AI Factory Austria AI:AT

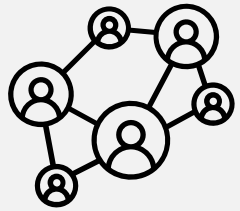
Trustworthy AI für Fortgeschrittene: Ethische Aspekte

Dr. Peter Biegelbauer
Co-Lead Legal, Regulatory and Ethics

Warum brauchen wir die AI Factory Österreich?

Souveränität

Ethik und Trustworthiness



KI-Ökosystem

Unsere Mission
Von der Idee zur Innovation

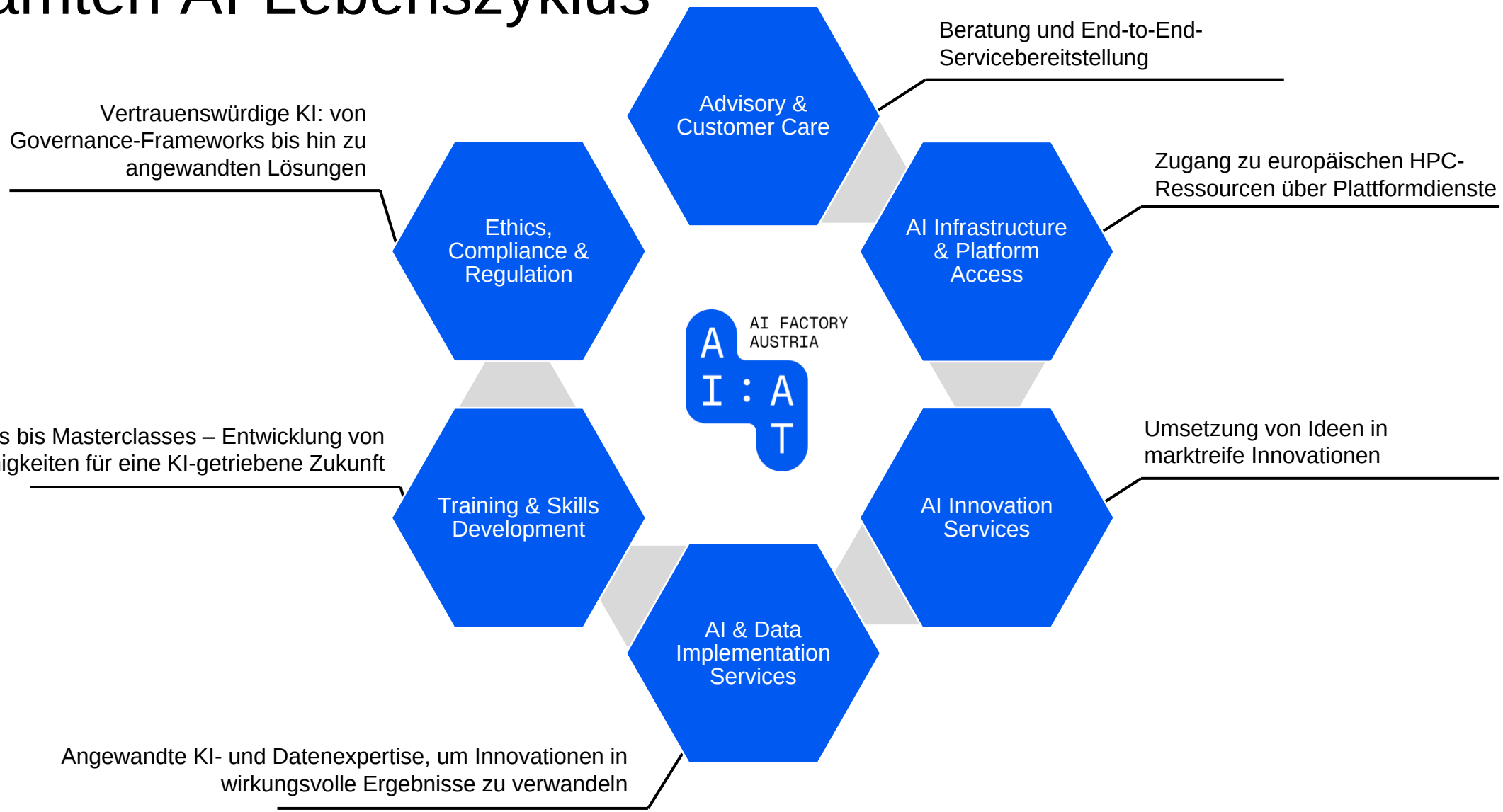
Etablierung eines One-Stop Shop für AI

Schließen des AI Ressourcen- und Wissens-Gap

Bewerbung von ethical und trustworthy AI

Launchpad Für AI-getriebene Innovation

Unsere Services unterstützen über den gesamten AI-Lebenszyklus



AI Factory Austria AI:AT – Team

AI Factory Austria AI:AT Consortium

Disclaimer:
The speakers are solely sharing their personal experiences. Therefore, this free seminar is not a substitute for professional/legal advice.

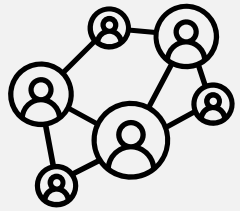
Beneficiaries

Affiliated Entities

Why do we need AI Factory Austria?

Sovereignty

Ethics and Trustworthiness



Connecting the Ecosystem

Fallstudie: Robodebt

Aufdecken von Sozialleistungsbetrug in Australien

🤖 **Automatisiertes System (Robodebt) sollte Sozialleistungsbetrug aufdecken**

- Fehlerhafte Datenabgleiche zwischen Steuer- und Sozialbehörden
- als Folge: Falschbeschuldigungen hunderttausender Bürger:innen
- als Folge: Massive finanzielle & psychische Belastungen, teils tragische Konsequenzen

🏛 **Politische und rechtliche Konsequenzen**

- Offizielle Entschuldigung der Regierung, Rückzahlungen & Entschädigungen > 1,8 Mrd. AUD
- Royal Commission (2022–23) mit umfassender Untersuchung
