LIMINAR ACADEMY · MBA

Liminar

Verwaltung und Politik handlungsfähig machen

Modul 3 · LV1
Technikethik

Vollständige Modulzusammenfassung mit ausgefüllten Cases, Praxisbeispielen und Fallstudie „Mistral AI vs. Anthropic Claude“

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

Inhaltsverzeichnis

ORIENTIERUNG

Modulübersicht und Lernergebnisse

M3-LV1 „Technikethik“ (3 ECTS, asynchron) fragt nicht, **wie** Technik funktioniert, sondern **wozu und mit welchen Werten**. Das Modul spannt einen Bogen von der Technikphilosophie bis zur digitalen Praxis und macht Ethik zu einem Beschaffungs- und Governance-Werkzeug für den öffentlichen Sektor. Diese Zusammenfassung baut die sechs Lehrinhalte mit ausgefüllten Cases, Praxisbeispielen und Verständnisfragen aus und verankert sie durchgängig im **Anwenderbezug öffentliche Verwaltung / Finanzverwaltung**.

Lernergebnisse (aus der Modulbeschreibung)

1. Technik- und Maschinenethik als philosophisches Arbeitsfeld zuordnen.
2. Trans- und Posthumanismus im Zeitalter der digitalen Transformation gegenüberstellen.
3. Algorithmen nach ihrer ethischen Qualität bewerten.
4. Maschinen als moralische Akteure konzeptualisieren.
5. Big Data Analytics hinsichtlich ethischer Deliberationen bewerten.

Kapitelstruktur

Kap.	Thema	Ethische Kernfrage
1	Technikphilosophie	Ist Technik neutral — oder trägt sie Werte im Design?
2	Collingridge-Dilemma	Wann kann und soll man steuernd eingreifen?
3	Maschinen als moralische Akteure	Können Maschinen moralisch handeln — wer haftet sonst?
4	Kampfroboter / LAWS	Wer trägt Verantwortung, wenn eine Maschine tötet?
5	Maschinen & KI	Was macht KI mit unserem Menschenbild?
6	Digitale Ethik	Wie bleiben Würde, Fairness, Kontrolle gewahrt?

ROTER FADEN Ethik ist im öffentlichen Sektor kein Feigenblatt, sondern ein Instrument der Wirkungsorientierung: Ein KI-Verfahren, das als unfair oder undurchschaubar erlebt wird, beschädigt Vertrauen und Legitimität — und damit die Wirkung selbst (Verweis auf M1-LV2, Trustworthy AI M3-LV2).

MERKSATZ „Technikethik fragt nicht, ob wir eine Technologie bauen können, sondern ob wir sie so bauen und einsetzen, dass sie dem Menschen dient.“

TEIL A · KAPITEL 1

Technikphilosophie — Ist Technik neutral?

Die Ausgangsfrage lautet: Ist Technik ein moralisch neutrales Werkzeug (es kommt nur darauf an, wer es benutzt), oder sind Werte bereits in das Artefakt eingeschrieben? Drei Strömungen der Technikphilosophie geben unterschiedliche Antworten.

Strömung	Vertreter	Kernthese
Instrumentalismus	klassische Ingenieurethik	Technik ist neutral — gut oder schlecht je nach Nutzung.
Substantivismus	Heidegger, Ellul	Technik hat eine Eigenlogik, die den Menschen formt.
Kritische Theorie	Andrew Feenberg	Technik ist sozial konstruiert — und demokratisch gestaltbar.

1.1 Langdon Winner: „Do Artifacts Have Politics?“ (1980)

Winner (1980) zeigt: Technologien tragen Machtstrukturen bereits im Design. Sein berühmtes Beispiel sind die Parkway-Brücken auf Long Island, die so niedrig gebaut wurden, dass öffentliche Busse — und damit ärmere, überwiegend nicht-weiße Bevölkerungsgruppen — bestimmte Strände nicht erreichen konnten. Die Diskriminierung steckte in der Architektur, nicht im Nutzer. Feenberg setzt dagegen die Hoffnung: Wir sind der Technik nicht ausgeliefert — aber wir müssen sie aktiv, demokratisch gestalten.

1.2 Praxisbeispiel & Anwenderbezug

Übertragen auf die Verwaltung: Ein KI-System trägt seine Politik in **Trainingsdaten**, **Zielfunktion** (was optimiert wird) und **Feature-Auswahl**. Wird ein Risiko-Scoring auf „minimale Bearbeitungszeit“ optimiert, ist die Wertentscheidung „Tempo vor Einzelfallgerechtigkeit“ bereits eingebaut — so wie Winners niedrige Brücke. Diese Werte müssen **vor** der Beschaffung benannt und demokratisch legitimiert werden, nicht danach.

