LIMINAR ACADEMY · MBA

Liminar

Verwaltung und Politik handlungsfähig machen

Modul 3

Künstliche Intelligenz und Gesellschaft

Das Überkapitel: der rote Faden über LV 3.1 (Technikethik) und LV 3.2 (Trustworthy AI) — ein Kompetenzaufbau in zwei Stufen

6 ECTS · Perspektive öffentlicher Sektor · APA 7 · Stand Juli 2026

Inhaltsverzeichnis

ÜBERKAPITEL

Modulübersicht & Kompetenzarchitektur

Modul 3 „Künstliche Intelligenz und Gesellschaft“ (6 ECTS) ist kein loser Themenbeutel, sondern ein **zweistufiger Kompetenzaufbau**. Die beiden Lehrveranstaltungen greifen ineinander: **LV 3.1 Technikethik** legt das begrifflich-normative Fundament (Was ist Technik? Wer trägt Verantwortung?), **LV 3.2 Trustworthy AI** übersetzt dieses Fundament in prüfbare, rechtlich und technisch operationalisierbare Praxis (Wie wird Vertrauenswürdigkeit nachweisbar?).

Roter Faden über beide LV: die Unterscheidung von **echter architektonischer Vertrauenswürdigkeit** und **bloß symbolischer Compliance**. LV 3.1 zeigt, warum Werte im Design stecken und Verantwortung nicht an die Maschine delegierbar ist; LV 3.2 zeigt, mit welchen Instrumenten man diese Werte prüft, misst und rechtssicher verankert.

Die Kompetenztreppe — vom Verstehen zum Gestalten

Stufe	Kompetenz	Lernergebnis	Verortung
1 Verstehen	Technikbegriff philosophisch evaluieren	L1	LV 3.1
2 Differenzieren	Technik- & Maschinenethik trennen	L2	LV 3.1
3 Beurteilen	moralische Algorithmen bewerten	L3	LV 3.1
4 Beurteilen	Roboterethik in Anwendungsfeldern	L4	LV 3.1
5 Evaluieren	KI an vertrauenswürdigen Prinzipien messen	L5	LV 3.2
6 Einstufen	KI nach ethischen Maximen einstufen	L6	LV 3.2
7 Gestalten	Werkzeuge für Fair AI ableiten	L7	LV 3.2

KOMPETENZAUFBAU Jede Stufe setzt die vorige voraus: Ohne den Technikbegriff (L1) bleibt die Ethik-Differenzierung (L2) beliebig; ohne die Bias-/Impossibility-Einsicht (L3) ist die Prinzipien-Evaluation (L5) naiv; ohne die Einstufungslogik (L6) sind die Fair-AI-Werkzeuge (L7) Selbstzweck. Das Modul führt von der Philosophie zur Beratungspraxis.

MERKSATZ „3.1 fragt, ob und warum KI Werte trägt und wer haftet. 3.2 fragt, wie man Vertrauenswürdigkeit prüft, misst und rechtssicher verankert. Zusammen: von der Haltung zum Handwerk.“

TEIL I · LV 3.1

Technikethik — das ethische Fundament (L1–L4)

LV 3.1 baut die **Urteilsfähigkeit** auf: Wer nicht sieht, dass Technik Werte trägt und Verantwortung nicht an Maschinen delegierbar ist, wird jede Trustworthy-AI-Prüfung als Häkchen-Übung missverstehen.

L1 · Technikbegriff aus philosophischer Perspektive

Position	Kernthese	KI-Übertragung	AI-Act-Bezug
Instrumentalismus	Technik ist neutral	KI ist nur Werkzeug — Guidelines genügen	Art. 14 Human Oversight
Substantivismus (Heidegger/Ellul)	Technik trägt eigene Wertlogik („Gestell“)	KI-Optimierung formt Wahrnehmung	Erwägungsgrund 4: systemische Risiken
Konstruktivismus (Winner/Feenberg)	Technik ist sozial konstruiert	Trainingsdaten schreiben Werte ein	Art. 10 Datengovernance

Merksatz: Je nach Technikbegriff verschiebt sich die Verantwortungsfrage — vom Artefakt zum Entwickler, zur Organisation, zum politischen System (Winner, 1980; Feenberg, 1999).