- Stärkung von Rechtsstaatlichkeit, Transparenz & Aufsicht
- Verbesserte Kontrollmechanismen für staatliche KI
- Klare Verantwortlichkeiten für Behörden & Politik

Robodebt: was hätte es gebraucht?

- **Robuste Datenvalidierung & -integration:**

  Aufbau sicherer Schnittstellen zwischen Sozialdatenbanken, Steuerbehörde und anderen Quellen mit automatisierter Fehlerprüfung vor Berechnungen.

- **Transparente Algorithmusgestaltung:**

  Offenlegung der Entscheidungslogik und Dokumentation der Gewichtungen; verpflichtende „explainability"-Mechanismen für jede automatisierte Entscheidung.

- **Human-in-the-Loop-Prinzip:**

  Automatisierte Vorschläge dürfen keine endgültigen Verwaltungsakte auslösen – verpflichtende menschliche Prüfung bei *jeder* Rückforderungsentscheidung.

- **Algorithmisches Impact Assessment (AIA):**

  Vor Einführung KI-basierter Systeme: Risiko- und Fairnessprüfung mit Fokus auf Diskriminierung, Fehlklassifikationen und Verfahrensgerechtigkeit.

Robodebt: was hätte es gebraucht?

- **Kontinuierliche Auditierung & Monitoring:**

Einrichtung unabhängiger Kontrollstellen zur Überwachung der Datenqualität, Modellperformance und der sozialen Auswirkungen automatisierter Entscheidungen.

- **Rechtsstaatliche Rückkopplungsschleifen:**

Implementierung effektiver Beschwerdemechanismen und automatischer Benachrichtigungssysteme bei algorithmisch generierten Bescheiden.

- **Ethik- und Governance-Rahmen:**

Verankerung ethischer Leitlinien (z. B. Fairness, Rechenschaftspflicht, Transparenz) im gesamten Lebenszyklus der KI – von Entwicklung bis Einsatz.

KI-Ethik braucht das Land!

