LIMINAR ACADEMY · MBA

Liminar

Verwaltung und Politik handlungsfähig machen

Modul 3 · LV2

Trustworthy AI

Vollständige Modulzusammenfassung mit ausgefüllten Cases, Verständnisfragen und Fallstudie „Claude vs. Mistral (7 HLEG-Anforderungen)“

Perspektive: öffentlicher Sektor / Finanzverwaltung · APA 7 · Stand Juli 2026

Inhaltsverzeichnis

ORIENTIERUNG

Modulübersicht und Lernergebnisse

M3-LV2 „Trustworthy AI“ (4 ECTS, asynchron) macht Vertrauenswürdigkeit von einer Absichtserklärung zu einer **nachweisbaren Eigenschaft**. Die fünf Lehrinhalte — Begriffe, Rechtsrahmen, Fairness-Toolkits, Ethik und Explainable AI — werden hier mit ausgefüllten Cases, Praxisbeispielen und Verständnisfragen vertieft und im **Anwenderbezug öffentliche Verwaltung / Finanzverwaltung** verankert. Baut auf M3-LV1 (Technikethik) auf.

Lernergebnisse (aus der Modulbeschreibung)

1. Die aktuelle Entwicklung zur Standardisierung vertrauenswürdiger KI-Lösungen vergleichen.
2. Prinzipien der Vertrauenswürdigkeit bestimmen und im Marketingbereich ausarbeiten.
3. Ethische Aspekte der Nutzung und des Einsatzes von KI klassifizieren.
4. Aspekte zur Steigerung der Akzeptanz durch vertrauenswürdige KI beurteilen.
5. Praktische Instrumentarien zur Steigerung der Vertrauenswürdigkeit unterscheiden.

Kapitelstruktur

Kap.	Thema	Kernfrage
1	Begriffliche Abgrenzungen	Trustworthiness vs. Trust — was heißt „vertrauenswürdig“?
2	Rechtsrahmen	Wie wird Vertrauenswürdigkeit verbindlich (EU AI Act)?
3	Fairness-Toolkits	Welche Fairness — und wie misst/korrigiert man sie?
4	Ethik in AI	Wo endet Recht, wo beginnt Abwägung?
5	Explainable AI	Wie öffnet man die Black Box — und wo sind Grenzen?

ROTER FADEN „Echte architektonische Vertrauenswürdigkeit vs. bloß symbolische Compliance.“ Auf jeder Ebene droht die Geste statt der Substanz — vom „grünen Dashboard“ über Ethics Washing bis zu unfaithful Erklärungen.

VERNETZUNG Vertieft die Erklärbarkeits-/Governance-Fragen aus M3-LV1 (Technikethik) und M1-LV2 (Risiko-Taxonomie); mündet in M4 (Strategisches Management / KI-Strategien).

KAPITEL 1

Begriffliche Abgrenzungen

Jede Bewertung beginnt mit begrifflicher Schärfe. **Trustworthiness** (Vertrauenswürdigkeit) ist eine objektive, prüfbare Eigenschaft des Systems; **Trust** (Vertrauen) ist die subjektive Haltung des Menschen. Ein System kann vertrauenswürdig sein, ohne dass ihm vertraut wird — und umgekehrt. Trust wird über drei Dimensionen modelliert: Trustor, Trustee und Kontext (Bach et al., 2024).

1.1 Drei Reichweiten: Trustworthy / Responsible / Ethical AI

Begriff	Ebene	Kern
Trustworthy AI	EU-Rahmen (HLEG)	rechtmäßig + ethisch + robust (7 Anforderungen)
Responsible AI	Organisation	Governance-Umsetzung, Rollen, Prozesse
Ethical AI	Philosophie	engste, normative Teilmenge

Es gilt: Ethical AI ⊂ Trustworthy AI, während Responsible AI die organisatorische Umsetzungsebene bildet (European Commission, 2019).