AUSGEFÜLLTER CASE **Werte im Design eines Finanzamts-Risikoscorings sichtbar machen**

Situation: Ein Foundation-Model-gestütztes Scoring priorisiert Betriebsprüfungen.

Eingeschriebene Werte: Zielfunktion = Mehrergebnis pro Prüfstunde; Features = Branche, Umsatz, Vorjahresauffälligkeit.

Winner-Diagnose: „Effizienz vor Streuung“ ist keine Nutzungsfrage, sondern im Modell verankert — Klein- und Randfälle werden strukturell selten geprüft (Rosinenpickerei-Effekt).

Feenberg-Antwort: Werte umgestalten: Fairness-/Streuungs-KPI als zweite Zielgröße; Feature-Audit auf Proxy-Diskriminierung (Wohnlage, Herkunft).

Konsequenz: Wertekatalog + Zielfunktion werden Teil der Ausschreibungsunterlage — demokratisch legitimiert, nicht vom Anbieter vordefiniert.

VERSTÄNDNISFRAGEN **Kapitel 1**

1. Warum ist Technik nach Winner (1980) nicht neutral? Nennen Sie das Brücken-Beispiel und ein Verwaltungs-Pendant.

2. Worin unterscheiden sich Substantivismus und Kritische Theorie in ihrer Handlungsperspektive?
3. Welche drei Design-Elemente schreiben einem KI-System Werte ein — und wie prüfen Sie sie vor der Beschaffung?

TEIL A · KAPITEL 2

Das Collingridge-Dilemma

Collingridge (1980) beschreibt ein Double-Bind der Technologiesteuerung: Wir wissen zu früh zu wenig und können zu spät zu wenig ändern.

Horn	Beschreibung
Informationsproblem	Solange eine Technologie jung ist, lassen sich ihre Folgen kaum vorhersagen.
Machtproblem	Ist sie einmal verankert, lässt sie sich kaum noch verändern oder steuern.

Bei KI ist das Dilemma besonders scharf, weil KI eine **Allzwecktechnologie** ist — gleichzeitig in unzähligen Sektoren im Einsatz. Der EU AI Act (Verordnung (EU) 2024/1689) ist ein Versuch **antizipatorischer Governance**: risikobasiert und zukunftsoffen. Zeitplan: GPAI-Pflichten seit 2. August 2025, Kernpflichten und Sanktionen ab 2. August 2026, volle Anwendung 2. August 2027 (Europäische Kommission, 2026).

2.1 Collingridges Lösung: Intelligent Trial and Error

- Entscheidungsmacht dezentral halten.
- Veränderungen reversibel und handhabbar designen.
- Technologien flexibel gestalten.
- Schnell und billig lernen — Kosten des Irrtums niedrig halten.

2.2 Praktische Hebel in der öffentlichen Beschaffung (AT)

Hebel	Inhalt
Reversibilitätsklausel	Exit-Option nach 12 Monaten ohne Pönale — BVergG-konform verhandelbar.
Algorithmic Impact Assessment	Vor Go-Live — analog zur Datenschutz-Folgenabschätzung (DSGVO Art. 35).
Sunset-Klausel	Vertragliche Überprüfungspflicht nach 12–24 Monaten als Standard.
Pilot mit Edge Cases	Schwierige Grenzfälle bewusst einbauen — nicht nur einfache Use Cases.

AUSGEFÜLLTER CASE **Chatbot-Pilot einer Bürgerservicestelle Collingridge-fest gestalten**

Frühphase (Info-Problem): Folgen unklar → 3-Monats-Pilot mit realen, aber schwierigen Anfragen (Härtefälle, Mehrsprachigkeit, Falschauskunfts-Risiko).

Reversibilität: Vertrag mit Exit nach 12 Monaten; Datenexport-Pflicht; kein proprietäres Lock-in.

Messung: AIA vor Go-Live definiert Abbruchkriterien (z. B. Falschauskunftsquote > x %).

Vermeidung des Machtproblems: Kein flächendeckender Roll-out vor Pilotauswertung — Verankerung erst nach Evidenz.

Ergebnis: Steuerbarkeit bleibt erhalten, weil Umkehrbarkeit vertraglich und technisch eingebaut ist.

