hochschule burgenland
University of Applied Sciences

FAIR-AI
Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence

# Workshop zu vertrauenswürdiger KI ein Auszug

*Mag. Dr. Julia Kreyler-Valsky, MPA & MMag. Dr. Verena Liszt-Rohlf*

3. Burgenländisches Zukunftssymposium

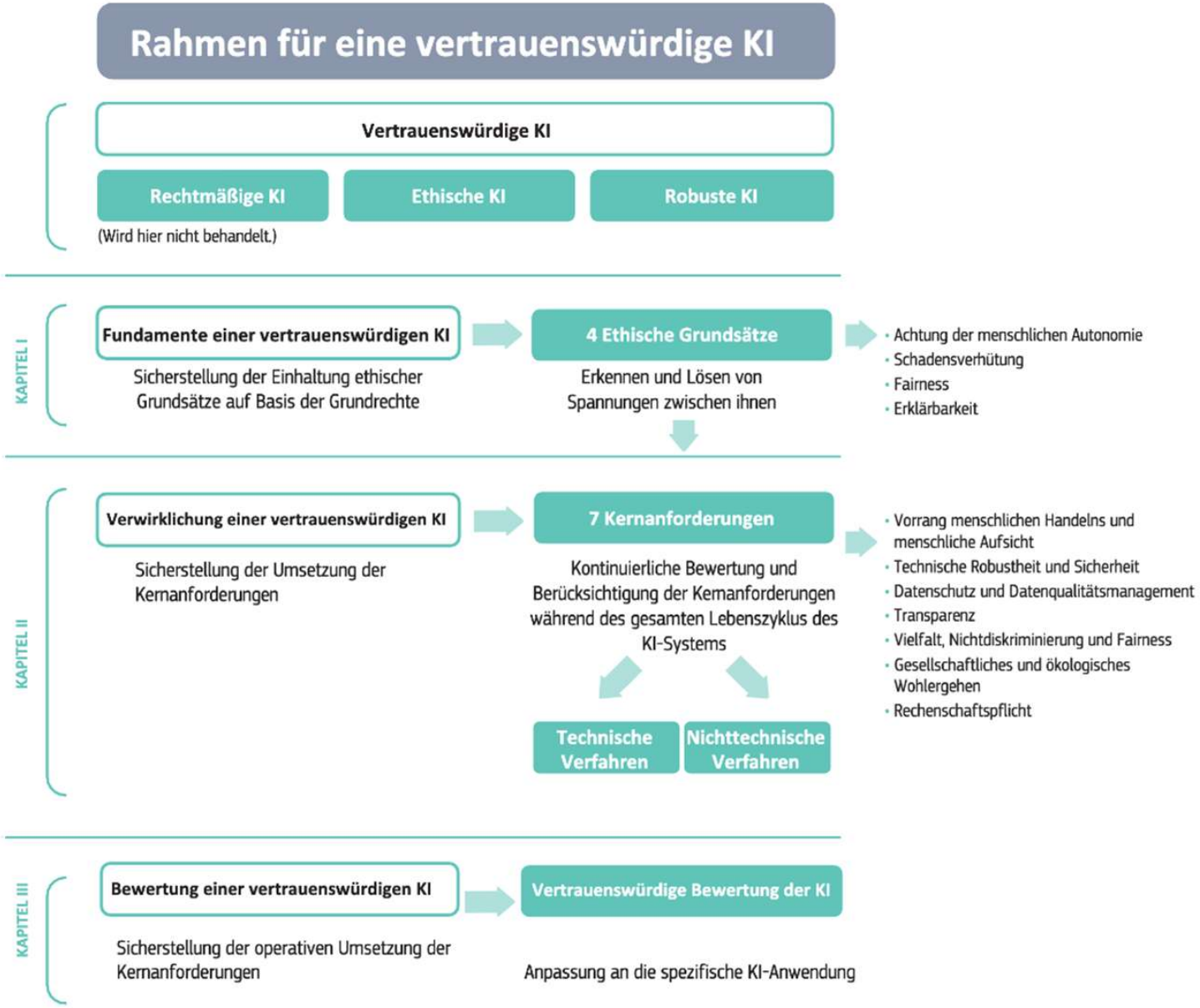
FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

hochschule burgenland
University of Applied Sciences

# Begriffsdefinition

KI =
*ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können.*
Art. 3 (1) AI-Act




FFG
Forschung wirkt.

hochschule burgenland
University of Applied Sciences

# Module für vertrauenswürdige KI

| KI-Technik | KI-Management | KI-Compliance | KI-Fairness |
|---|---|---|---|
| • Daten verstehen<br>• Modelle meistern<br>• Zuverlässige KI | • Entscheidungs-verantwortung<br>• Menschliche Kontrolle<br>• Prozess-verantwortung | • KI-VO<br>• DSGVO<br>• Zivilrecht | • Bias<br>• Diskriminierung<br>• Diversity & Inclusion |