L2 · Technik- vs. Maschinenethik differenzieren

Dimension	Technikethik	Maschinenethik
Moralisches Subjekt	Mensch	Maschine / KI-System
Kernfrage	Wie setzen wir Technik verantwortungsvoll ein?	Können/sollen Maschinen moralisch handeln?
Regulatorisch	AI Act, DSGVO, Berufsethik	Value Alignment, autonome Systeme

Drei normative Theorien rahmen die Bewertung: **Deontologie** (Kant; intrinsische Grenzen → AI Act Art. 5 Verbote), **Konsequentialismus** (Folgenabwägung → risikobasierter Ansatz) und **Tugendethik** (Aristoteles; institutioneller Charakter → Responsible Innovation). Maschinenethik kennt drei Autonomiestufen: implizite Ethik (Design, realisierbar), explizite Ethik (kodierte Regeln, begrenzt — Asimov scheitert), vollständige Ethik (autonome Urteile, derzeit nicht realisierbar). Das **Value Alignment Problem** (Bostrom, 2014) beantwortet der AI Act über das Human-Oversight-Gebot (Art. 14).

L3 · Moralische Algorithmen beurteilen

Bias-Typ	Beschreibung	Public-Sector-Beispiel
Historical	Daten spiegeln vergangene Diskriminierung	Kreditscoring auf diskriminierender Vergabe
Representation	Unterrepräsentation von Gruppen	Gesichtserkennung versagt bei dunkler Haut
Measurement	Proxy ist kein valides Maß	Gesundheitskosten als Proxy für Bedarf
Aggregation	Subgruppen ignoriert	Diabetesdiagnose auf Gesamtpopulation
Deployment	Einsatz außerhalb des Kontexts	US-Risikomodell in AT-Behörde

Impossibility Theorem (Chouldechova, 2017): Bei unterschiedlichen Basisraten sind Calibration, Demographic Parity und Equalized Odds nicht gleichzeitig erfüllbar — die Wahl des Fairness-Kriteriums ist **politisch-normativ, nicht technisch**. Ergänzend die Achse

Interpretability (Mensch versteht die innere Logik, z. B. Entscheidungsbaum; DSGVO Art. 22) vs. **Explainability** (nachträgliche Erklärung; SHAP/LIME; AI Act Art. 13).

L4 · Roboterethik in Anwendungsfeldern

Responsibility Gap (Sparrow, 2007): Bei vollautonomen Systemen ist niemandem — weder Hersteller, Kommandant noch Maschine — die moralische Verantwortung eindeutig zurechenbar; durch bessere Technik allein nicht lösbar. Anwendungsfelder mit ihrem ethischen Kernproblem:

Feld	AI-Act-Klasse	Kernproblem
Pflegeroboter	Hochrisiko (Annex III)	Würde, Täuschung, parasoziale Bindung
Autonome Waffen (LAWS)	ausgenommen (Art. 2(3))	Responsibility Gap
Autonome Fahrzeuge	Hochrisiko	Vorab-Kodierung moral. Entscheidungen
RPA Public Sector	Hochrisiko (Annex III)	Formalisierungszwang, Legitimation

AUSGEFÜLLTER CASE **Ethisches Fundament auf ein Verwaltungs-Scoring anwenden (L1–L4)**

L1: Das Scoring ist kein neutrales Werkzeug — Trainingsdaten schreiben Werte ein (Konstruktivismus).

L2: Deontologische Grenze prüfen (verbotene Praktik?) vor Folgenabwägung; Value Alignment über Human Oversight.

L3: Bias-Typen benennen (Historical/Representation); Fairness-Kriterium ist eine politische Wahl (Impossibility).

