LIMINAR ACADEMY · MBA · M3-LV2

Liminar

Trustworthy AI

Vertrauenswürdige KI zwischen Recht, Technik und Ethik — mit Fallstudie „Claude vs. Mistral“

Perspektive: öffentlicher Sektor · APA 7 · Juli 2026

Fünf Themen — ein roter Faden

Leitfrage: echte architektonische Vertrauenswürdigkeit vs. bloß symbolische Compliance?

1	**Begriffliche Abgrenzungen**	Trustworthiness vs. Trust
2	**Rechtsrahmen**	HLEG → EU AI Act · Fairness-Privacy
3	**Fairness-Toolkits**	konkurrierende Definitionen · SHAP
4	**Ethik in AI**	Prinzipien · Konflikte · Ethics Washing
5	**Explainable AI**	SHAP/LIME · Grenzen · mechanistische Wende

MERKSATZ *Auf jeder Ebene droht die Geste statt der Substanz — vom „grünen Dashboard“ bis zur unfaithful Erklärung.*

Begriffliche Abgrenzungen

Trustworthiness

Eigenschaft des Systems: objektiv, prüfbar, technisch belegbar.

Trust

Haltung des Menschen: subjektiv, psychologisch, kontextabhängig (Trustor × Trustee × Kontext).

Ethical AI

engste, philosophisch-normative Teilmenge

Trustworthy AI

EU-Rahmen: rechtmäßig + ethisch + robust

Responsible AI

organisatorische Umsetzungsebene

VERNETZUNG Akzeptanz (Lernergebnis 4) ist primär eine Trust-Frage — ein einwandfreies System kann an Akzeptanz scheitern.

Rechtsrahmen: der risikobasierte EU AI Act

Von freiwilligen HLEG-Leitlinien (2019) zum verbindlichen EU AI Act (in Kraft seit 08/2024).

Unacceptable	verboten — z. B. behördliches Social Scoring (seit 02/2025)
High Risk	Verwaltung, Beschäftigung — Conformity Assessment + CE (Art. 9,10,13,14,27)
Limited Risk	Transparenzpflichten — Kennzeichnung (Art. 50)
Minimal Risk	keine besonderen Pflichten — Großteil der Anwendungen

VERNETZUNG GPAI eigener Strang (Art. 53, 55). Zeitleiste: GPAI 08/2025 · Kern/Hochrisiko 08/2026 · voll 08/2027.

Das Fairness-Datenschutz-Paradox

AI Act · Art. 10

verlangt repräsentative, bias-geprüfte Daten — kann Verarbeitung geschützter Merkmale erfordern.

DSGVO · Art. 9

verbietet Verarbeitung besonderer Kategorien (Herkunft, Gesundheit …) grundsätzlich.

Auflösung — Art. 10 Abs. 5 AI Act

Sensible Daten eng umzäunt — nur zur Bias-Korrektur, mit Zugriffskontrolle, Zweckbindung, Löschpflicht. DSFA (Art. 35 DSGVO) + FRIA (Art. 27 AI Act) verzahnen; PETs (Pseudonymisierung, synthetische Daten, Differential Privacy).

MERKSATZ *„Fairness through unawareness“ versagt — Proxys rekonstruieren die Merkmale; Diskriminierung wird nur unsichtbar.*

Es gibt nicht eine Fairness

Demographic Parity

gleiche Rate positiver Ergebnisse über alle Gruppen

Equalized Odds

gleiche Treffer- und Fehlerraten über alle Gruppen

Calibration

ein Score bedeutet für jede Gruppe dasselbe

Unmöglichkeitsresultat (COMPAS-Debatte)

Kalibrierung und gleiche Fehlerraten sind bei ungleichen Basisraten nicht zugleich erreichbar.

MERKSATZ *Fairness ist eine normative & politische Wahl — vor der Implementierung getroffen, begründet, dokumentiert.*

Wo man eingreift — und womit

Pre-Processing	an den Daten — Reweighting, Resampling (modellunabhängig)
In-Processing	am Modell — adversarial debiasing (braucht Trainingszugriff)
Post-Processing	am Ergebnis — Schwellen je Gruppe (auch bei Black-Box)

Souveränität

Black-Box erlaubt oft nur Post-Processing. Open-Weight/Self-Hosting eröffnet alle drei Ebenen.

Toolkits: AI Fairness 360 (IBM) · Fairlearn (MS) · What-If Tool (Google) · Aequitas (Public Sector)

Sie messen & korrigieren — die normative Wahl nehmen sie nicht ab. „Grünes Dashboard“ ≠ echte Fairness.

Ethik: konvergierende Prinzipien, die kollidieren

- Transparenz
- Gerechtigkeit
- Nicht-Schaden
- Verantwortung
- Privatsphäre

Meta-Analyse von 84 Leitlinien (Jobin et al., 2019); + „Explicability“ (Floridi & Cowls, 2019).

- **Transparenz ↔ Privatsphäre**
- **Genauigkeit ↔ Erklärbarkeit**
- **Autonomie ↔ Benefizienz**

MERKSATZ *Ethik ist Abwägung unter Zielkonflikten — kein Abhaken einer Checkliste.*

Ethics Washing — wenn Ethik zur Tarnung wird

„From What to How“ (Mittelstadt, 2019; Morley et al., 2020)

Unzählige Prinzipien-Listen (what), kaum erprobte Methoden zur Übersetzung in Praxis (how). Ethik wird erst real in überprüfbaren Prozessen, benannten Rollen mit Eingriffsmacht und echten Konsequenzen.