Bisher: 100 Mitarbeiter:innen in KMUs aus den Bereichen (Projekt)Management, IT Softwareengineering, HR

Zeitaufwand: ~ 3h individuelle Vorbereitung (Podcast, Quizzes) + 8h-Präsenzworkshop + ~ xh Transfer mit Vorlagen

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



**FAIR = UNBIASED + INCLUSIVE**

- Bias = unbewusste Denkmuster bzw. Vorurteile, die zu einer systematische Abweichung von einer neutralen und "objektiven" Entscheidungsfindung führen
- Beispiele: Automation bias, extroversion bias, similarity bias etc.
- In Zusammenhang mit KI sind drei Arten von Biases relevant:

1. **Data Bias**
2. **Model (or Algorithm) Bias**
3. **User Bias**

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



## DATA BIAS

- Problem: unvollständige und/oder verzerrte Datensätze
- Garbage in, garbage out
- **Historical bias**
  - AMS Berufsinfomat 2024
  - Amazon HR AI 2018
  - Kreditvergabealgorithmen
  - Racial Profiling, Gender Medicine …
- Repräsentationsbias: WEIRD Sample

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



## MODEL BIAS

- Problem: Verzerrungen bei der Gestaltung von AI features
- Wer entwickelt AI? Und wer nicht?
- Wie interdisziplinär und inklusiv sind Teams?
- Beispiele:
  - Hautsensoren bei Händetrocknern
  - Menschen mit Behinderungen & Verkehrsplanung
  - Chatbots


FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



## USER BIAS

- Problem: Der Zugang zu und die Nutzung von KI
- Wer nutzt AI? Und wer nicht?
- Beispiele:
  - WEIRD, again
  - Bewertungsplattformen
  - Soziale Medien: Echo chambers und Polarisierung
- Wer interpretiert die Ergebnisse?
- Wer bestimmt über die Anwendung und Weiterentwicklung von AI?

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



## SCHRITT FÜR SCHRITT ZU BIAS-FREE AI

- Modell 1: World Economic Forum, Global Future Council 2022

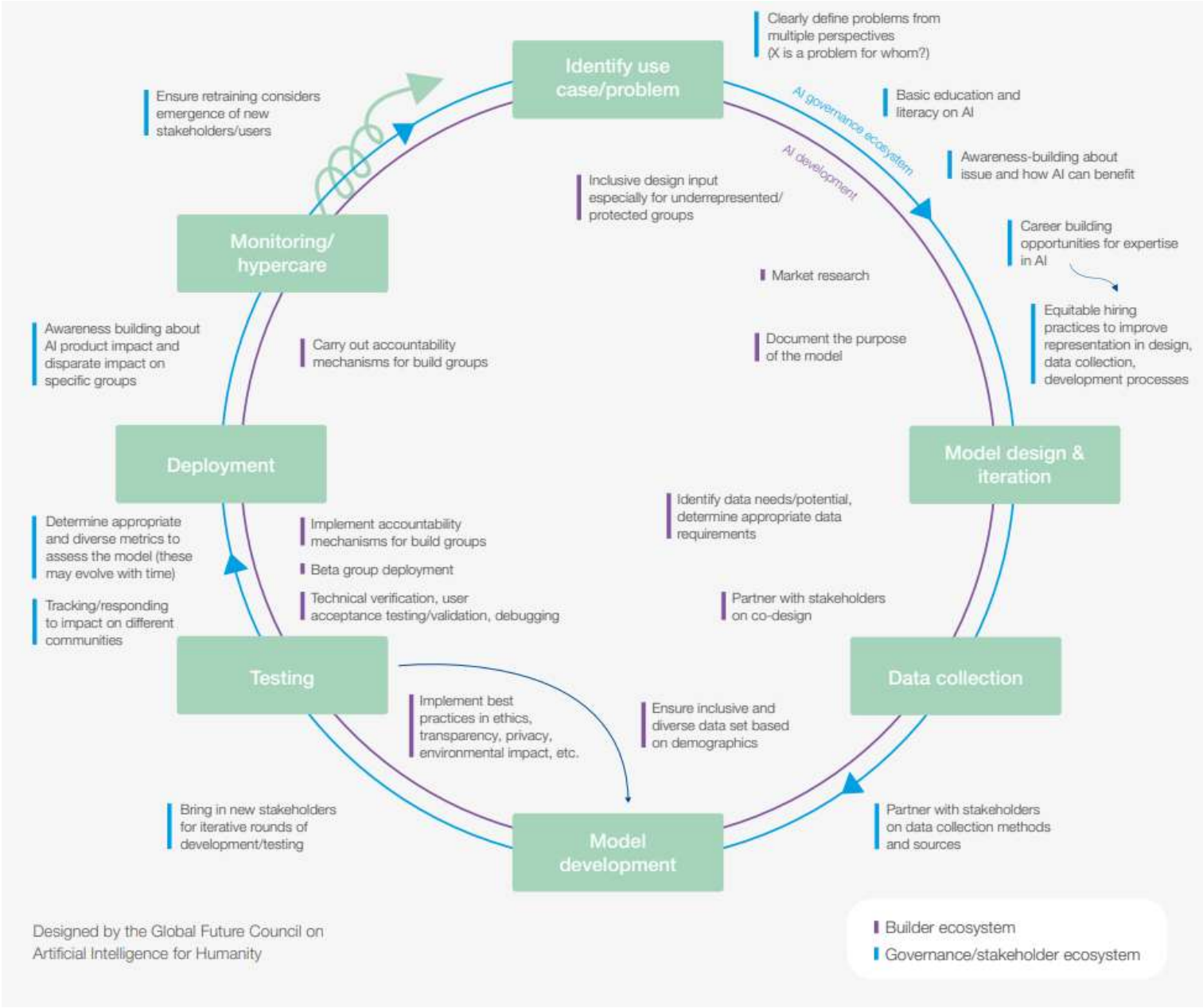
Ziel: 7 Schritte im AI-Entwicklungs- und Nutzungsprozess zu identifizieren und so zu gestalten, dass die Perspektiven aller Menschen Berücksichtigung finden