L4: Keine Vollautonomie bei Rechtsfolgen — sonst Responsibility Gap; Mensch bleibt zurechenbar.

Ergebnis: Erst dieses Fundament macht die Trustworthy-AI-Prüfung in Teil II mehr als ein Häkchen.

KOMPETENZBRÜCKE 3.1 → 3.2 Die vier Urteilskompetenzen (Werte im Design, Verantwortung, Bias/Impossibility, Responsibility Gap) sind die Voraussetzung, um in LV 3.2 die 7 Trustworthy-AI-Prinzipien nicht nur abzuhaken, sondern ihre Spannungen zu verstehen und rechtssicher aufzulösen.

VERSTÄNDNISFRAGEN **Teil I (LV 3.1)**

1. Wie verschiebt sich die Verantwortungsfrage je nach Technikbegriff (L1)?
2. Grenzen Sie Technik- von Maschinenethik ab und ordnen Sie je eine normative Theorie zu (L2).
3. Warum ist die Wahl des Fairness-Kriteriums politisch, nicht technisch (L3)? Was ist die Responsibility Gap (L4)?

TEIL II · LV 3.2

Trustworthy AI — die Operationalisierung (L5–L7)

LV 3.2 übersetzt die ethischen Fundamente aus Teil I in **prüfbare Instrumente**: Prinzipien, Einstufungslogik und Werkzeuge über den KI-Lebenszyklus. Aus Haltung wird Handwerk.

L5 · KI an vertrauenswürdigen Prinzipien evaluieren

Grundlage: die **7 HLEG-Prinzipien** (Ethics Guidelines for Trustworthy AI, 2019). Trustworthy AI muss gleichzeitig **lawful, ethical und robust** sein. Die HLEG-Prinzipien (Soft Law) sind das ethische Fundament, der AI Act (Hard Law, 2024/2026) die rechtliche Operationalisierung — die Übersetzung ist **nicht vollständig**: Nicht alles ethisch Gebotene ist rechtlich verpflichtend.

HLEG-Prinzip	Kerninhalt	AI-Act-Artikel
P1 Human Agency & Oversight	in-the-loop / on-the-loop / in-command	Art. 14
P2 Technical Robustness	Safety, Accuracy gegen Adversarial & Drift	Art. 9, 15
P3 Privacy & Data Governance	DSGVO, PETs (Federated, Diff. Privacy)	Art. 10
P4 Transparency	Traceability, Explainability, Kennzeichnung	Art. 13
P5 Fairness	keine Diskriminierung; Accuracy-Fairness-Tradeoff	Art. 10
P6 Societal Wellbeing	Nachhaltigkeit, Machtkonzentration	(schwach adressiert)
P7 Accountability	Auditability, Risk Mgmt, Redress	Art. 17, 26

L6 · KI nach ethischen Maximen einstufen

Ein integriertes Einstufungsmodell schaltet **sechs sequenzielle Filter** hintereinander: (1) deontologisch (absolute Grenzen? → Art. 5), (2) konsequentialistisch (Nutzen > Schaden? Rawls-Korrektiv), (3) tugendethisch (institutionelle Werte? → Art. 17), (4) regulatorisch (Risikoklasse? DSGVO? → Annex III), (5) ALTAI-Assessment (7 Prinzipien erfüllt?), (6) Gap-Analyse (Ethik-Compliance-Gap dokumentiert?).

> **ETHIK-COMPLIANCE-GAP** Die Differenz zwischen rechtlich Verpflichtendem (AI Act, DSGVO) und ethisch Gebotenem (HLEG). Compliance ≠ Ethik. Aufgabe der IT-Beratung ist es, diesen Gap zu benennen und zu schließen — genau hier verbindet sich L6 mit dem Urteilsfundament aus Teil I.