- Internationaler Review von **KI-Guidelines** (Jobin et al. 2019):
  - transparency
  - justice and fairness
  - non-maleficence
  - responsibility
  - privacy
  - beneficence
  - freedom and autonomy
  - trust
  - sustainability
  - dignity and solidarity

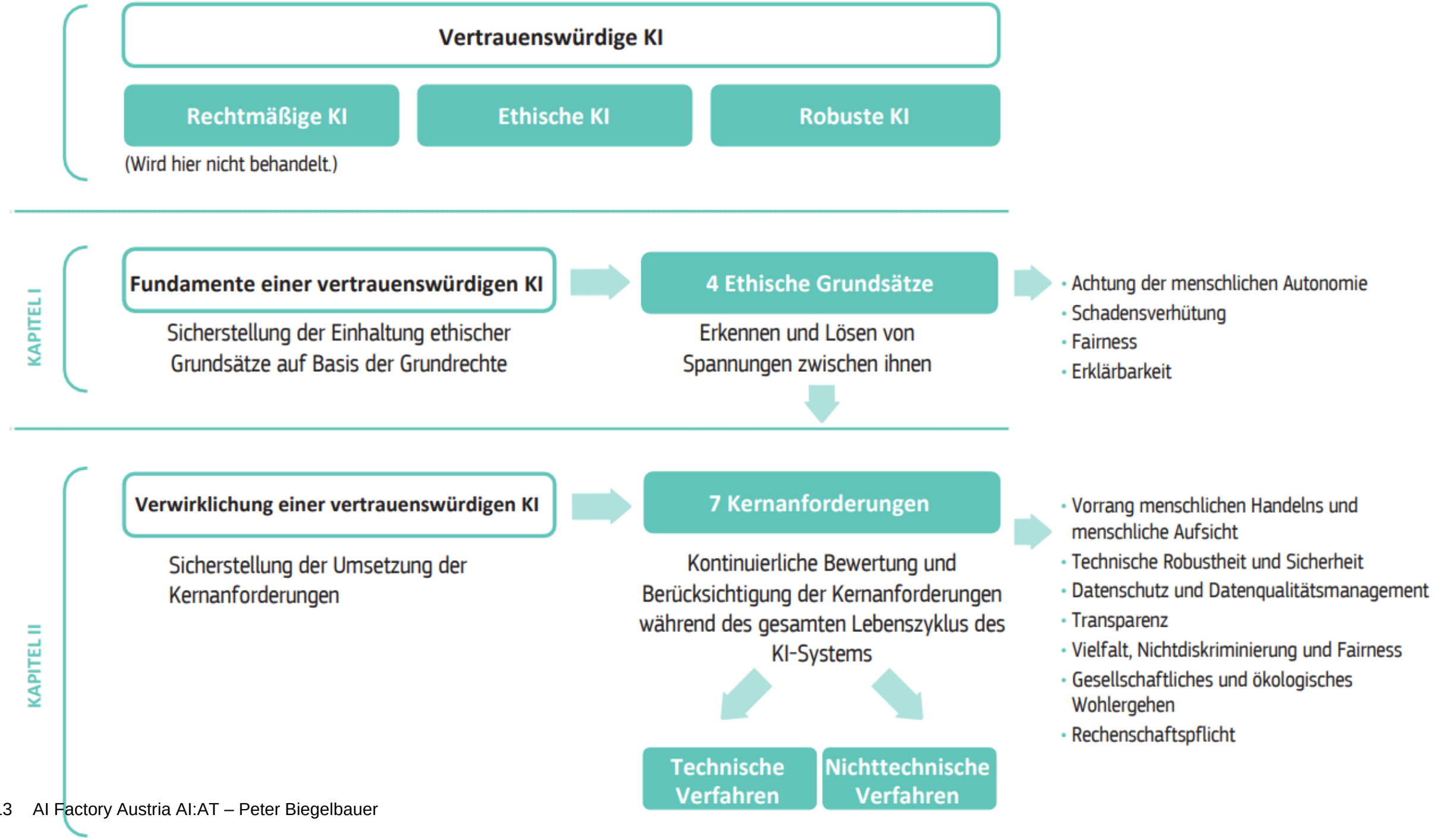
  - Transparenz
  - Gerechtigkeit und Fairness
  - Schadensvermeidung
  - Verantwortung
  - Privatsphäre
  - Wohlwollen
  - Freiheit und Autonomie
  - Vertrauen
  - Nachhaltigkeit
  - Würde und Solidarität

Ethische Prinzipien im AI-Act

Integration der **HLEG-Prinzipien für vertrauenswürdige KI** in den AI-Act (Erwägungsgründe):