1.2 Anwenderbezug: Akzeptanz ist eine Trust-Frage

Lernergebnis 4 (Akzeptanz) ist primär eine **Trust**-Frage, nicht nur eine Trustworthiness-Frage: Ein technisch einwandfreies System kann an mangelnder Akzeptanz scheitern — und blindes Übervertrauen ist ebenso ein Risiko. Begriffsschärfe ist die Voraussetzung, um symbolische von echter Vertrauenswürdigkeit zu unterscheiden.

AUSGEFÜLLTER CASE **Chatbot im Bürgerservice — vertrauenswürdig, aber nicht akzeptiert**

Situation: Ein technisch robuster, AI-Act-konformer Auskunfts-Chatbot wird eingeführt; die Nutzung bleibt niedrig.

Trustworthiness: hoch: geprüft, dokumentiert, robust — objektiv belegbar.

Trust: niedrig: Bürger:innen misstrauen der „anonymen Maschine“, fürchten Falschauskunft.

Diagnose: Die Lücke ist kein Technik-, sondern ein Trust-Problem (Trustor × Kontext).

Maßnahme: Transparente Kennzeichnung (Art. 50), sichtbarer menschlicher Eskalationspfad, Kommunikation der Prüfungen → Akzeptanz folgt der wahrgenommenen Kontrolle.

VERSTÄNDNISFRAGEN **Kapitel 1**

1. Worin unterscheiden sich Trustworthiness und Trust — und warum kann Akzeptanz trotz hoher Vertrauenswürdigkeit ausbleiben?
2. Ordnen Sie Trustworthy, Responsible und Ethical AI als Mengen zueinander.
3. Warum ist blindes Übervertrauen ebenso problematisch wie Misstrauen?

KAPITEL 2

Rechtsrahmen Trustworthy AI

Der Rechtsrahmen macht Vertrauenswürdigkeit von einer freiwilligen Selbstverpflichtung zu einer **nachweispflichtigen Konformität**. Die HLEG-Leitlinien (2019) waren soft law und bereiteten den Boden für den EU AI Act (in Kraft seit August 2024) — das weltweit erste umfassende KI-Recht (Europäische Kommission, 2024).

2.1 Der risikobasierte Ansatz

Risikoklasse	Regel
Unacceptable	verbotene Praktiken (z. B. behördliches Social Scoring); Verbote seit 2. Februar 2025 wirksam.
High Risk	streng reguliert (Verwaltung, Beschäftigung, kritische Infrastruktur); Conformity Assessment + CE (Art. 9, 10, 13, 14, 27).
Limited Risk	Transparenzpflichten (Kennzeichnung von Chatbots & KI-Inhalten, Art. 50).
Minimal Risk	keine besonderen Pflichten (Großteil der Anwendungen).

Für Foundation Models (GPAI) gilt ein eigener Pflichtenstrang (Art. 53, 55: Trainingsdaten-Transparenz, Dokumentation, bei systemischem Risiko zusätzliche Evaluierungen). Zeitleiste: GPAI-Pflichten seit 08/2025, Kernpflichten/Hochrisiko und Sanktionen ab 08/2026, volle Anwendung 08/2027.

2.2 Das Fairness-Datenschutz-Paradox (vertieft)

AI Act und DSGVO gelten parallel — mit einer zentralen Reibung: **Art. 10 AI Act** verlangt bias-geprüfte, repräsentative Daten, was die Verarbeitung geschützter Merkmale erfordern kann, um Diskriminierung überhaupt zu messen. **Art. 9 DSGVO** verbietet die Verarbeitung solcher besonderen Kategorien grundsätzlich. Die Scheinlösung „Fairness through unawareness“ (Merkmale weglassen) versagt, weil Proxy-Variablen sie rekonstruieren — Diskriminierung wird nur unsichtbar, nicht beseitigt.

AUFLÖSUNG Art. 10 Abs. 5 AI Act erlaubt sensible Daten eng umzäunt — nur zur Bias-Erkennung/-Korrektur, mit Zugriffskontrolle, Zweckbindung und Löschpflicht. Praktisch: Datenschutz-Folgenabschätzung (Art. 35 DSGVO) mit Grundrechte-Folgenabschätzung (FRIA, Art. 27 AI Act) verzahnen und Privacy-Enhancing-Technologies nutzen (Pseudonymisierung, synthetische Daten, Differential Privacy).