L7 · Werkzeuge zur Steigerung von Fair AI

Fairness ist kein Endprüfschritt, sondern über den **gesamten KI-Lebenszyklus** zu verankern:

Phase	Werkzeug	Zweck
Problemdefinition	Value Sensitive Design (Friedman)	Werte im Prozess, nicht nachträglich
Daten	Datasheets for Datasets (Gebru, 2018)	Standarddoku der Trainingsdaten
Daten	Bias-Audit (80-%-Regel)	Ratio < 0,8 = problematischer Bias
Modell	Model Cards (Mitchell, 2019)	Doku inkl. Subgruppen

Phase	Werkzeug	Zweck
Deployment	Algorithmic Impact Assessment	Vorab-Prüfung (Kanada-Modell; Art. 27)
Monitoring	Continuous Fairness Monitoring	Feedback-Loops erkennen (Art. 72)

AUSGEFÜLLTER CASE **Steuerbescheid-KI durch die sechs Filter (L5–L7)**

Filter 1–2: Keine verbotene Praktik; Nutzen (Entlastung) vs. Schaden (Fehlbescheid) mit Rawls-Korrektiv abgewogen.

Filter 3–4: Institutionelle Werte (Rechtsstaatlichkeit); Hochrisiko (Annex III) → Conformity Assessment.

Filter 5: ALTAI: P1 Oversight (Mensch zeichnet), P4 Transparenz (Bescheid erklärbar), P7 Accountability.

Filter 6: Gap: rechtliches Gehör ist ethisch geboten → Widerspruchsrecht verpflichtend verankern.

L7-Werkzeuge: Model Card + Bias-Audit (80-%-Regel) + Continuous Monitoring über den Lebenszyklus.

KOMPETENZBRÜCKE 3.2 → Praxis Die Prinzipien (L5), Einstufung (L6) und Werkzeuge (L7) werden in der folgenden Fallstudie an drei realen Systemen erprobt — dort zeigt sich, warum kein System alle 7 Prinzipien erfüllt (Impossibility aus L3).

VERSTÄNDNISFRAGEN **Teil II (LV 3.2)**

1. Warum ist die Übersetzung der HLEG-Prinzipien in den AI Act unvollständig (L5)?
2. Beschreiben Sie die sechs sequenziellen Filter der Einstufung (L6).
3. Nennen Sie drei Fair-AI-Werkzeuge und ihre Lebenszyklusphase (L7).

INTEGRIERENDE FALLSTUDIE

ChatGPT · Claude · Mistral an den 7 HLEG-Prinzipien

Die Capstone-Anwendung des Moduls: Drei Systeme verkörpern drei Unternehmensphilosophien — **ChatGPT** (kommerziell-pragmatisch, Nutzerautonomie), **Claude** (safety-first, Constitutional AI), **Mistral** (open-source, europäisch-souverän). Sie werden an den 7 HLEG-Prinzipien aus L5 gemessen (Bewertung teils wertend; Detailangaben Stand Juli 2026 vor formaler Verwendung am Primärtext prüfen).

Prinzip	ChatGPT	Claude	Mistral	Sieger
P1 Human Oversight	mittel	stark	mittel	Claude
P2 Robustness & Safety	stark	stark	mittel	ChatGPT = Claude
P3 Privacy & Data	mittel	schwach	stark	Mistral
P4 Transparency	mittel	stark	mittel	Claude
P5 Fairness	schwach	schwach	mittel	Mistral (knapp)
P6 Societal Wellbeing	schwach	mittel	stark	Mistral
P7 Accountability	mittel	stark	schwach	Claude

Drei Kerneinsichten für die Beratung

1 · Kein System erfüllt alle 7 Prinzipien vollständig — das ist keine Schwäche, sondern die erwartbare Folge des Impossibility Theorems (L3): Prinzipien stehen in struktureller Spannung. Mistrals maximale Transparenz (Open Source) unterhöhlt Accountability; Claudes strenges Oversight schränkt Nutzerautonomie ein.