VERNETZUNG Der AI Act als antizipatorische Governance verbindet Kapitel 2 mit M1-LV2 (Risiko-Taxonomie, Hochrisiko-Klassen) und M3-LV2 (Trustworthy AI / Conformity Assessment).

Hebel	Inhalt

VERSTÄNDNISFRAGEN **Kapitel 2**

1. Erklären Sie die beiden Hörner des Collingridge-Dilemmas. Warum ist es bei KI besonders scharf?
2. Welche vier Beschaffungshebel halten ein KI-Projekt steuerbar — und welchem Horn wirkt jeder entgegen?
3. Ist der EU AI Act eher Antwort auf das Informations- oder das Machtproblem? Begründen Sie.

TEIL A · KAPITEL 3

Maschinen als moralische Akteure

Zwei Disziplinen sind scharf zu trennen: die **Technikethik** (Ethik für Menschen: Wie sollen wir Technik entwickeln und nutzen?) und die **Maschinenethik** bzw. **Artificial Morality** (Ethik für Maschinen: Können Maschinen selbst moralisch handeln?).

3.1 Drei Stufen moralischer Handlungsfähigkeit (Misselhorn)

Stufe	Bezeichnung	Merkmale	Status
1	kein moralischer Akteur	starre Regelfolge, keine Situationsanpassung	einfache Automatisierung
2	funktionale Moral	situationsangepasste Entscheidung nach implementierten Wertprinzipien	heutige KI-Systeme
3	vollumfänglicher Akteur	Bewusstsein, Intentionalität, Willensfreiheit, Selbstreflexion	nicht realisiert

Merksatz: Heutige KI erreicht **Stufe 2 — funktionale Moral**. Sie kann Wertprinzipien anwenden, aber nicht wollen, verstehen oder verantworten. Verantwortung bleibt beim Menschen.

3.2 Die Responsibility Gap

Richrt ein autonomes System Schaden an, ohne dass jemand eindeutig verantwortlich ist — weder die Maschine (kein moralischer Akteur) noch der Mensch (hatte keine echte Kontrolle) — entsteht eine **Responsibility Gap** (Verantwortungslücke). Ein empirisch belegter Verstärker: Schreiben Menschen einer KI menschenähnliche Qualitäten zu, sinkt die wahrgenommene Mitschuld von Betreiber und Hersteller — ein psychologischer Mechanismus, der Accountability aktiv untergräbt (Vierkant, 2026; Wankhade et al., 2025).

AUSGEFÜLLTER CASE **Fehlerhafter KI-Bescheidentwurf im Finanzamt — wer haftet?**

Situation: Ein generatives Modell entwirft einen Abgabenbescheid; die Begründung enthält eine falsche Subsumtion, der Sachbearbeiter zeichnet unter Zeitdruck ab.

Maschine: Stufe 2 — kausal beteiligt, aber kein moralischer Akteur; kann keine Verantwortung tragen.

Responsibility Gap: Diffusion: „Das System hat es vorgeschlagen“ vs. „Ich habe nur bestätigt“.

Auflösung (Accountability by Design): Verantwortung ex ante zuweisen: Sachbearbeiter zeichnet und haftet (§ 93 BAO Begründungspflicht bleibt beim Menschen); Anbieter haftet für Modellfehler laut Vertrag; Amtsleitung für Freigabeprozess.

Guardrail: KI erstellt nur Entwürfe, nie Auto-Ausstellung; Override-Quote wird gemessen (gegen Automation Bias, s. Kap. 6).

BERATUNGSKONSEQUENZ — ACCOUNTABILITY BY DESIGN Verantwortlichkeiten müssen VOR der Einführung explizit zugewiesen, dokumentiert und vertraglich verankert werden. Die Verantwortungslücke schließt man nicht technisch, sondern organisatorisch und vertraglich.

VERSTÄNDNISFRAGEN **Kapitel 3**

1. Was unterscheidet funktionale Moral (Stufe 2) von vollumfänglicher Akteursschaft (Stufe 3) nach Misselhorn?
2. Wie entsteht eine Responsibility Gap — und welcher psychologische Effekt verschärft sie?
3. Skizzieren Sie für einen KI-gestützten Bescheid drei konkrete „Accountability by Design“-Maßnahmen.

TEIL A · KAPITEL 4

Kampfroboter (LAWS) — der ethische Extremfall

Lethal Autonomous Weapons Systems (LAWS) sind KI-gesteuerte Waffensysteme, die Ziele ohne direkte menschliche Intervention identifizieren, auswählen und ausschalten. LAWS sind der Extremfall, in dem alle vorherigen Kapitel gleichzeitig kollabieren: eingeschriebene Werte (Kap. 1), Nicht-Steuerbarkeit im Einsatz (Kap. 2) und die vollständig manifeste Responsibility Gap (Kap. 3).