Vier Prüffragen für jedes behördliche KI-System

- Wird die Zielgröße als „neutral“ dargestellt, obwohl sie verteilungspolitisch ist?
- Ist die menschliche Aufsicht real handlungsfähig — oder durch Automation Bias ausgehöhlt?
- Können Betroffene die Einstufung verstehen und anfechten?
- Wurde die normative Entscheidung offen & demokratisch legitimiert?

Der AMS-Algorithmus (Österreich)

Ab ~2018 stufte ein KI-System Arbeitssuchende nach Wiedereingliederungschance ein. Datenschutzbehörde untersagte 2020; BVwG hob teilweise auf.

Fairness

Alter, Geschlecht, Gesundheit → reproduzierte Nachteile; kalibriert, aber diskriminierungsübertragend

Autonomie vs. Benefizienz

Score formal beratend (Art. 14) — drohte durch Automation Bias zur Fiktion zu werden

Transparenz

Variablengewichtung intransparent; Einstufung kaum nachvollziehbar/anfechtbar

Ethics Washing

verteilungspolitische Wahl als technisch-neutrale Effizienzlogik gerahmt

VERNETZUNG Illustrativ; Falldetails vor formaler Verwendung verifizieren. Vgl. Chouldechova (2017); Mittelstadt (2019).

Explainable AI: die Black Box öffnen

Methoden

- SHAP — spieltheoretisch, lokal & global
- LIME — lokale Approximation
- Counterfactuals — „was hätte sich ändern müssen?“

Grundachsen

intrinsisch ↔ post-hoc · lokal (eine Entscheidung) ↔ global (Modellverhalten). Treiber: Genauigkeit ↔ Erklärbarkeit (Art. 13 AI Act).

Technisch

keine Kausalität; instabil bei korrelierten Merkmalen

Faithfulness

korrelative Zusammenfassung statt treuer Abbildung — über-interpretiert

Human-centered

Qualität hängt von Adressatengerechtheit ab, nicht nur Mathematik

VERNETZUNG Drei fundamentale Grenzen (Salih et al., 2025; UST, 2026) — SHAP anhängen ≠ echte Erklärbarkeit.

Die mechanistische Wende

Bisher: Feature-Attribution

„Welche Eingabe-Merkmale korrelieren mit der Ausgabe?“ — modell-agnostisch, black-box (SHAP/LIME).

Neu: Schaltkreis-Kartierung

„Welche internen Komponenten treiben das Verhalten wirklich?“ — modell-spezifisch, white-box.

MERKSATZ *White-box setzt Zugriff auf Modellinternals voraus — bei Hyperscaler-Modellen fehlt er. Erklärbarkeit = Kontrolle über Infrastruktur.*

VERNETZUNG Noch keine schlüsselfertige Governance-Schicht (Luo & Specia, 2024; UST, 2026). → Modellwahl M3-LV1.

Claude vs. Mistral an den 7 HLEG-Anforderungen

HLEG-Anforderung	Claude (Anthropic)	Mistral AI
1 Menschl. Aufsicht	erfüllt — agentic-safety-Tests	teilw. — Self-Hosting, weniger Doku
2 Robustheit & Sicherheit	erfüllt — RSP/ASL, CBRN-Evals	unklar — keine öffentl. Audits
3 Privatsphäre & Governance	teilw. — CLOUD Act trotz EU-Hosting	erfüllt — EU-nativ, 30-Tage-Löschung
4 Transparenz	erfüllt — System Cards, Hub	teilw. — Open Weights, wenig Prozess-Doku
5 Nicht-Diskriminierung	teilw./erf. — Bias-Evals	unklar — keine Fairness-Audits
6 Gesellschaftl./ökolog. Wohl	teilw. — keine CO_2-Audits	teilw. — Edge-Effizienz, EU-Souveränität
7 Rechenschaftspflicht	erfüllt — RSP, Risk-Reports	teilw. — EU-AI-Act-Pflichten

MERKSATZ *Claude führt bei Safety & Transparenz — Mistral bei EU-Jurisdiktion & Datensouveränität. CLOUD Act greift bei Claude.*

Vier Leitsätze für vertrauenswürdige KI

01 Fairness-Definition zuerst wählen

normativ wählen, begründen, dokumentieren — Leitungs-/Politikentscheidung, nicht Entwickler-Default.

02 Fairness-Privacy sauber lösen

Bias-Messung nur über Art. 10 Abs. 5 (abgeschottet); DSFA + FRIA verzahnen, PETs einsetzen.

03 Erklärbarkeit ehrlich betreiben

SHAP/LIME als Näherung deklarieren; adressatengerechte Begründung; white-box für echte Auditierbarkeit.

04 Mehrdimensional beschaffen

entlang der 7 HLEG-Anforderungen prüfen; CLOUD-Act-Risiko explizit bewerten — kein „einer Sieger“.

VERNETZUNG Von der Vertrauenswürdigkeit zur Managemententscheidung → M4-LV1, M4-LV3. Fairness/Recht → M1-LV2, M3-LV1.

SYNTHESE

Wie die fünf Themen ineinandergreifen

Begriffe

schaffen Klarheit

Rechtsrahmen

macht verbindlich

Fairness-Toolkits

liefern Instrumente

Explainable AI

liefert Instrumente

Ethik

liefert Begründung & Wachsamkeit

Auf jeder Ebene droht symbolische Compliance — echte Vertrauenswürdigkeit verlangt Substanz, nicht die Geste.

KERNBOTSCHAFT

Vertrauenswürdigkeit ist kein Häkchen — sondern der nachweisbare, unbequemere Weg.