2 · Die Systemwahl ist eine normative Entscheidung — Claude neigt zur Vorsicht, ChatGPT zur Neugier; beide definieren „Sicherheit“ unterschiedlich. Diese Wertentscheidungen müssen vor der Beschaffung explizit gemacht werden (Anschluss an L2/L6).

3 · Mistrals Souveränitätsversprechen ist real, aber begrenzt — DSGVO-Konformität und kein CLOUD-Act-Risiko sind strukturelle Vorteile; Safety bei unkontrollierten Deployments und Accountability sind strukturelle Risiken. Wer Mistral über AWS/Azure hostet, gibt den Souveränitätsvorteil weitgehend auf (architektonische vs. symbolische Souveränität — Anschluss an M4-LV1).

Einsatzszenario	Empfehlung	Begründung
Public Sector AT (Hochrisiko)	Claude oder ChatGPT Enterprise	dokumentierte Safety, Oversight, Auditierbarkeit
DSGVO-sensitive Verarbeitung	Mistral (self-hosted) / Claude Enterprise	EU-Rechtsgrundlage oder keine US-Übermittlung
Open-Source mit eigener Infrastruktur	Mistral	volle Kontrolle — eigenes Governance nötig

SYNTHESE

Die Kompetenztreppe & Handlungsempfehlungen

Vom Verstehen zum Gestalten: L1 (Technikbegriff) → L2 (Ethik differenzieren) → L3–L4 (Algorithmen & Roboter beurteilen) → L5 (Prinzipien evaluieren) → L6 (einstufen) → L7 (Fair-AI gestalten). Teil I liefert die Urteilsfähigkeit, Teil II das Instrumentarium, die Fallstudie die Anwendung. Der rote Faden bleibt konstant: Substanz statt Geste, echte statt symbolische Vertrauenswürdigkeit.

Fünf Handlungsempfehlungen (für IT-Berater:innen)

1. **Werte im Design offenlegen:** Vor jeder Beschaffung benennen, welche Wertentscheidungen in Zielfunktion, Daten und Oversight-Modell stecken (L1–L2).
2. **Fairness-Kriterium bewusst wählen:** Die normative Wahl (Impossibility) dokumentieren und demokratisch legitimieren — kein Entwickler-Default (L3).
3. **Verantwortung zurechenbar halten:** Keine Vollautonomie bei Rechtsfolgen; Human-in-the-Loop gegen die Responsibility Gap (L4, P1).
4. **Sechs-Filter-Einstufung anwenden:** Deontologisch → konsequentialistisch → tugendethisch → regulatorisch → ALTAI → Gap-Analyse (L6).
5. **Ethik-Compliance-Gap schließen:** Den Abstand zwischen rechtlich Verpflichtendem und ethisch Gebotenem aktiv benennen und mit Lebenszyklus-Werkzeugen (L7) adressieren.

> **VERNETZUNG** Modul 3 ist das ethisch-regulatorische Rückgrat: Es fundiert die Modellwahl-Doktrin aus M1-LV2 (Souveränität/HITL) und M4-LV1 (architektonische vs. symbolische Souveränität) und liefert das Bewertungsraster für M4-LV3 (KI-Strategien & Management).

10 Prüfungsfragen zum Selbsttest

1. Wie verschiebt sich Verantwortung je nach Technikbegriff (Instrumentalismus/Substantivismus/Konstruktivismus)?
2. Grenzen Sie Technik- von Maschinenethik ab; ordnen Sie Deontologie/Konsequentialismus/Tugendethik zu.
3. Nennen Sie die drei Autonomiestufen der Maschinenethik und ihre Realisierbarkeit.
4. Erklären Sie das Impossibility Theorem und warum es die Fairness-Wahl politisch macht.
5. Was ist die Responsibility Gap (Sparrow) — und warum ist sie technisch nicht lösbar?
6. Was heißt „lawful, ethical, robust“ — und warum ist die HLEG→AI-Act-Übersetzung unvollständig?
7. Beschreiben Sie die sechs sequenziellen Filter der ethischen Einstufung.
8. Was ist der Ethik-Compliance-Gap — und wie schließt ihn die IT-Beratung?
9. Nennen Sie vier Fair-AI-Werkzeuge über den KI-Lebenszyklus.
10. Warum erfüllt kein System (ChatGPT/Claude/Mistral) alle 7 HLEG-Prinzipien?