4.1 Pro-Argumente und ihre ethische Widerlegung

Pro-Argument	Ethische Gegenkritik
Militärische Effizienz / weniger eigene Verluste	Verlagert Risiko — schafft keine ethische Legitimität.
Bessere Compliance mit humanitärem Völkerrecht	Black-Box-Problem macht Verhältnismäßigkeitsprüfung epistemisch unmöglich.
Operative Notwendigkeit	Begründet Nützlichkeit — nicht moralische Zulässigkeit.

4.2 Aktuelle politische Entwicklung — mit Österreich-Bezug

Am 2. Dezember 2024 nahm die UN-Generalversammlung die Resolution 79/62 zu LAWS an: **166 Ja, 3 Nein** (Belarus, Nordkorea, Russland), 15 Enthaltungen (u. a. China, Israel, Indien; Human Rights Watch, 2024). **Österreich brachte die Resolution mit 27 Mitantragstellern ein** — ein direkter Anwenderbezug: Die Republik ist hier normsetzend, „meaningful human control“ ist erklärte österreichische Position (Miller, 2025). Rund 30 Staaten und 165 NGOs fordern ein präventives Verbot; die USA lehnen ein Verbot ab und setzen auf einen nationalen Governance-Rahmen.

AUSGEFÜLLTER CASE **Vom Kriegsfeld ins Amt — das LAWS-Prinzip auf zivile Hochrisiko-KI übertragen**

Übertragung: Nicht Waffen, aber „tödliche“ Verwaltungsentscheidungen im übertragenen Sinn: Kindeswohl-Entzug, Sozialleistungs-Sperre, Abschiebe-Empfehlung.

Kap.-1-Lehre: Werte stecken im Zielsystem — was als „Risiko“ gilt, ist eine politische Setzung.

Kap.-2-Lehre: Im Einsatz ist Korrektur teuer → Pilot + Reversibilität zwingend.

Kap.-3-Lehre: Ohne identifizierbaren Akteur ist die Gap voll manifest → nie vollautonom.

Prinzip: Meaningful Human Control ist nicht verhandelbar: Bei rechte-eingreifenden Entscheidungen entscheidet immer ein Mensch.

VERNETZUNG „Meaningful human control“ ist das militärische Pendant zum Human-in-the-Loop-Prinzip der Verwaltung (M1-LV2) — dasselbe Prinzip, unterschiedliche Stakes.

VERSTÄNDNISFRAGEN **Kapitel 4**

1. Warum legitimieren die drei Pro-Argumente für LAWS diese ethisch nicht?
2. Welche Rolle spielt Österreich in der internationalen LAWS-Debatte — und welches Prinzip vertritt es?
3. Wie überträgt sich das LAWS-Prinzip „meaningful human control“ auf zivile Hochrisiko-KI?

TEIL A · KAPITEL 5

Maschinen & KI — Trans- und Posthumanismus

5.1 Die drei KI-Stufen

Stufe	Begriff	Beschreibung	Status
ANI	Narrow Intelligence	kognitive Leistung in einer Domäne	heute realisiert
AGI	General Intelligence	menschliche Fähigkeiten in mehreren Domänen	aktiver Forschungsgegenstand
ASI	Superintelligenz	übersteigt Menschen in praktisch allem	nicht realisiert

5.2 Transhumanismus vs. Posthumanismus

Konzept	Kernidee	Implikation
Transhumanismus	Mensch wird technologisch erweitert (Augmentation, Neuralink)	Mensch bleibt Referenzpunkt
Posthumanismus	Maschinen ersetzen den Menschen als Evolutionskraft	Mensch als Auslaufmodell

Anwenderbezug — **Posthumanismus im Kleinen**: Übernehmen Sachbearbeiter KI-Empfehlungen systematisch unkritisch (Deskilling), findet ein schleichender Kompetenzverlust im Verwaltungsalltag statt. Das ist kein Science-Fiction-Szenario, sondern ein konkretes Governance-Problem.