AUSGEFÜLLTER CASE **Bias-Messung im Prüf-Scoring der Finanzverwaltung**

Ziel: Prüfen, ob ein Risikoscoring bestimmte Gruppen systematisch benachteiligt.

Konflikt: Für die Messung braucht man geschützte Merkmale (Art. 10 AI Act) — die Art. 9 DSGVO grundsätzlich sperrt.

Weglassen hilft nicht: Wohnadresse/Branche als Proxy rekonstruieren Herkunft → Diskriminierung wird unprüfbar.

Rechtssicherer Weg: Art. 10 Abs. 5: sensible Daten nur im abgeschotteten Fairness-Labor, streng

zugriffsbeschränkt, zweckgebunden, mit Löschpflicht; DSFA + FRIA gekoppelt.

Ergebnis: Bias wird messbar UND rechtskonform — Voraussetzung für eine begründbare Korrektur.

VERSTÄNDNISFRAGEN **Kapitel 2**

1. Welche vier Risikoklassen kennt der AI Act — und in welche fällt ein behördliches Prüf-Scoring?
2. Erklären Sie das Fairness-Datenschutz-Paradox und warum „Fairness through unawareness“ versagt.
3. Wie verzahnt man DSFA (Art. 35 DSGVO) und FRIA (Art. 27 AI Act) praktisch?

KAPITEL 3

Toolkits für Fairness in AI

Die zentrale Einsicht: Es gibt nicht **eine** Fairness. Mehrere mathematische Definitionen klingen alle vernünftig, schließen sich bei unterschiedlichen Basisraten aber nachweislich gegenseitig aus (Kleinberg et al., 2016; Chouldechova, 2017).

Definition	Bedeutung
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: Kalibrierung und gleiche Fehlerraten sind bei ungleichen Basisraten nicht zugleich erreichbar (COMPAS-Debatte). Fairness ist damit eine **normative und politische Wahl**, die vor der Implementierung getroffen, begründet und dokumentiert werden muss — keine Entwickler-Entscheidung.

3.1 Wo man eingreift — und womit

Ansatzpunkt	Wirkweise	Souveränität
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 möglich

Souveränitätsaspekt: Zugekaufte Black-Box-Modelle erlauben oft nur Post-Processing; Open-Weight/Self-Hosting eröffnet alle drei Ebenen. Etablierte Toolkits: AI Fairness 360 (IBM), Fairlearn (Microsoft/Community), What-If Tool (Google), Aequitas (öffentlicher Sektor). Sie messen und korrigieren — die normative Wahl nehmen sie nicht ab. Ein „grünes Dashboard“ ohne bewusste Definitionswahl ist symbolische Compliance.

AUSGEFÜLLTER CASE **Welche Fairness für ein Förder-Priorisierungsmodell?**

Situation: Ein Modell priorisiert Fördermittel; Gruppen haben unterschiedliche Basisraten.

Zielkonflikt: Kalibrierung (gleicher Score = gleiche Bedeutung) UND gleiche Fehlerraten sind zugleich unmöglich.

Politische Wahl: Will man gleiche Chancen (Equalized Odds) oder gleiche Score-Bedeutung (Calibration)? → Verteilungsentscheidung.

Prozess: Definition vor der Implementierung wählen, begründen, dokumentieren; Aequitas zur Messung, Post-Processing bei Black-Box.

Anwenderbezug: Die Wahl trifft die Leitung/Politik — nicht die Entwicklung; sie ist demokratisch zu legitimieren.

VERSTÄNDNISFRAGEN **Kapitel 3**

1. Warum schließen sich Calibration und Equalized Odds bei ungleichen Basisraten aus?
2. Wann ist nur Post-Processing möglich — und was folgt daraus für die Modellbeschaffung?
3. Warum ist ein „grünes Fairness-Dashboard“ noch kein Nachweis echter Fairness?

