AI Factory Austria AI:AT

AI FACTORY AUSTRIA

Trustworthy AI Advanced: Ethical Aspects

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

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.

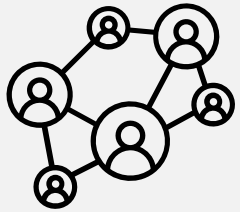
Beneficiaries

Affiliated Entities

Why do we need AI Factory Austria?

Sovereignty

Ethics and Trustworthiness



Connecting the Ecosystem

Case: Robodebt

Detecting social security fraud in Australia

🤖Automated system (Robodebt) was intended to detect social security fraud

- Incorrect data comparisons between tax and social security authorities
- Result: false accusations against hundreds of thousands of citizens
- Result: massive financial and psychological stress, sometimes with tragic consequences (suicides!)

🏛️ Political and legal consequences

- Official apology from the government, repayments and compensation > AUD 1.8 billion
- Royal Commission (2022–23) with comprehensive investigation
- Strengthening rule of law, transparency and oversight
- Improved control mechanisms for government AI
- Clear responsibilities for authorities and politicians

Robodebt: What would have been needed?

- **Robust data validation and integration:**

Establishment of secure interfaces between social databases, tax authorities and other sources

with automated error checking prior to calculations

- **Transparent algorithm design:**

Disclosure of decision-making logic and documentation of weightings;

mandatory explainability mechanisms for every automated decision

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

Before introducing AI-based systems:

risk and fairness assessment with a focus on discrimination misclassification and procedural fairness

- **Human-in-the-Loop-Principle:**

Automated suggestions must not trigger final administrative acts:

mandatory human review for every recovery decision

Robodebt: What would have been needed?

- **Human-in-the-Loop-Principle:**

Automated suggestions must not trigger final administrative acts:

mandatory human review for every recovery decision

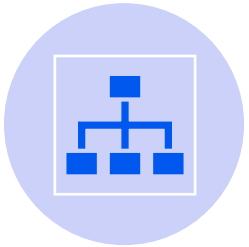
→CASE ‘Human in the Loop’ as catch-all remedy?

- Profiling algorithm (ADM) in the Polish labour market authority for assigning unemployed persons to training
- Formal final decision made by authority staff aka ‘Human in the Loop’
- Despite widespread scepticism about accuracy and fairness of algorithm, staff rarely correct algorithmic decisions
- Corrections in x<4% of cases!!
- Article: Karolina Sztandar-Sztanderska (2025)

Case ‘Human in the Loop’: Explanations

Human factors: automation bias

Policy-related factors: unclear legal regulations



Organisational culture: more corrections in autonomous organisational cultures than in hierarchical settings

Working conditions: high caseloads reduced willingness to correct, as deviations had to be justified

Profession-related factors: Employees with further training (e.g. counselling, psychology) corrected ADM more

Technology-related factors: lack of transparency in decision-making logic increased fear of sanctions

Back to Robodebt: What would have been needed?

- **Continuous auditing and monitoring**:

Establishment of independent control bodies to monitor data quality, model performance and the social impact of automated decisions

- **Rule of law feedback loops**:

Implementation of effective complaint mechanisms and automatic notification systems for algorithmically generated decisions

- **Ethics and governance framework**:

Ethics guidelines (e.g., fairness, accountability, transparency) throughout the entire AI lifecycle – from development to deployment

Peter Biegelbauer, Lone Pine Tree Sactuary, Brisbane

AI Ethics: Principles

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

Peter Biegelbauer, train station, Sydney

Ethics Principles: AI-Act

Integration of the HLEG principles for Trustworthy AI into the AI Act (see recitals):

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Social and environmental well-being

