrolle?



Die Kultur des "move fast and break things"

"If people are just cutting corners all the time, if people are not in a culture that prioritizes safety and has appropriate incentives and structures, [accidents] are kind of bound to happen."

- **David Krueger**, computer science professor and prominent alignment expert, University of Montreal

Anthropomorphisierung plus Wettlauf-mindset lässt unverantwortliche KI-Entwicklungspraktiken zu.

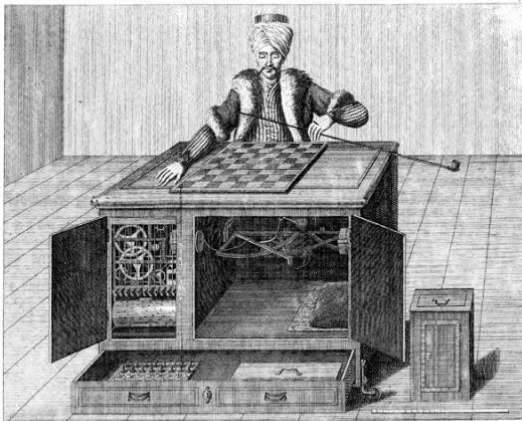
Schnelligkeit und Effizienz werden priorisiert über Genauigkeit, Reproduzierbarkeit, Transparenz, und Sicherheit.

Image: Sineem Gördü /
<https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>
 C. Hucks, "The Hugging Face hack could indicate cultural issues at OpenAI", 31.08.2026,
<https://www.technologyreview.com/2026/08/31/1143180/huggingfacehack-could-indicate-cultural-issues-at-openai/>

62

62

Überblick über zentrale technologische Risiken bei KI



*Il s. Kempelen del
Die Schachtürke, ein wunderbares Kunstwerk, das seit 1770 in Wien zu sehen ist. Dieses ist das Original, das in der Welt bekannt ist.
F.G. Tritz, fecit*

Schachtürke
Erbaut von Wolfgang von Kempelen, 1770

Kupferstich aus: Briefe über den Schachspieler des Hrn. von Kempelen,
Karl Gottlieb von Windisch [Public Domain](#)

Agenda

1. Verschiedene Arten von KI
2. Cybersecurity & Lock-in Effekte
3. Unverlässlichkeit
4. Wo bleibt die Qualitätskontrolle?

5. KI, aber richtig

63



KI, aber richtig

- Es muss nicht immer LLM oder Agentic sein. Manchmal sind andere Arten der KI besser
- Es gibt governance Prozesse, Standards, und technische Methoden um KI zu testen, und im Einsatz zu kontrollieren ...
- ... und auch diese werden ständig weiterentwickelt



Image: Lone Thomasky & Bics&Bäume /
<https://betterimagesofai.org/> / <https://creativecommons.org/licenses/by/4.0/>

64

64

AI Act: VERORDNUNG (EU) 2024/1689

Artikel 1

Zweck dieser Verordnung ist es, das Funktionieren des Binnenmarkts zu verbessern und 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, einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz**, vor schädlichen Auswirkungen von KI-Systemen in der Union zu gewährleisten und die Innovation zu unterstützen.

65

65

Standardisation request: C(2023) 3215 final

Risk Management
Systems

- CEN/CLC RMS

Data and data
governance

- CEN/CLC Bias
- CEN/CLC Datasets

Record keeping
through logging

- CEN/CLC TwF-1
- ISO/IEC Logging

Transparency for
users

- CEN/CLC TwF-1
- ISO/IEC
Transparency

Human oversight

- CEN/CLC TwF-1

Accuracy

- CEN/CLC TwF-2
- ISO/IEC Metrics
- ISO/IEC NLP
- CEN/CLC CV
- CEN/CLC Bias

Robustness

- CEN/CLC TwF-2
- ISO/IEC
Robustness

Cybersecurity

- CEN/CLC
Cybersecurity

Conformity
assessment

- CEN/CLC CA

Quality management
system

CEN/CLC QMS

66

66

Wichtige Projekte

NoLeFa: Vorbereitung der Umsetzung des EU AI Acts

COMBO: Automated, Multilingual Compliance Assistance

Roadguard: Digital Road User Safeguarding: A regulatory-compliant, contextual driver state assessment & road user warning system