- Vorrang menschlichen Handelns und menschliche Aufsicht
- Technische Robustheit und Sicherheit
- Datenschutz und Datenqualitätsmanagement
- Transparenz
- Vielfalt, Nichtdiskriminierung und Fairness
- Gesellschaftliches und ökologisches Wohlergehen

Das fehlende Prinzip: Accountability (Rechenschaftspflicht) → AI Liability Directive (geplant)


Vertrauenswürdige KI
Rechtmäßige KI
Ethische KI
Robuste KI
(Wird hier nicht behandelt.)
KAPITEL I
Fundamente einer vertrauenswürdigen KI
Sicherstellung der Einhaltung ethischer Grundsätze auf Basis der Grundrechte
4 Ethische Grundsätze
Erkennen und Lösen von Spannungen zwischen ihnen
Achtung der menschlichen Autonomie
Schadensverhütung
Fairness
Erklärbarkeit
KAPITEL II
Verwirklichung einer vertrauenswürdigen KI
Sicherstellung der Umsetzung der Kernanforderungen
7 Kernanforderungen
Kontinuierliche Bewertung und Berücksichtigung der Kernanforderungen während des gesamten Lebenszyklus des KI-Systems
Technische Verfahren
Nichttechnische Verfahren
Vorrang menschlichen Handelns und menschliche Aufsicht
Technische Robustheit und Sicherheit
Datenschutz und Datenqualitätsmanagement
Transparenz
Vielfalt, Nichtdiskriminierung und Fairness
Gesellschaftliches und ökologisches Wohlergehen
Rechenschaftspflicht

KI-Prinzipien in Österreich

Leitfaden digitale Verwaltung: KI, Ethik und Recht

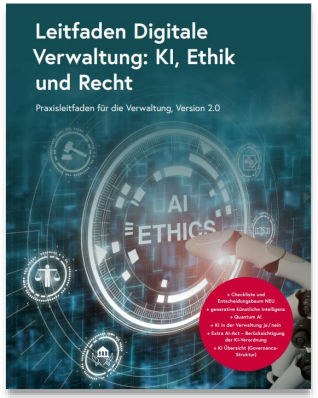
Eine Arbeit des AIT AI Ethics Lab für BMKÖS / BKA

Webpage

Leitfaden

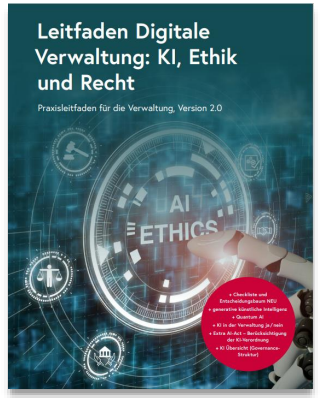
YouTube

Kriterien im Leitfaden digitale Verwaltung I



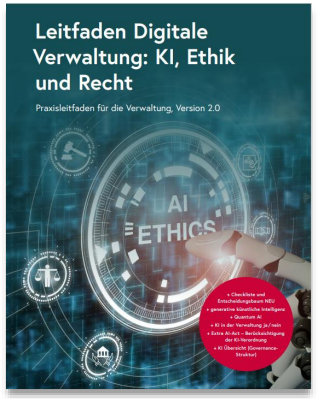
1. **Recht –** Einhaltung des geltenden Rechts, KI-Anwendung muss die einschlägigen Gesetze und Vorschriften einhalten, einschließlich der Grundrechte
2. **Transparenz –** Informationen über die KI-Anwendung verfügbar und zugänglich machen, Transparenz in KI-Entscheidungsprozessen fördern, Öffentlichkeit und Verwaltungsbedienstete über die Ziele und der KI-Anwendung informieren, Offenlegung der Entscheidungsergebnisse
3. **Unparteilichkeit und Fairness –** KI-Anwendung muss unvoreingenommene und vielfältige Daten und Modelle verwenden, Vermeidung der Aufrechterhaltung bestehender Vorurteile, Fairness der KI-Anwendung im Kontext der öffentlichen Verwaltung
4. **Effektivität und Effizienz –** Einsatz von KI-Anwendungen in der Verwaltung muss deren Effektivität und Effizienz nachhaltig verbessern, ohne die Arbeitssituation der im öffentlichen Dienst tätigen Menschen zu verschlechtern