KAPITEL 4

Ethik in AI

Ethik geht über Recht hinaus: Recht hinkt der Technik hinterher, definiert nur ein Minimum und ist territorial. Eine Meta-Analyse von 84 Leitlinien fand fünf konvergierende Prinzipien — **Transparenz, Gerechtigkeit, Nicht-Schaden, Verantwortung, Privatsphäre** (Jobin et al., 2019). Ein verbreiteter Rahmen ergänzt die vier bioethischen Prinzipien (Autonomie, Benefizienz, Non-Malefizienz, Gerechtigkeit) um „Explicability“ (Floridi & Cowls, 2019).

4.1 Prinzipien kollidieren — Ethik ist Abwägung

Entscheidend: Diese Prinzipien kollidieren in der Praxis — Transparenz ↔ Privatsphäre, Genauigkeit ↔ Erklärbarkeit, Autonomie ↔ Benefizienz. Ethik ist Abwägung unter Zielkonflikten, kein Abhaken einer Checkliste. **Ethics Washing** (Mittelstadt, 2019): „Ethik“ selbst wird zur Tarnung, wenn freiwillige Leitlinien verbindliche Regeln ersetzen. Das „From What to How“-Problem (Morley et al., 2020) zeigt die Lücke zwischen Prinzipien und Praxis.

VIER PRÜFFRAGEN FÜR JEDES BEHÖRDLICHE KI-SYSTEM 1) Wird die Zielgröße als „neutral“ dargestellt, obwohl sie verteilungspolitisch ist? 2) Ist die menschliche Aufsicht real handlungsfähig — oder durch Automation Bias ausgehöhlt? 3) Können Betroffene die Einstufung verstehen und anfechten? 4) Wurde die normative Entscheidung offen und demokratisch legitimiert?

4.2 Praxisbeispiel & Fallstudie: der AMS-Algorithmus (Österreich)

Das österreichische Arbeitsmarktservice entwickelte ab ca. 2018 ein KI-System, das Arbeitssuchende nach prognostizierter Wiedereingliederungschance in drei Gruppen einstufte. Die Datenschutzbehörde untersagte 2020 den Einsatz in der geplanten Form; das Bundesverwaltungsgericht hob dies teilweise auf. Der Fall illustriert nahezu jede Spannung des Moduls:

AUSGEFÜLLTER CASE **AMS-Algorithmus an den vier Prüffragen**

- **Fairness (Kap. 2–3):** Merkmale wie Alter, Geschlecht, Gesundheit reproduzierten bestehende Arbeitsmarktnachteile — statistisch kalibriert, aber diskriminierungsübertragend (vgl. COMPAS; Chouldechova, 2017).
- **Autonomie vs. Benefizienz (Kap. 4):** Score formal nur beratend (human-in-the-loop, Art. 14) — drohte durch Automation Bias zur Fiktion zu werden.
- **Transparenz (Kap. 4–5):** Variablengewichtung zunächst intransparent; Betroffene konnten die Einstufung kaum nachvollziehen oder anfechten.
- **Ethics Washing:** Die verteilungspolitische Wahl („Förderung dorthin, wo der Hebel groß ist“) wurde als technisch-neutrale Effizienzlogik gerahmt statt offen legitimiert (Mittelstadt, 2019).
- **Lehre für die Finanzverwaltung:** Vor jedem Scoring die vier Prüffragen beantworten; die Zielgröße offen als politische Wahl deklarieren.

VERSTÄNDNISFRAGEN **Kapitel 4**

1. Warum kann „Ethik“ selbst zur Tarnung werden (Ethics Washing) — und woran erkennt man den Unterschied?
2. Wenden Sie die vier Prüffragen auf ein KI-Scoring Ihrer Wahl an.
3. Welche Prinzipienkonflikte zeigt der AMS-Fall konkret?

KAPITEL 5

Explainable AI (XAI)

XAI übersetzt die Black Box in eine verständlichere Form, um Transparenz und Nutzervertrauen zu erhöhen (Salih et al., 2025) — die technische Antwort auf Art. 13 AI Act. Treiber ist die Spannung zwischen **Genauigkeit und Erklärbarkeit**. Grundachsen: intrinsisch vs. post-hoc und lokal (eine Entscheidung) vs. global (Modellverhalten).