Algorithmic Bias Risk Radar: Erkennung von Bias Risiken



67

67

*Danke für Ihre
Aufmerksamkeit!*

rania.wazir@leiwand.ai
www.leiwand.ai

68

68

KI-Governance in der APA

Verena Krawarik, Stabsstelle
Innovationsmanagement

69

AI-Governance: Struktur und Rollen

APA-KI-Leitlinie

KI-Kennzeichnungspflicht

Was wir von ihnen benötigen

Bildmetadaten und EU-Label

Internationale Learnings

Next Steps und Fragen



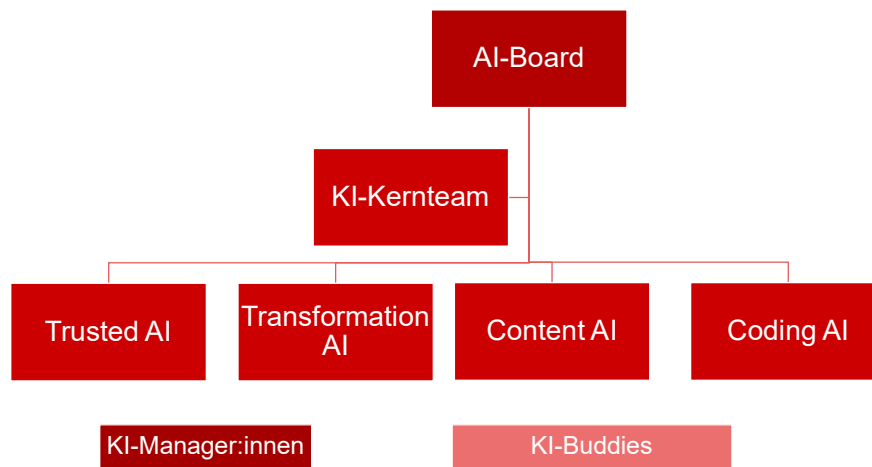
70



KI-Governance

71

AI-Governance: Struktur und Rollen



72

Spezielle Rollen

- **KI-Manager:innen:** unterstützen Geschäftsführer:innen und Bereichsverantwortliche in Entscheidungsprozessen und leiten KI-Projekte in den APA-Units.
 - Zertifiziert bei **Austrian Standards**
 - 13 Kolleg:innen

- **KI-Buddies:** unterstützen Abteilungsleiter:innen und Teams im Umgang mit KI-Tools, bei der Evaluierung neuer Plattformen sowie bei Weiterbildungsmaßnahmen (Multiplikatoren).
 - Ausbildung im Bereich Gen-AI
 - 36 Kolleg:innen (Abbild der APA im Kleinen)

73

APA KI-Leitlinie

Laufend aktualisierter Handlungsrahmen für die Entwicklung und den Einsatz von KI in der APA.

- Allgemeine Leitlinien (z.B. sensibler Umgang mit firmeninternen, datenschutz- oder urheberrechtlich geschützten Inhalten; ausschließlich geprüfte KI-Tools einsetzen)
- Bereichsspezifika (z.B. APA-Redaktion erstellt keine KI-generierten Bilder oder Videos und verbreitet diese auch nicht)
- 3. Update der KI-Leitlinie seit Juli 2026 in Kraft



74



Kennzeichnungspflicht

75

Ausgangslage

Beim Thema „**KI-Kennzeichnung**“ im Medienbereich treffen **unterschiedliche Erwartungen und Standards** aufeinander.

- **EU AI Act**

Der EU AI Act regelt in seinen Transparenzverpflichtungen (Artikel 50) die Kennzeichnung bestimmter KI-Erzeugnisse (Text, Bild, Video,...) und zwar für Hersteller von KI-Systemen als auch für deren Anwender. Insbesondere **fotorealistische Bilder oder Videos**, die mit KI erzeugt oder bearbeitet wurden, sind vom Anwender von Tools für Menschen zu kennzeichnen. Für Informationstexte gelten spezielle Regelungen.