Kriterien im Leitfaden digitale Verwaltung II



5. **Sicherheit –** KI-Anwendung muss sicher eingesetzt werden, Schutz sensibler Informationen, Vermeidung unbefugten Zugriffs
6. **Barrierefreiheit und Inklusion –** KI-Anwendung muss für Menschen mit unterschiedlichen Fähigkeiten, Hintergründen und Kulturen zugänglich und integrativ sein, Angebot von Alternativen zur KI-Technologie für gleichberechtigten Zugang zu öffentlichen Dienstleistungen
7. **Rechenschaftspflicht –** Klare Zuständigkeiten und Verantwortlichkeiten, Bewusstsein bei Verantwortlichen über ihre Verantwortung
8. **Digitale Souveränität –** Die Verwaltung muss in der Lage sein die Entwicklung von KI-Lösungen zu beeinflussen, unabhängig anzuwenden und vertrauliche Daten in ihrem eigenen Einflussbereich zu halten

Checkliste im Leitfaden digitale Verwaltung



Basierend auf den Kriterien

Menschliche Aufsicht

Wurde die KI-Anwendung so entwickelt, dass menschliche Aufsicht möglich ist (z.B. human-in-the-loop, human-on-the-loop)? (Siehe Wissen: Human in the Loop, Human on the Loop)	◯ Ja	◯ Nein
Wird das KI-System in regelmäßigen Abständen überprüft (zumindest in Bezug auf Leistung/Qualität, Sicherheit, Einhaltung der geltenden Gesetze und Vorschriften)?	◯ Ja	◯ Nein

Rechenschaftspflicht

Sind klare Verantwortlichkeiten für Entwickler:innen, Betreiber:innen und Nutzer:innen der KI-Anwendung festgelegt?	◯ Ja	◯ Nein
Wurde festgelegt, wer die letztendliche Verantwortung und Rechenschaftspflicht für den KI-Einsatz sowie die Ausgaben des KI-Systems trägt?	◯ Ja	◯ Nein

- Vier Seiten mit Fragen zu den Bereichen
  - Recht,
  - Transparenz,
  - Unvoreingenommenheit und Fairness,
  - Effektivität und Effizienz,
  - Sicherheit,
  - Zugänglichkeit und Inklusion,
  - menschliche Aufsicht,
  - Rechenschaftspflicht,
  - digitale Souveränität.
- Geschlossene Fragen, die als Indikatoren der Erfüllung der wesentlichsten ethischen Kriterien gelten können.

Zum Beispiel: Fairness

Aktuelle Debatten rund um Fairness und KI:

- Unfaire Behandlung von Individuen oder Gruppen aufgrund von Verzerrungen in KI-Entscheidungen
- Undurchsichtigkeit von KI-Entscheidungen („Black Box“ KI)
- Verschiedene Gefahren für die Demokratie und das gesellschaftliche Wohlergehen
- Ungleichheiten auf dem Markt durch die Macht von Big Tech

Fairness im AI Act: “*Diversity, non-discrimination and fairness means that AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.”*

Herausforderung:

- Kein einheitliches Fairnesskonzept in einer pluralistischen Gesellschaft
- Technologie als Werkzeug, nicht als Allheilmittel

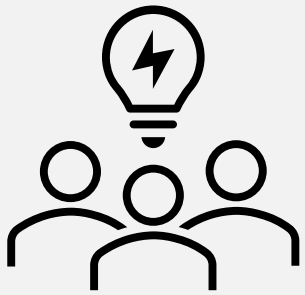
Verschiedene Perspektiven auf das Kriterium Fairness

- Anerkennung **moralischer Gleichheit**: Entscheidungen sollten gleiche Achtung und Würde für alle Menschen widerspiegeln
- Schutz vor **struktureller Diskriminierung**: Historische Ungleichheiten dürfen nicht verstärkt oder verfestigt werden
- Transparenz als **Voraussetzung für Gerechtigkeit**: Faire Entscheidungen erfordern Nachvollziehbarkeit und öffentliche Rechenschaft
- **Verteilungsgerechtigkeit** mit Blick auf Machtverhältnisse: Nutzen und Schaden sollte gerecht (z.B. über soziale Gruppen) verteilt sein
- Berücksichtigung **kontextueller Gerechtigkeit** - Fairness ist nicht universell einheitlich: Was als fair gilt, muss sich an kulturellen und sozialen Kontexten orientieren