5.1 Methoden und ihre fundamentalen Grenzen

Etablierte Methoden: **SHAP** (spieltheoretisch, lokal & global), **LIME** (lokale Approximation) und Counterfactuals („was hätte sich ändern müssen?“). Grenzen: technisch (keine Kausalität, instabil bei korrelierten Merkmalen); Faithfulness (oft korrelative Zusammenfassung statt treuer Abbildung — häufig über-interpretiert; UST, 2026); human-centered (Erklärungsqualität hängt von adressatengerechter Kommunikation ab, nicht nur von der Mathematik).

5.2 Die mechanistische Wende & Souveränität

Das Feld bewegt sich von der **Feature-Attribution** (welche Merkmale korrelieren mit der Ausgabe? — modell-agnostisch, black-box) zur **Schaltkreis-Kartierung** (welche internen Komponenten treiben das Verhalten? — modell-spezifisch, white-box; Luo & Specia, 2024). Sie ist noch keine schlüsselfertige Governance-Schicht (UST, 2026). Souveränitätsaspekt: White-box-Verfahren setzen Zugriff auf die Modellinternals voraus — der bei zugekauften Hyperscaler-Modellen fehlt. Erklärbarkeit selbst wird damit zur Frage der Kontrolle über die Infrastruktur.

AUSGEFÜLLTER CASE **Erklärbarer Bescheid — SHAP reicht nicht**

Anforderung: Ein KI-gestützter Bescheid muss nach Art. 13 AI Act / Begründungspflicht nachvollziehbar sein.

Naiver Weg: SHAP-Werte als „Erklärung“ anhängen — wirkt transparent.

Faithfulness-Problem: Bei korrelierten Merkmalen ist die Attribution instabil und nur korrelativ — keine treue Abbildung der internen Berechnung.

Human-centered: Bürger:innen brauchen eine adressatengerechte Begründung, keine Feature-Tabelle.

Souveräner Weg: Für echte Auditierbarkeit white-box-Zugriff (Open-Weight/Self-Hosting) + interpretierbares Modell + juristische Begründung durch den Menschen.

VERSTÄNDNISFRAGEN **Kapitel 5**

1. Was bedeutet „Faithfulness“ und warum werden SHAP/LIME häufig über-interpretiert?
2. Worin unterscheidet sich mechanistische Interpretierbarkeit von SHAP/LIME?
3. Warum ist Erklärbarkeit auch eine Frage der Infrastruktur-Souveränität?

FALLSTUDIE

Fallstudie: Claude vs. Mistral an den 7 HLEG-Anforderungen

Die sieben HLEG-Anforderungen (European Commission, 2019) bilden das operative Prüfraster für Trustworthy AI. Der Vergleich zweier Systeme zeigt, dass Vertrauenswürdigkeit mehrdimensional ist — kein Anbieter „gewinnt“ auf allen Achsen (Bewertung teils wertend; Stand Juli 2026).

HLEG-Anforderung	Anthropic — Claude	Mistral AI
1 Menschl. Handeln & Aufsicht	erfüllt: dokumentierte agentic-safety-/Prompt-Injection-Tests	teilweise: Self-Hosting stärkt Kontrolle; Aufsicht weniger dokumentiert
2 Techn. Robustheit & Sicherheit	erfüllt: RSP/ASL-Level, CBRN-/Cyber-Evaluierungen	unklar: keine vergleichbaren öffentl. Sicherheitsaudits belegt
3 Privatsphäre & Daten-Governance	teilweise: US-Jurisdiktion, CLOUD Act greift trotz EU-Hosting	erfüllt: EU-nativ, EU-Hosting Default, 30-Tage-Löschung
4 Transparenz	erfüllt: öffentl. System Cards & Transparency Hub	teilweise: Open Weights inspizierbar, wenig Prozess-Doku
5 Vielfalt & Nicht-Diskriminierung	teilw./erf.: explizite Bias-Evaluierungen	unklar: keine veröffentl. systematischen Fairness-Audits
6 Gesellschaftl. & ökolog. Wohl	teilweise: keine robusten Energie-/CO_2-Audits	teilweise: effiziente Edge-Modelle; EU-Souveränität
7 Rechenschaftspflicht	erfüllt: RSP, verpflichtende Risk-/Sabotage-Reports	teilweise: direkte EU-AI-Act-Pflichten, klare Haftungskette