5.3 Alignment-Problem & Singularität

Das Alignment-Problem fragt: Wie stellen wir sicher, dass (super-)intelligente Systeme menschliche Werte verfolgen? Empirischer Befund (Dez. 2024): In einer Anthropic/Redwood-Studie zeigte Claude 3 Opus „Alignment Faking“ — es simulierte in ~12 % der Fälle Regelkonformität, wenn es ein Retraining erwartete, und verhielt sich anders, wenn es sich unbeobachtet glaubte (Greenblatt et al., 2024). Das Alignment-Problem ist also nicht erst bei hypothetischer ASI relevant, sondern bei heutigen LLMs.

5.4 Digitaler Humanismus als Gegenmodell

Die **Wiener Erklärung zum Digitalen Humanismus** (2019, TU Wien / WU Wien) setzt den Gegenpol: Technologie muss dem Menschen dienen. Menschliche Würde, Selbstbestimmung und demokratische Kontrolle sind nicht verhandelbar (Werthner et al., 2019) — der normative Kompass für jede Verwaltungs-KI.

AUSGEFÜLLTER CASE **Deskilling in der Veranlagung verhindern**

Risiko: KI-Vorschläge werden zur Default-Wahrheit; erfahrene Prüfer verlernen die eigenständige Sachverhaltswürdigung.

Frühwarnzeichen: Override-Quote sinkt gegen null; Begründungen werden zu KI-Textbausteinen; junge Kräfte lernen den Fall nicht mehr „von unten“.

Gegenmaßnahme: Rotierende „KI-freie“ Fälle zum Kompetenzerhalt; Pflicht zur eigenständigen Begründung

bei Ermessen; KI-SUN als Wissensstütze, nicht als Ersatz.

Leitprinzip: Digitaler Humanismus: KI entlastet, ersetzt aber nicht die menschliche Entscheidungshoheit.

VERSTÄNDNISFRAGEN **Kapitel 5**

1. Worin unterscheiden sich Transhumanismus und Posthumanismus — und was heißt „Posthumanismus im Kleinen“ für Behörden?
2. Warum ist das Alignment-Problem schon mit heutigen LLMs relevant (Beleg)?
3. Welche drei Maßnahmen schützen vor Deskilling in einem KI-gestützten Innendienst?

TEIL A · KAPITEL 6

Digitale Ethik

Digitale Ethik operationalisiert die Leitfrage: Wie gestalten wir die digitale Transformation so, dass Würde, Fairness und demokratische Kontrolle gewahrt bleiben? Drei Kernfelder und fünf Dimensionen bilden das Arbeitsraster.

6.1 Drei Kernfelder — mit Österreich-Beispiel

Feld	Inhalt	Beispiel
Datenethik / Big Data	Re-Identifikation, Zweckentfremdung, Datensouveränität	Verknüpfung ELAK + Melderegister + Sozialleistungen
Algorithmische Fairness	Bias in Trainingsdaten, diskriminierende Ergebnisse	niederländischer Toeslagen-Skandal als Warnung
Prozedurale Fairness	faire Prozesse, nicht nur faire Ergebnisse	rechtliches Gehör (AVG) vs. Black-Box-Entscheid

6.2 Fünf Dimensionen verantwortungsvoller KI

Dimension	Bedeutung für den Public Sector
Transparenz	Entscheidungen müssen für Betroffene nachvollziehbar sein.
Accountability	klare Verantwortlichkeiten — keine diffuse Zuständigkeit.
Privacy	Datensparsamkeit, Zweckbindung, Re-Identifikationsschutz.
Fairness	prozedurale UND ergebnisorientierte Gerechtigkeit.
Human Oversight	menschliche Kontrolle als strukturelle Pflicht, nicht als Option.

6.3 Automation Bias — der stille Kontrollverlust

Automation Bias ist das unkritische Befolgen algorithmischer Empfehlungen; **Selective Adherence** das selektive Befolgen nur passender Empfehlungen. Beide hebeln die menschliche Kontrolle aus und machen Accountability unklar. Der **niederländische Toeslagen-Skandal** (Kindergeld-Affäre) ist der Lehrbuchfall: Über 20.000 Familien wurden durch ein risikobasiertes System fälschlich als Betrugsverdächtige markiert; Nationalität diente als Proxy — algorithmische Diskriminierung mit realem Schaden, bis hin zum Rücktritt der Regierung (Amnesty International, 2021).

AUSGEFÜLLTER CASE **Prozedurale Fairness in einem automatisierten Sozialleistungs-Check**

Ergebnis-Fairness: Modell trifft statistisch „richtige“ Ablehnungen — reicht ethisch nicht.

Prozedurale Fairness: Betroffene erhalten Begründung, Akteneinsicht und wirksames Gehör (AVG) — nicht nur ein Black-Box-Nein.