- **Code of Conducts von Medien für den Umgang mit KI**

Medienunternehmen haben sich in den letzten Jahren in Form von KI-Leitlinien oder Redaktionsstatuten Regeln zum transparenten Umgang mit KI auferlegt. Diese sehen oft vor, KI-Content dem User gegenüber auszuweisen oder auch gar kein KI generiertes Material zu veröffentlichen.

76

Auszüge aus KI-Richtlinien öst. Medien: Standard

- *Fotojournalistische Bilder sind frei von inhaltsverändernden Eingriffen. Wenn wir über **KI-generierte Fake-Fotos** berichten, die den Eindruck eines fotojournalistischen Bildes erzeugen sollen, kennzeichnen wir diese. Illustrationen, die wir mit KI-Tools erstellen, kennzeichnen wir ebenfalls. ...Und wir drängen vertraglich darauf, dass uns Bildagenturen, deren Fotos und Illustrationen wir übernehmen, keine ungekennzeichneten KI-generierten Bilder und Illustrationen zur Verfügung stellen.*

<https://about.derstandard.at/redaktion/redaktionelle-ki-richtlinien-der-standard/> (abgerufen am 22.06.2026)

77

Auszüge aus KI-Richtlinien öst. Medien: ORF

▪ ERLAUBTE ANWENDUNGSFÄLLE

Verwendung von KI-generierten Bildern in der Berichterstattung zum Thema KI In der Berichterstattung zu Themen rund um künstliche Intelligenz können KI-generierte Bilder und Illustrationen eingesetzt werden.

- *Das KI-generierte Element muss durch einen schriftlichen Hinweis „KI-generiert“ eindeutig gekennzeichnet sein.*
- *Diese Kennzeichnung und, sofern möglich, eine Quellenangabe müssen unmittelbar beim entsprechenden Objekt platziert werden.*

▪ NICHT ERLAUBTE ANWENDUNGSFÄLLE

Verwendungen fotorealistischer Bilder als Ersatz für Fotojournalismus

<https://der.orf.at/unternehmen/leitbild-werte/ki-guidelines/ki-guidelines104.pdf> (abgerufen am 22.06.2026)

78

EU AI Act: Deepfake - ein umstrittener Begriff

- KI-erzeugte oder -bearbeitete Bild-, Audio- und Videoinhalte, die reale Personen, Orte oder Ereignisse **täuschend echt** darstellen
- Weiter gefasst als in der Öffentlichkeit oft diskutiert: auch KI-Produktbilder, fingierte Pressefotos, synthetische Stock-Bilder
- Beurteilung aus Publikumssicht: **keine Täuschungsabsicht nötig**
- Technische Standardbearbeitung (Normalisierung, Rauschunterdrückung, Kompression) erzeugen in der Regel kein Deepfake-Bild



→ Deepfakes müssen klar als KI-erzeugt oder bearbeitet gekennzeichnet werden – verständlich für Menschen, ohne Spezialtool.

79

Forschungsinputs



Austrian Lab for AI Trust* Branchen Report 2

Voraussetzungen für den vertrauenswürdigen Einsatz von Künstlicher Intelligenz in Nachrichtenmedien

Branchen-Report mit konkreten Handlungsvorschlägen für Führungskräfte, KI- und Strategie-Verantwortliche sowie Journalistinnen

Executive Summary

Künstliche Intelligenz (KI) verändert fundamental, wie Nachrichtenmedien arbeiten und wie ihre Produkte in der Gesellschaft empfangen und wahrgenommen werden. Ihr Einsatz beeinflusst den journalistischen Prozess der Erarbeitung von Text-, Bild- und Video-Inhalten in Zeitungs-, Zeitschrift- und Rundfunk-Unternehmen ebenso wie deren Verbreitung in verschiedenen Formaten über digitale Plattformen und Kanäle. KI-Werkzeuge dirigieren und filtern Nachrichtenflüsse. Damit entscheiden sie mit, welche Inhalte Konsumentinnen wo und wann zu sehen und zu hören bekommen.