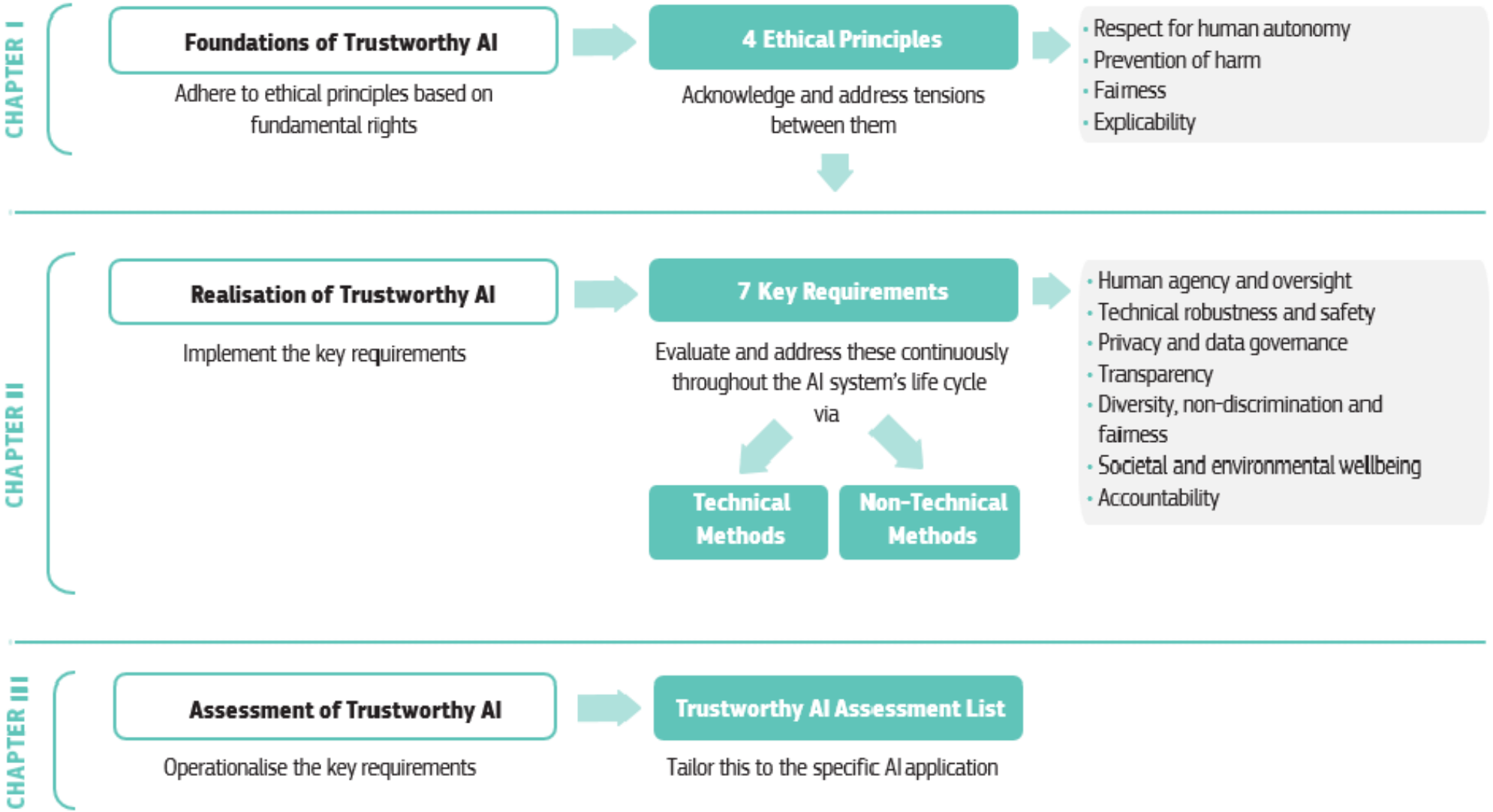
The missing principle: Accountability → AI Liability Directive (planned)

→Ethics Guidelines for Trustworthy Artificial Intelligence (2019)

→Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (2020)

→Also ALTAI Webtool

Ethics Guidelines for Trustworthy Artificial Intelligence



AI Principles in Austria

Guidelines Digital Public Administration: AI, Ethics, and Law

AIT AI Ethics Lab for BMKÖS / BKA

Webpage

Guideline

YouTube

Criteria for Trustworthy AI / Responsible AI

1. **Law** – Compliance with applicable law; AI applications must comply with relevant laws and regulations, including fundamental rights

2. **Transparency** – Making information about AI applications available and accessible; promoting transparency in AI decision-making processes; informing the public and administrative staff about the objectives and use of AI applications; disclosure of decision outcomes

3. **Impartiality and fairness** – AI applications must use unbiased and diverse data and models, avoid perpetuating existing biases, and ensure fairness in the context of public administration

4. **Effectiveness and efficiency** – The use of AI applications in administration must improve their effectiveness and efficiency in a sustainable manner without worsening the working conditions of public servants

Criteria for Trustworthy AI / Responsible AI

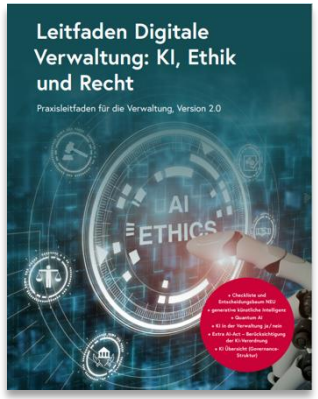
5. **Security** – AI applications must be used securely, protecting sensitive information and preventing unauthorised access