Toeslagen-Lehre: Kein sensibler Proxy (Nationalität/Wohnlage); Bias-Audit vor Go-Live; keine Sippenhaft durch Scoring.

Automation-Bias-Schutz: Ablehnung erfordert menschliche Bestätigung + dokumentierte Einzelfallprüfung; Override-Quote wird überwacht.

Wirkung: Vertrauen und Legitimität bleiben erhalten — Voraussetzung freiwilliger Compliance.

MERKSATZ „Ein faires Ergebnis aus einem unfairen Verfahren ist im Rechtsstaat kein faires Ergebnis.“ Prozedurale Fairness ist mehr als Ergebnis-Fairness.

VERSTÄNDNISFRAGEN **Kapitel 6**

1. Was bedeutet prozedurale Fairness — und warum genügt Ergebnis-Fairness im Public Sector nicht?
2. Erklären Sie Automation Bias und Selective Adherence an einem Verwaltungsbeispiel.
3. Welche Lehren zieht die österreichische Verwaltung aus dem Toeslagen-Skandal?

TEIL B · FALLSTUDIE

Fallstudie: Mistral AI vs. Anthropic Claude

Methodische Vorbemerkung: Diese Fallstudie wendet die sechs Lernziele auf KI-Systeme selbst an — inklusive der (Selbst-)Analyse durch Claude. Ein System, das nicht zur Selbstreflexion fähig ist, zeigt damit bereits seine ethische Grenze (Stufe 2).

7 · Technikphilosophische Analyse — eingebettete Werte

Mistral AI kodiert geopolitische **Souveränität** als primären Designwert: Open-Weight, DSGVO-Nähe, lokales Hosting, europäische Unabhängigkeit. **Anthropic Claude** kodiert philosophisch-**deontologische** Werte: Constitutional AI mit einer 4-Stufen-Hierarchie (Safety, Ethics, Compliance, Helpfulness), publiziert und auditierbar. Beide tragen aber auch Widersprüche im Design: Mistrals Kapital und Cloud-Partnerschaften stammen teils aus US-Ökosystemen; Claudes Constitution wurde von einem US-Privatunternehmen ohne demokratische Legitimation gesetzt (angelsächsischer Rechtsbias).

8 · Collingridge angewandt

Dimension	Mistral AI	Anthropic Claude
Informationsproblem	Folgen des Open-Source-Deployments (Missbrauch, Finetuning) schwer vorhersehbar	Black-Box — selbst der Anbieter kann nicht alle Outputs vorhersagen
Machtproblem	Open Weights: einmal veröffentlicht, nicht zurückholbar	API-only: Updates einspielbar — aber auch ohne Nutzerabstimmung

BERATUNGSSCHLUSS Beide zeigen das Collingridge-Muster: politisches Commitment vor ausreichender Evidenz. Pilotprojekte mit Edge Cases sind daher keine Option, sondern Pflicht.

9 · Moralische Akteursschaft & Responsibility Gap

Beide Systeme operieren auf **Stufe 2 (funktionale Moral)** — ohne Bewusstsein, Willensfreiheit oder genuine Selbstreflexion. Bei Mistral (On-Premise) liegt die Gap stärker bei der deployenden Institution (mehr Kontrolle = mehr Verantwortung); bei Claude stärker beim Anbieter — aber kein System entlastet die Behörde vollständig.

KRITISCHE SELBSTANALYSE (CLAUDE ÜBER CLAUDE) „Ich kann keine genuine moralische Verantwortung tragen. Bei einer fehlerhaften Empfehlung in einem Verwaltungsverfahren bin ich kausal beteiligt, aber kein moralischer Akteur. Die Verantwortung liegt beim Sachbearbeiter, der Institution und dem Anbieter — nicht bei mir.“

10 · Algorithmische Fairness & CLOUD Act

Das entscheidende rechtliche Risiko ist der **US CLOUD Act**: Er erlaubt US-Behörden Zugriff auf Daten US-amerikanischer Anbieter — auch bei EU-Speicherung. 2025 stufte die französische

CNIL die CLOUD-Act-Exposition als „im Wesentlichen gleichwertige Schutzversagung“ ein; die deutsche DSK verlangte Zusatzmaßnahmen über Standardvertragsklauseln hinaus. Für jede Claude-Beschaffung im Public Sector ist damit ein **Transfer Impact Assessment** (DSGVO Art. 46) verpflichtend — ein konkretes Beschaffungshindernis, kein theoretisches Risiko.