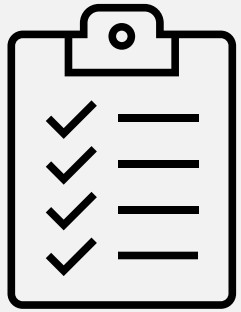
…und auf das Thema Fairness in Bezug auf KI

- **Demografische Parität**: Gleiche positive Entscheidungsraten für alle Gruppen, unabhängig von sensiblen Attributen wie Geschlecht oder Ethnie
- **Gleichheit der Fehlerraten**: Gleiche Wahrscheinlichkeiten für False Positives und False Negatives zwischen Subgruppen, um faire Fehlerverteilungen zu gewährleisten
- **Individuelle Fairness**: Ähnliche Individuen erhalten ähnliche Entscheidungen
- **Gleichheit der Chancen**: Gleiche True Positive Rate über Gruppen hinweg
- **Kontextualisierte Fairness**: Berücksichtigt gesellschaftliche, rechtliche und kulturelle Rahmenbedingungen; ev. domänenspezifische Gerechtigkeitskonzepte berücksichtigen

Wie wird Fairness implementiert?



Ex-ante: *System Design*
Ex-post: *Produkt Nutzungskontrolle*



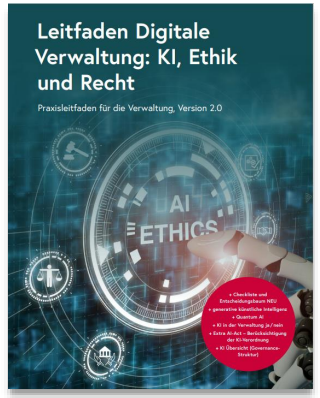
Impact/Risk Assessment

Beteiligung von Interessensgruppen

Qualitätsstandards

Fairnessmetriken

Und im Leitfaden digitale Verwaltung?



Checkliste: Kriterium Fairness

Unvoreingenommenheit und Fairness

Sind die Daten, die zum Training des KI-Systems verwendet werden, vielfältig und repräsentativ für den jeweiligen Kontext? (Siehe 7.1 Wissen: Bias; Anwendung: Geschlechterbias)	◯ Ja	◯ Nein
Gibt es einen Prozess, um verwendete Datenquellen auf mögliche Verzerrungen und Ungenauigkeiten zu prüfen?	◯ Ja	◯ Nein
Ist die KI-Anwendung so konzipiert, dass sie die Entmenschlichung, Diskriminierung, Stereotypisierung oder Manipulation von Menschen vermeidet? (Siehe 5 Anwendungsfall: Chatbot)	◯ Ja	◯ Nein
Gibt es ein Verfahren, mit dem Personen gegen den Einsatz bzw. die Ausgabe des KI-Systems Einspruch oder sonstige Rechtsmittel dagegen erheben können?	◯ Ja	◯ Nein

Fallstudie: Apple Pay

KI-gestütztes Risikomodell der Apple Card (entwickelt mit Goldman Sachs)

- Algorithmische Diskriminierung bei Kreditlimits zeigte systematische Verzerrungen: Frauen erhielten häufig deutlich niedrigere Kreditlimits als Männer bei vergleichbarer Bonität
- Intransparente Modelllogik & fehlende Erklärbarkeit: zugrundeliegende Entscheidungsalgorithmen waren nicht nachvollziehbar; Kundinnen erhielten keine Begründungen
- Regulatorische und ethische Defizite: Fehlende interne Fairness-Tests, unzureichende Compliance-Kontrollen führten Verletzungen des Equal Credit Opportunity Acts
- Lehren für KI-Governance: KI-Modelle ohne Fairness-Audits und Bias-Monitoring verdeutlichen Notwendigkeit von Explainable AI, Fairness-Metriken und menschlicher Aufsicht

https://www.bbc.com/news/business-50365609

Q & A

Funded by

under discussion with

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).

Contact

Dr. Peter Biegelbauer

Co-Lead Legal, Regulatory and Ethics
AI Factory Austria AI:AT

+43 664 88390033
peter.biegelbauer@ai-at.eu

AI Factory Austria AI:AT
Schwarzenbergplatz 2
1010 Wien, Austria

info@ai-at.eu
ai-at.eu@ai-factory-austria