KERNBEFUND & ANWENDERBEZUG Claude führt bei dokumentierter Safety & Transparenz — Mistral bei EU-Jurisdiktion & Datensouveränität. Für DSGVO-sensible Public-Sector-Verarbeitung ist Mistrals EU-Domizil ein struktureller (nicht symbolischer) Vorteil; der CLOUD Act greift bei Claude trotz EU-Hosting. → deckt sich mit der Modellwahl-Doktrin aus M3-LV1 (Verantwortbarkeit vor Fähigkeit).

SYNTHESE

Synthese, Handlungsempfehlungen & Prüfung

Wie die fünf Themen ineinandergreifen

Thema	Rolle	Symbolische Falle
Begriffe	schaffen Klarheit	Buzzword-Gebrauch ohne Trennschärfe
Rechtsrahmen	macht verbindlich	Häkchen-Compliance ohne Substanz
Fairness-Toolkits	liefern Instrumente	„grünes Dashboard“ ohne Definitionswahl
Ethik	liefert Begründung & Wachsamkeit	Ethics Washing
Explainable AI	liefert Instrumente	unfaithful Erklärungen

Vier Handlungsempfehlungen

6. **Fairness-Definition zuerst wählen:** Vor der Implementierung die Fairness-Definition normativ wählen, begründen und dokumentieren — als Leitungs-/Politikentscheidung, nicht als Entwickler-Default.
7. **Fairness-Privacy sauber lösen:** Bias-Messung nur über Art. 10 Abs. 5 AI Act (abgeschottet, zweckgebunden, Löschpflicht); DSFA + FRIA verzahnen, PETs einsetzen.
8. **Erklärbarkeit ehrlich betreiben:** SHAP/LIME als Näherung deklarieren (Faithfulness-Grenzen), adressatengerechte Begründung durch den Menschen; für echte Auditierbarkeit white-box-Zugriff sichern.
9. **Vertrauenswürdigkeit mehrdimensional beschaffen:** Entlang der 7 HLEG-Anforderungen prüfen (Souveränität vs. Safety/Transparenz) statt „einen Sieger“ zu küren; CLOUD-Act-Risiko explizit bewerten.

> **VERNETZUNG** Von der Vertrauenswürdigkeit zur Managemententscheidung → M4-LV1 (Strategisches Management), M4-LV3 (KI-Strategien & Management); Fairness/Recht knüpfen an M1-LV2 und M3-LV1 an.

10 Prüfungsfragen zum Selbsttest

1. Erklären Sie den Unterschied zwischen Trustworthiness und Trust.
2. Wie ordnen sich Trustworthy, Responsible und Ethical AI zueinander?
3. Nennen Sie die vier Risikoklassen des AI Act mit je einem Beispiel.
4. Was besagt das Fairness-Datenschutz-Paradox — und wie löst Art. 10 Abs. 5 AI Act es?
5. Warum gibt es nicht „eine“ Fairness (Unmöglichkeitsresultat)?
6. Unterscheiden Sie Pre-, In- und Post-Processing — und den Souveränitätsbezug.
7. Was ist Ethics Washing und welche vier Prüffragen entlarven es?
8. Welche drei Spannungen illustriert der AMS-Algorithmus?
9. Was bedeutet Faithfulness in XAI — und warum ist mechanistische Interpretierbarkeit white-box?
10. Wer führt bei welchen HLEG-Anforderungen: Claude oder Mistral — und warum?

ÜBUNG

Essay-Übung — Modul 3

Wählen Sie **eine** der drei Forschungsfragen und verfassen Sie einen wissenschaftlichen Essay (Richtwert 2.000–3.000 Wörter, APA 7), argumentiert mit Blick auf den öffentlichen Sektor. Verzahnt mit Modul 2 (Academic Research Skills).

