M19b - Ethik und Generative KI

Stand: 04.2025

1 Ethische Dimensionen

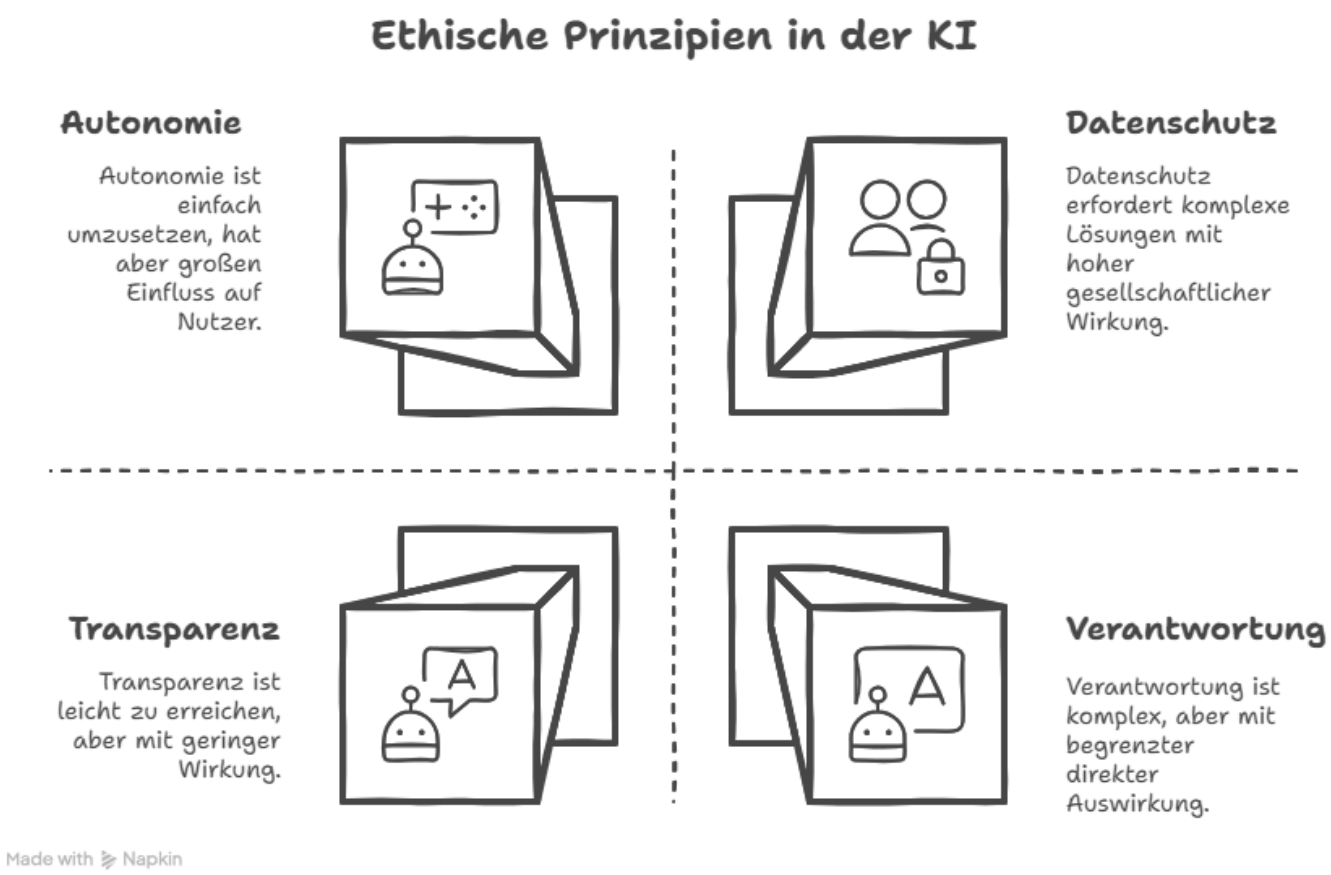
Definition & Abgrenzung:

- Generative KI (GenAI) erzeugt neue Inhalte (Texte, Bilder, Musik, Code) auf Basis gelernter Muster und stellt damit eine kreative, wenn auch nicht bewusste, Nachbildung menschlicher Ausdrucksformen dar.
- Abgrenzung zu anderen KI-Typen:
  - *Analytische oder prädiktive KI* analysiert bestehende Daten, um Vorhersagen zu treffen (z. B. Kredit-Scoring). Ethikfragen betreffen hier v. a. Fairness und Nachvollziehbarkeit der Entscheidungskriterien.
  - *Regelbasierte oder symbolische KI* folgt festen, menschenkodierten Entscheidungsregeln (z. B. Expertensysteme) und ist meist gut erklärbar.
  - *AGI* (Artificial General Intelligence) beschreibt eine hypothetische KI mit menschenähnlicher, allgemeiner Intelligenz. Sie existiert aktuell nicht, ist aber zentrales Thema in der KI-Ethik-Forschung.

  Die Abgrenzung ist wichtig, da **generative KI besonders intransparente, kreative Outputs erzeugt**, was neue ethische Fragestellungen aufwirft – etwa zur Originalität, Verantwortung und Manipulationsgefahr.

Zentrale ethische Prinzipien:

- **Verantwortung:** Wer haftet bei Fehlentscheidungen? Eine klare juristische und ethische Zuweisung ist meist schwierig.
- **Fairness:** Gefahr der Reproduktion sozialer Ungleichheiten durch Daten-Bias; bedarf systematischer Überprüfung.
- **Transparenz:** Undurchschaubarkeit der Modelle verhindert Vertrauen und kontrollierte Anwendung.
- **Datenschutz:** Besonders problematisch bei sensiblen Daten wie Gesundheitsdaten oder intimen Nutzereingaben.
- **Autonomie:** Nutzer:innen dürfen nicht entmündigt werden; Systeme müssen überschreibbar bleiben.
- **Sicherheit:** Technisch wie gesellschaftlich müssen Risiken minimiert werden, etwa durch Missbrauchsprävention.



Akteure:

- Technologieunternehmen (oft marktgetrieben), Forschung (wissensgetrieben), Politik (regulierend), Zivilgesellschaft (wertorientiert), Nutzer:innen (praktisch-orientiert) prägen gemeinsam das öffentliche Verständnis und die Entwicklungspfade von KI.

2 Rahmenwerke & Praxis

Regulatorische Grundlagen:

- Der **EU AI Act** ist das weltweit erste umfassende Gesetz zur Regulierung von KI und unterteilt Systeme in vier Risikokategorien. Besonders generative KI mit hohem Einfluss auf Meinungsbildung und Kreativbereiche steht dabei unter besonderer Beobachtung.
- **OECD- und UNESCO-Richtlinien** setzen wichtige normative Standards, die Fairness, Erklärbarkeit und Rechenschaftspflicht als universelle Prinzipien formulieren.

Umsetzung in der Industrie:

- Unternehmen wie OpenAI, Google und Meta haben ethische Leitlinien entwickelt, die Aspekte wie Moderation, Red Teaming und Prompt-Engineering einschließen.
- Tools wie SynthID oder Wasserzeichen-Lösungen fördern Transparenz und Fälschungsschutz.
- Trotzdem bleibt die Selbstverpflichtung oft hinter regulatorischen Anforderungen zurück.

Organisatorische Umsetzung:

- Interne Ethikboards, Compliance-Beauftragte und Prozesse zur Ethikfolgenabschätzung gewinnen an Bedeutung.
- Wichtig ist nicht nur die Existenz, sondern die Integration ethischer Reflexion in agile Entwicklungsprozesse.