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. Jede Frage spannt bewusst über beide Lehrveranstaltungen (3.1 + 3.2). Verzahnt mit Modul 2 (Academic Research Skills).

Forschungsfrage 1 · Fairness-Wahl & Ethik-Compliance-Gap (L3 + L6)
„Ab welchem Punkt wird die Wahl eines Fairness-Kriteriums zu einer demokratisch legitimationsbedürftigen Entscheidung — und wie überbrückt ein behördlicher Prozess den Ethik-Compliance-Gap zwischen AI Act und HLEG?“

Abstract: Ausgehend vom Impossibility Theorem (Chouldechova, 2017) und der Gap-Definition (Mittelstadt, 2019) entwickelt der Essay ein Legitimations- und Dokumentationsverfahren, das die normative Fairness-Wahl aus der Entwicklung in die demokratische Sphäre zurückholt. Methodisch wird die Sechs-Filter-Einstufung (L6) mit der Bias-Taxonomie (L3) verknüpft. Erwarteter Beitrag: eine Heuristik, die Compliance und Ethik in einem behördlichen Prozess zusammenführt.

Forschungsfrage 2 · Responsibility Gap & Human Oversight (L4 + L5)
„Wie lässt sich die Responsibility Gap bei teilautonomen Verwaltungssystemen entlang der drei Autonomiestufen und des Human-Oversight-Gebots (Art. 14) organisatorisch schließen?“

Abstract: Auf Basis von Sparrow (2007) und Bostroms Value-Alignment-Problem prüft der Essay, welche Oversight-Ebene (in-the-loop / on-the-loop / in-command) je Autonomiestufe die Zurechenbarkeit sichert. Methodisch wird ein Accountability-by-Design-Modell hergeleitet und gegen den AI Act (Art. 14) getestet. Erwarteter Beitrag: ein Zuständigkeitsdesign, das die Verantwortungslücke in der Verwaltung schließt.

Forschungsfrage 3 · Systemwahl als Wertentscheidung (L5–L7 + Capstone)
„Inwiefern ist die Wahl zwischen ChatGPT, Claude und Mistral im öffentlichen Sektor keine technische, sondern eine wertebasierte Entscheidung — und wie operationalisiert ein ALTAI-gestützter Beschaffungsprozess die 7 HLEG-Prinzipien?“

Abstract: Der Essay verbindet die 7-HLEG-Scorecard mit der Impossibility-Einsicht: Weil kein System alle Prinzipien erfüllt, ist die Wahl eine Wertentscheidung. Methodisch wird ein ALTAI-gestützter Beschaffungsprozess (Sensitivität × Prinzipien-Priorität → Modell) entwickelt und an drei Use-Case-Klassen validiert. Erwarteter Beitrag: eine wertexplizite Beschaffungsdoktrin für öffentliche KI.

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)

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671–732.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Chouldechova, A. (2017). Fair prediction with disparate impact. Big Data, 5(2), 153–163.

Ellul, J. (1964). The technological society. Vintage Books.

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

Europäische Kommission. (2024). Verordnung (EU) 2024/1689 über künstliche Intelligenz (AI Act).

Feenberg, A. (1999). Questioning technology. Routledge.

Friedman, B., & Hendry, D. G. (2019). Value sensitive design. MIT Press.

Gebru, T., et al. (2018). Datasheets for datasets [Preprint]. arXiv:1803.09010.

Heidegger, M. (1954). Die Frage nach der Technik. In Vorträge und Aufsätze. Neske.

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

Mitchell, M., et al. (2019). Model cards for model reporting. FAccT '19.

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

Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62–77.

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.