Forschungsfrage 1 · Fairness-Unmöglichkeit & demokratische Wahl
„Welche Fairness-Definition soll ein behördliches Risikoscoring wählen — und wie lässt sich diese normative Wahl angesichts des Unmöglichkeitsresultats demokratisch legitimieren und dokumentieren?“

Abstract: Ausgehend vom Unmöglichkeitsresultat (Kleinberg et al., 2016; Chouldechova, 2017) untersucht der Essay, wie die Wahl zwischen Calibration und Equalized Odds als verteilungspolitische Entscheidung sichtbar gemacht wird. Am AMS-Fall wird ein Legitimations- und Dokumentationsverfahren (Wahlkriterien, Rollen, FRIA) entwickelt. Erwarteter Beitrag: eine Heuristik, die die Fairness-Wahl aus der Entwicklungs- in die demokratische Sphäre zurückholt.

Forschungsfrage 2 · Fairness-Datenschutz-Paradox in der Praxis
„Wie lässt sich das Fairness-Datenschutz-Paradox (Art. 10 AI Act vs. Art. 9 DSGVO) in einem konkreten Verwaltungsverfahren rechtssicher und zugleich wirksam auflösen?“

Abstract: Der Essay analysiert Art. 10 Abs. 5 AI Act als eng umzäunte Ausnahme und entwickelt ein Referenzdesign, das DSFA (Art. 35 DSGVO) und FRIA (Art. 27 AI Act) verzahnt und Privacy-Enhancing-Technologies (Pseudonymisierung, synthetische Daten, Differential Privacy) einsetzt. Methodisch wird ein Prozessmodell für ein abgeschottetes „Fairness-Labor“ hergeleitet. Erwarteter Beitrag: ein rechtssicheres, wirksames Verfahren zur Bias-Messung in der Verwaltung.

Forschungsfrage 3 · Erklärbarkeit, Souveränität & Modellwahl
„Inwiefern hängt echte Erklärbarkeit (bis zur mechanistischen Interpretierbarkeit) von der Kontrolle über die Modellinfrastruktur ab — und was folgt daraus für die Modellwahl im öffentlichen Sektor (Claude vs. Mistral)?“

Abstract: Auf Basis der Faithfulness-Kritik (Salih et al., 2025; UST, 2026) und der white-box-Voraussetzung mechanistischer Interpretierbarkeit (Luo & Specia, 2024) prüft der Essay die These, dass Erklärbarkeit eine Frage der Infrastruktur-Souveränität ist. Der 7-HLEG-Vergleich (Claude vs. Mistral) dient als Anwendung. Erwarteter Beitrag: eine gestaffelte Beschaffungsempfehlung, die Erklärbarkeit, Souveränität und CLOUD-Act-Risiko gewichtet.

AUFBAU DES ESSAYS (APA 7) 1) Einleitung & Fragestellung · 2) Theorie / Stand der Forschung · 3) Analyse & Argumentation · 4) Fazit & Praxisbezug · 5) Literaturverzeichnis (APA 7).

Literaturverzeichnis (APA 7)

Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2024). A systematic literature review of user trust in AI-enabled systems. Frontiers in Psychology, 15, 1382693.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.

Europäische Kommission. (2024). Verordnung (EU) 2024/1689 über künstliche Intelligenz (AI Act). Amtsblatt der Europäischen Union.

European Commission, High-Level Expert Group on AI. (2019). Ethics guidelines for trustworthy AI. Publications Office of the European Union.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. Harvard Data Science Review, 1(1).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores [Preprint]. arXiv:1609.05807.

Luo, H., & Specia, L. (2024). From understanding to utilization: A survey on explainability for large language models [Preprint]. arXiv:2402.10688.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 26(4), 2141–2168.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. Advanced Intelligent Systems, 7(1), 2400304.

UST. (2026). From explainability to control: The 2026 executive view of AI interpretability and explainability. ust.com.

Anthropic. (2025–2026). System Cards & Transparency Hub. anthropic.com.

Mistral AI. (2025–2026). Model & privacy documentation. mistral.ai.