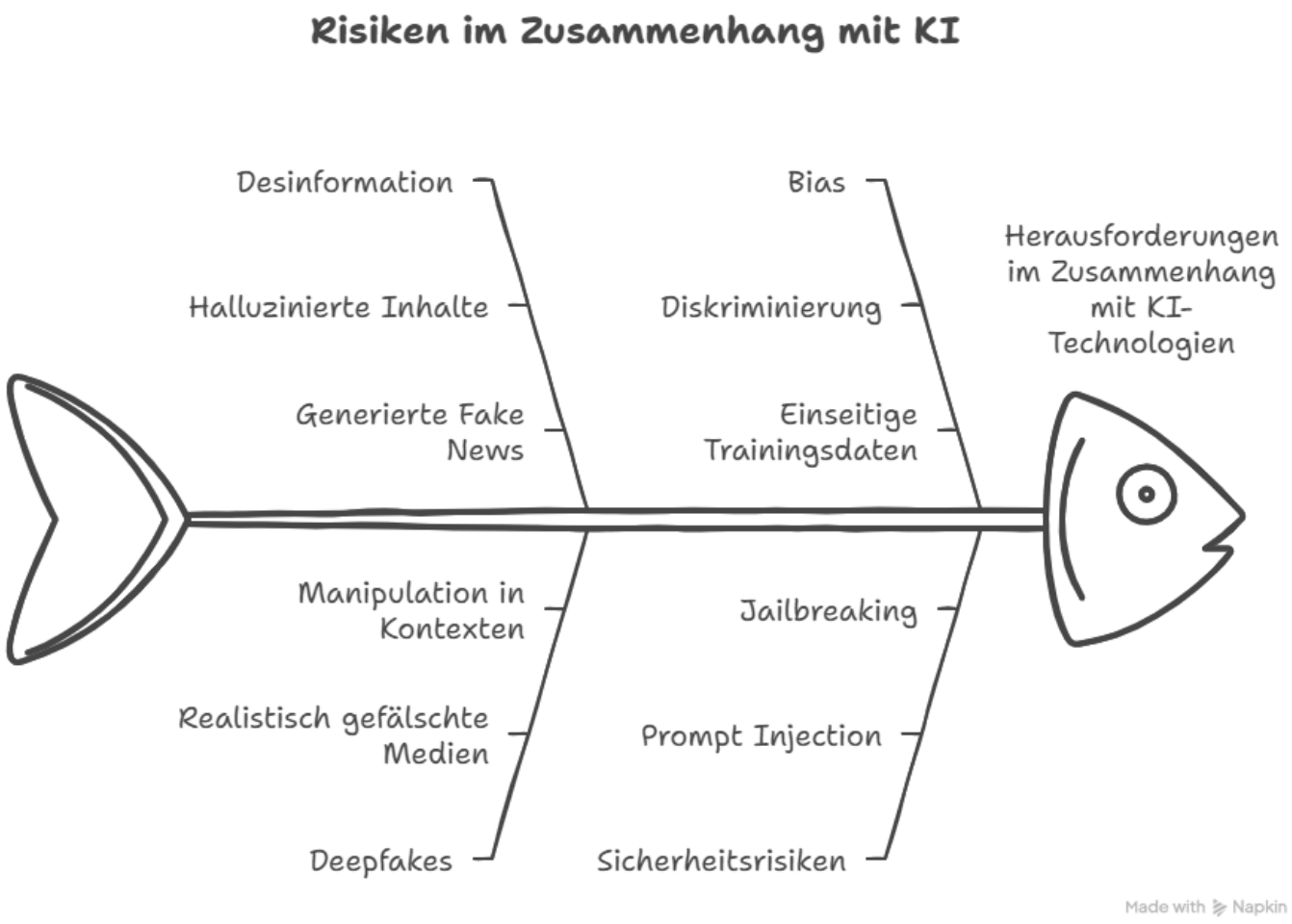
Rolle von Bildung & Forschung:

- Hochschulen und Ausbildungsinstitutionen entwickeln spezialisierte Curricula zu KI-Ethik.
- Praxisnahe Formate wie Fallanalysen, Planspiele oder interdisziplinäre Projektarbeit sind besonders wirksam.