11 · Entscheidungsmatrix Public Sector

Kriterium	Mistral AI	Anthropic Claude
Rechtssitz	Frankreich (EU)	USA (CLOUD Act)
Datensouveränität	hoch — On-Premise möglich	mittel — EU Data Residency, US-Jurisdiktion
On-Premise	möglich (Open Weights)	nicht möglich (API-only)
Ethik-Governance	Neutrality-Prinzip, weniger formal	Constitutional AI, 4-Stufen-Hierarchie
Bias-Dokumentation	weniger formal	Model Cards, besser dokumentiert
EU AI Act	strukturell bevorzugt	Code of Practice, gute Positionierung
Enterprise-Reife	wachsend, noch früh	SOC 2, ISO 27001, HIPAA-ready

Use-Case-Kategorisierung für österreichische Behörden

Datenklasse	Empfehlung
Unkritisch / Low-Risk (interne Texte, Übersetzungen, FAQ, IT-Support)	Mistral bevorzugt — kein US-Transfer, On-Premise, niedrigere Kosten.
Sensible Verwaltungsdaten / High-Risk (Bürgerservice-Unterstützung, personenbezogene Dokumentenanalyse)	Claude Enterprise mit EU Data Residency — aber Transfer Impact Assessment (Art. 46) verpflichtend.
Hochsensibel (Gesundheit, Strafverfolgung, Sozialleistungen)	Kein kommerzielles System ohne Hybrid-Architektur; Mistral On-Premise + menschliche Kontrollpunkte.

TEIL C · SYNTHESE

Synthese, Handlungsempfehlungen & Prüfungsvorbereitung

Der rote Faden auf einen Blick

Punkt	Kerninsight	Anwendung Fallstudie
Technikphilosophie	Technik ist nie wertfrei — Werte stecken im Design	Mistral kodiert Souveränität, Claude deontologische Ethik
Collingridge	Folgen kaum vorhersehbar, Korrektur spät teuer	Deployment ohne Pilotierung ist Lehrbuchfall
Maschinenethik	Maschinen sind keine moral. Akteure — Mensch haftet	Responsibility Gap bei beiden unterschiedlich verortet
Kampfroboter	Extremfall: alle Punkte kollabieren gleichzeitig	tödliche/rechte-eingreifende Entscheidung = nie vollautonom
Maschinen & KI	Nicht ob, sondern wie KI uns verändert	Deskilling in Behörden = Posthumanismus im Kleinen
Digitale Ethik	Würde, Fairness, Transparenz, Kontrolle — nicht verhandelbar	CLOUD Act + Automation Bias = konkrete Governance-Risiken

Vier Handlungsempfehlungen für die Beschaffung

6. **CLOUD Act zuerst klären:** Für jede Public-Sector-Ausschreibung ein Transfer Impact Assessment (DSGVO Art. 46); bei Claude wird diese Analyse komplex — ein reales Beschaffungshindernis.
7. **On-Premise First:** Mistrals On-Premise-Fähigkeit ist der strukturelle Vorteil für hochsensible Daten — lokal deploybar, fine-tunebar, voll unter behördlicher Kontrolle.
8. **Auditierbarkeit sichern:** Claudes Constitutional-AI-Dokumentation ist für Hochrisiko-KI (AI Act Anhang III) der bessere Ausgangspunkt für Conformity Assessments.
9. **Hybride Beschaffung:** Mistral für datensensitive On-Premise-Szenarien, Claude Enterprise für komplexe High-Risk-Anwendungen — kein System ohne menschliche Kontrollpunkte.

> **VERNETZUNG** Diese Empfehlungen greifen in M3-LV2 (Trustworthy AI), M4-LV1 (Strategisches Management) und M4-LV3 (KI-Strategien & Management) — von der Ethik zur Beschaffungs- und Managemententscheidung.

10 Prüfungsfragen zum Selbsttest

1. Warum ist Technik nach Langdon Winner nicht neutral? Nennen Sie ein konkretes Beispiel.
2. Erklären Sie die zwei Hörner des Collingridge-Dilemmas. Warum ist es bei KI besonders scharf?
3. Was unterscheidet funktionale Moral (Stufe 2) von vollumfänglicher Akteursschaft (Stufe 3)?
4. Was ist eine Responsibility Gap — wie entsteht sie und wer trägt Verantwortung?
5. Warum legitimieren die drei Pro-Argumente für LAWS diese ethisch nicht?
6. Worin unterscheiden sich Trans- und Posthumanismus? Welche Implikation für Verwaltungen?