Journalist Fellowship Paper Autonomie und Autorenschaft: KI-Transparenz im hybriden Journalismus

Katharina Schell

Dezember 2024
Michaelus Term
Sponsor: APA - Austria Presse Agentur

80

Circle of Trust: Transparenz im Ökosystem

Der **APA** als nationale Nachrichtenagentur kommt als Verarbeiter und Distributor von täglich tausenden Text-, Bild- und Videomeldungen eine besondere Sorgfalt im Umgang mit Inhalten zu.

Im Hinblick auf KI hat die APA deshalb eine eigene Leitlinie veröffentlicht. Diese sieht vor, dass **in den redaktionellen Kanälen der APA kein KI erzeugtes oder wesentlich bearbeitetes Material** veröffentlicht wird.

Wo wird die APA künftig die Distribution von gekennzeichnetem, KI erzeugtem oder KI bearbeitetem Material unterstützen?

- APA-Science Partnermeldungen
- APA-OTS

81

Maßnahmen aus dem APA-Universum

- KI-Erklärung für redaktionelle Partner im Bereich KI-Bild/Video
- Technische Unterstützung für PR bei Versand von KI-Bild/Videomaterial über APA-OTS
- Kennzeichnung bei der Interaktion mit KI im Produkt Pressespiegel
- Addendum zur KI-Leitlinie zum Thema KI-Transparenz für interne Arbeit
- Anpassung der AGBs
- KI-Kompetenzschulung zum Thema Kennzeichnung

82



APA NEWS

Vielen Dank!

Verena Krawarik

Leitung APA-
Innovationsmanagement und
APA-medialab

+43 1 36060 3020
verena.krawarik@apa.at

LinkedIn:
<https://www.linkedin.com/in/verena-krawarik-b405b3156/>

83

Fazit: Wie können Organisationen mehr Sicherheit im Umgang mit KI-Systemen erlangen?



Systemisches Verständnis von KI-Risiken

Technische, organisatorische, rechtliche, gesellschaftliche und menschliche Aspekte gemeinsam betrachten



Offene Diskussionskultur über KI-Risiken etablieren

menschliche Kontrolle, kritisches Denken, moralische Integrität und Verantwortungsbewusstsein



KI-Risiko-Kompetenz aufbauen

Menschen / Mitarbeitende befähigen, die Grenzen von KI zu erkennen; AI Risk Literacy fördern

Foto: unsplash

84

84



**Vertrauen durch Wissen:
Gemeinsam KI verstehen
und gestalten**

Forschungs- und Entwicklungsprojekt ALAIT - Austrian Lab for Artificial Intelligence Trust

Das F&E-Projekt ALAIT entwickelt ein innovatives ALAIT Technology Impact Assessment für Künstliche Intelligenz und fördert durch das ALAIT Labor und ein Train-the-Trainer-Netzwerk den öffentlichen Diskurs über KI-Vertrauen. Ergebnisse werden in AI-Trust-Dossiers veröffentlicht und breit gestreut. Unser Ziel ist es, Vertrauen in KI-Technologie durch Transparenz und Information zu stärken.

Projektlaufzeit: September 2024 – Jänner 2027

Die Projektergebnisse – ALAIT-Trust-Dossiers & Branchen-Reports – stehen zum Download bereit.

Besuchen Sie gerne unsere Projekt-Website!

6 Dossiers und 2 Branchen Reports
stehen zum Download bereit:

<https://science.apa.at/project/alait/>



Weitere Materialien in Vorbereitung

85

85

Zum Nachlesen



AILEX - Fachzeitschrift für KI-Recht
Ausgabe erscheint in den nächsten Wochen:

Ex-ante Risikobewertung von KI-Systemen

Unterstützung von Entscheidungs- und Gestaltungsprozessen in Organisationen

Gertraud Leimüller, Brigitte Ömer-Rieder, Sabine Köszegi

Austrian Lab for AI Trust (ALAIT)

86

86

Hinweis



<https://www.imagine-ikt.at/>



87

87



leiwand.ai



Herzlichen Dank !

Wir freuen uns über Ihr Feedback und Ihre Anfragen:

gertraud.leimueller@leiwand.ai

brigitte.oemer-rieder@winnovation.at

Austrian Lab for AI Trust (ALAIT)

88

88