3 Risiken & Fehlerquellen

Kernrisiken:

- **Desinformation:** Halluzinierte Inhalte und generierte Fake News untergraben Vertrauen in Informationen.
- **Deepfakes:** Realistisch gefälschte Medien können zur Manipulation in politischen oder wirtschaftlichen Kontexten führen.
- **Bias:** Diskriminierung aufgrund einseitiger Trainingsdaten ist ein zentrales Problem.
- **Rechtsunsicherheit:** Unklarheiten bei Urheberrecht und Datenschutz hemmen klare Verantwortungszuweisung.
- **Sicherheitsrisiken:** Prompt Injection, Jailbreaking oder Training mit vergifteten Daten sind reale Angriffsvektoren.



Ethische Spannungsfelder:

- Innovation vs. Regulierung
- Transparenz vs. Datenschutz oder geistiges Eigentum
- Open Source vs. Missbrauch
- Automatisierung vs. Arbeitsplatzverlust

Fehlerquellen im Lebenszyklus:

- *Daten:* Verzerrung durch schlechte oder einseitige Quellen.
- *Modell:* Fehlende Robustheit, Halluzinationen, Black-Box-Verhalten.
- *Prozess:* Mangelnde Tests, nicht-diverse Teams, unklare Verantwortlichkeiten.
- *Nutzung:* Missbrauch für Desinformation, unreflektierte Übernahme von KI-Outputs.

Verantwortung:

- Verteilte Verantwortung in Wertschöpfungsketten erschwert Haftung und Governance.
- Neue regulatorische Ansätze wie der AI Act definieren Rollen und Pflichten neu.

4 Chancen & Potenziale

Gesellschaftlicher Mehrwert:

- **Bildung:** Automatisierte Lernpfade, intelligente Nachhilfe und adaptive Lernumgebungen.
- **Barrierefreiheit:** Text-zu-Sprache, visuelle Erkennung, einfache Sprache für mehr Teilhabe.
- **Wissenschaft:** Hypothesengenerierung, Datenanalyse, automatisierte Literaturauswertung.
- **Kreativität:** Unterstützung für künstlerische Prozesse und Demokratisierung kreativer Mittel.
- **Wirtschaft:** Automatisierung von Routineaufgaben, Entlastung von Fachkräften.
- **Nachhaltigkeit:** Umweltmonitoring, Klimamodellierung, Analyse von ESG-Daten.

Ethics by Design:

- Ethische Reflexion bereits in der Designphase einbetten.
- Interdisziplinäre Teams, Impact Assessments und diverse Perspektiven sind zentral.

Gemeinwohlorientierte KI:

- Open-Source-Modelle, öffentliche KI-Infrastrukturen (z. B. EU AI Factories), transparente Standards.
- Ziel: Technologische Souveränität und gerechter Zugang zu KI.

5 Best Practices

Technische Maßnahmen:

- **Explainable AI (XAI):** Methoden wie LIME, SHAP, RAG-basierte Erklärungen fördern Transparenz.
- **Bias-Mitigation:** In allen Phasen (Pre-, In-, Postprocessing), begleitet von Fairness-Audits.
- **Sicherheit:** Red Teaming, Input-Validierung, Zugriffskontrollen, Inhaltsfilter.
- **Datenschutz:** Anonymisierung, Pseudonymisierung, differenzielle Privatsphäre.
- **Transparenz:** Dokumentation, Wasserzeichen, klare Kommunikation der KI-Nutzung.

Organisatorische Strategien:

- Etablierung klarer Verantwortlichkeiten für KI im Unternehmen.
- Entwicklung und Pflege von KI-Ethik-Kodizes.
- Integration von ethischer Reflexion in Produktentwicklung und -bewertung.

Bildungs- und Schulungsinitiativen:

- Schulung für Entwickler (z. B. Bias, Datenschutz, XAI).
- Aufklärung von Anwendern über Grenzen, Risiken und verantwortungsvollen Umgang mit KI.

Rahmenwerke und Tools:

- Nutzung internationaler Standards (z. B. NIST AI RMF, EU AI Act, ISO 42001).
- Checklisten für ethisch orientierte Entwicklung und Deployment.