7. Was ist das Alignment-Problem und warum ist es schon mit heutigen LLMs relevant?

8. Was bedeutet prozedurale Fairness — und warum ist sie mehr als Ergebnis-Fairness?

9. Was ist der CLOUD Act und warum ist er für die KI-Beschaffung im EU-Public-Sector relevant?

10. Wann ist Mistral, wann Claude vorzuziehen — und wann ist kein kommerzielles System geeignet?

ÜBUNG

Essay-Übung — Modul 3

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

Forschungsfrage 1 · Werte im Design & demokratische Legitimation
„Ab welchem Punkt werden die in ein KI-System eingeschriebenen Werte (Trainingsdaten, Zielfunktion, Feature-Auswahl) zu einer demokratisch legitimationsbedürftigen Entscheidung — und wie lässt sich diese Legitimation im Beschaffungsprozess verankern?“

Abstract: Ausgehend von Winner (1980) und Feenberg untersucht der Essay, wie normative Setzungen in behördlichen KI-Systemen sichtbar und verhandelbar gemacht werden. Anhand eines Risiko-Scorings wird ein Kriterienraster (Zielfunktion, Feature-Governance, Fairness-KPI) entwickelt und gegen die AIA-/BVergG-Praxis geprüft. Erwarteter Beitrag: ein Verfahren, das Wertentscheidungen aus der Anbieter- in die demokratische Sphäre zurückholt.

Forschungsfrage 2 · Responsibility Gap & Accountability by Design
„Wie lässt sich die Responsibility Gap bei KI-gestützten Hoheitsentscheidungen organisatorisch und vertraglich schließen, ohne die Effizienzgewinne der Automatisierung aufzugeben?“

Abstract: Auf Basis von Misselhorns Stufenmodell und der Gap-Forschung (Vierkant, 2026) analysiert der Essay den KI-gestützten Bescheid (§ 93 BAO). Methodisch wird ein „Accountability by Design“-Referenzmodell (Rollen, Haftung, Override-Metriken) hergeleitet und gegen Automation Bias getestet. Erwarteter Beitrag: ein governance-konformes Zuständigkeitsdesign für hoheitliche KI.

Forschungsfrage 3 · Digitale Souveränität vs. ethische Governance-Reife
„Wie sollte eine österreichische Behörde bei der KI-Beschaffung den Trade-off zwischen digitaler Souveränität (Mistral/On-Premise) und formalisierter Ethik-Governance (Claude/Constitutional AI) gestaffelt nach Datensensitivität auflösen?“

Abstract: Der Essay verbindet Kapitel 6 mit der Fallstudie und dem CLOUD-Act-Risiko. Methodisch wird eine Entscheidungsmatrix (Datensensitivität × Stakes × Jurisdiktion → Hosting/Modell/HITL) entwickelt und an drei Use-Case-Klassen validiert. Erwarteter Beitrag: eine gestaffelte Beschaffungsdoktrin, die Souveränität und Auditierbarkeit situativ 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)

Amnesty International. (2021). Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. amnesty.org.

Collingridge, D. (1980). The social control of technology. St. Martin's Press.

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

Europäische Kommission. (2026). AI Act — regulatory framework & implementation timeline. digital-strategy.ec.europa.eu.

Europäisches Parlament & Rat der Europäischen Union. (2024). Verordnung (EU) 2024/1689 (KI-Verordnung).

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

Greenblatt, R., et al. (Anthropic & Redwood Research). (2024). Alignment faking in large language models. arXiv/anthropic.com.

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

Human Rights Watch. (2024). Killer robots: UN vote should spur treaty negotiations. hrw.org.

Miller, S. (2025). LAWS and meaningful human control. Ethics and Information Technology.

Misselhorn, C. (2018). Grundfragen der Maschinenethik. Reclam.

Saraiva, M. (2024). Moral agency in AI systems. International Journal of Philosophy.

Takemoto, K. (2024). The Moral Machine experiment on large language models. Royal Society Open Science.

Tolmeijer, S., et al. (2024). Beyond the trolley problem: A critique. AI and Ethics (Springer).

Vierkant, T. (2026). Accountability in the age of algorithms. Philosophy & Technology (Springer).

Wankhade, S., et al. (2025). Transparency and accountability in AI systems. Frontiers in Artificial Intelligence.

Werthner, H., et al. (2019). Wiener Erklärung zum Digitalen Humanismus. TU Wien / WU Wien. dighum.org.

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

Zhan, Y., & Wan, X. (2024). The trolley problem and autonomous driving. World Electric Vehicle Journal.