6. **Accessibility and inclusion** – AI applications must be accessible and inclusive for people with different abilities, backgrounds and cultures, offering alternatives to AI technology for equal access to public services

7. **Accountability** – Clear responsibilities and accountabilities, awareness among those responsible of their responsibilities

8. **Digital sovereignty** – The administration must be able to influence the development of AI solutions, apply them independently and keep confidential data within its own sphere of influence

Checklist in Guidelines Digital Public Administration



Based on criteria

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

- Four pages of questions covering the areas of
  - law
  - transparency
  - impartiality and fairness
  - effectiveness and efficiency
  - security
  - accessibility and inclusion
  - human oversight
  - accountability
  - and digital sovereignty
- Closed questions that can be considered indicators of compliance with the most essential ethical criteria

Fairness

Current debates surrounding fairness and AI:

- Unfair treatment of individuals or groups due to biases in AI decisions
- Opacity of AI decisions ('black box' AI)
- Various threats to democracy and social welfare
- Market inequalities due to the power of Big Tech

Fairness in 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.*"

Challenges:

- No uniform concept of fairness in a pluralistic society
- Technology as a tool, not a panacea

Different Perspectives on Fairness…

- **Recognition of moral equality:**
  Decisions should reflect equal respect and dignity for all people
- **Protection from structural discrimination:**
  Historical inequalities must not be reinforced or perpetuated
- **Transparency as a prerequisite for justice:**
  Fair decisions require traceability and public accountability
- **Distributive justice with regard to power relations:**
  benefits and harms should be distributed fairly (e.g., across social groups)
- **Consideration of contextual justice – fairness is not universally uniform:**
  what is considered fair must be based on cultural, social, and political contexts

GettyImages-904420104

…and on Fairness in Relation to AI

- **Demographic parity:**

  Equal positive decision rates for all groups, regardless of sensitive attributes such as gender or ethnicity

- **Equality of error rates:**

  Equal probabilities for false positives and false negatives between subgroups to ensure fair error distributions

- **Individual fairness:**

  Similar individuals receive similar decisions

- **Equality of opportunity:**

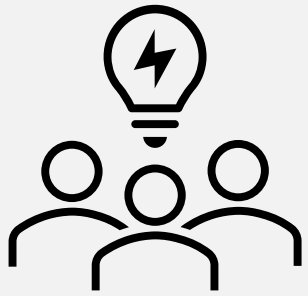
  Same true positive rate across groups

- **Contextualised fairness:**

  Takes into account social, legal, and cultural conditions; possibly considers domain-specific concepts of fairness

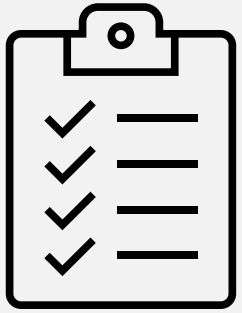
GettyImages-904420104

How is Fairness Implemented?



Ex-ante: *System Design*

Ex-post: *Product usage control*



Impact/Risk Assessment

Stakeholder involvement

Quality standards

Fairness metrics

Fairness: Guidelines Digital Public Administration?

Checklist: 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

- **Impartiality and Fairness**
  - Are the data used to train the AI system diverse and representative for the respective context?
  - Is there a process in place to check the data sources used for possible biases and inaccuracies?
  - Is the AI application designed in such a way that it avoids dehumanization, discrimination, stereotyping, or manipulation of people?
  - Is there a procedure that allows individuals to lodge objections or pursue other legal remedies against the use or output of the AI system?

Case: Apple Pay

Apple Card AI-powered risk model (developed with Goldman Sachs)

- **Algorithmic discrimination in credit limits revealed systematic biases:**

Women often received significantly lower credit limits than men with comparable credit ratings

- **Non-transparent model logic & lack of explainability:**

Algorithms underlying decision were not clear; female customers received no explanations

- **Regulatory and ethical shortcomings:**

Lack of internal fairness tests and insufficient compliance controls led to violations of the Equal Credit Opportunity Act

- **Lessons for AI governance:**

AI models without fairness audits and bias monitoring highlight the need for explainable AI, fairness metrics and human oversight

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

Contact

Dr. Peter Biegelbauer

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

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

linkedin.com/in/peterbiegelbauer/

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

training@ai-at.eu
info@ai-at.eu

ai-at.eu

@ai-factory-austria

Funded by

under discussion with

EuroHPC
Joint Undertaking

Funded by
the European Union

Federal Ministry
Innovation, Mobility
and Infrastructure
Republic of Austria

FFG
Promoting Innovation.

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).