"A systems view of equality and inclusion in AI"

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

hochschule burgenland
University of Applied Sciences




FFG
Forschung wirkt.

# Modul 4: KI Fairness



## SCHRITT FÜR SCHRITT ZU BIAS-FREE AI

- Modell 2: Avellan et al. (2020)

Zurück zu den drei relevanten Dimensionen: Data, Algorithms, Users

AcademicMindtrek '20, January 29–30, 2020, Tampere, Finland

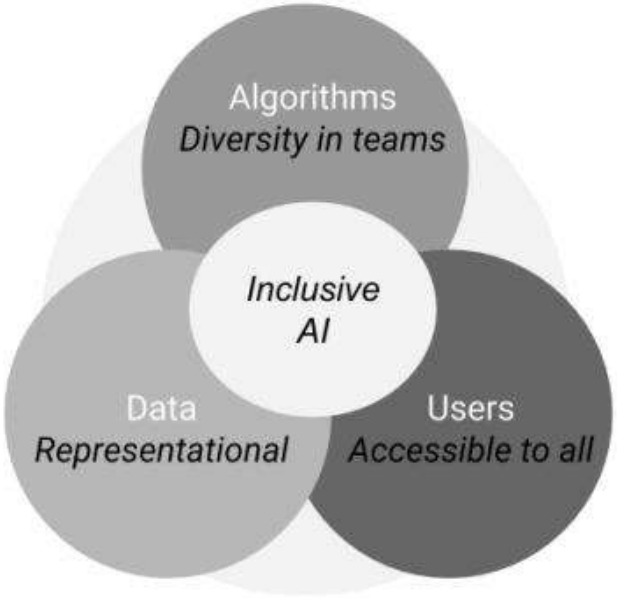



**Figure 1: The three components of Inclusive AI: diverse teams work on the algorithms, data used for training is representational of diverse users, and the application themselves are accessible to diverse user groups.**

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

# Modul 4: KI Fairness



## Nützliche Tools & Links für Expert:innen

- **Fairlearn:** Python-Bibliothek zur Analyse und Verbesserung der Fairness von KI-Modellen https://fairlearn.org
- **AI Fairness 360 (AIF360):** Umfassendes Open-Source-Toolkit von IBM mit über 70 Fairness-Metriken AI Fairness 360
- **Google What-If Tool:** Visualisierung zur Untersuchung von Modellverhalten bei Merkmalsänderungen https://pair-code.github.io/what-if-tool
- **EU Assessment List for Trustworthy AI:** https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

FFG Forschung wirkt.

*Diese Unterlagen sind im Rahmen des von der FFG geförderten Leitprojekts FairAI – „Fostering Austria's Innovative strength and Research excellence in Artificial Intelligence" entstanden.*

hochschule burgenland
University of Applied Sciences

# Bleiben wir im Gespräch...


## Mag.a Dr.in Julia Kreyler-Valsky, MPA

Lecturer, Kollegiumsmitglied

Dep. Wirtschaft

Campus Eisenstadt

julia.kreyler-valsky(at)hochschule-burgenland.at


Evidenzbasiertes Diversity Management und wissenschaftliche Methoden


## Dr.in Verena Liszt-Rohlf

Senior Researcher

Dep. Wirtschaft

Campus Eisenstadt

Verena.Liszt-Rohlf(at)hochschule-burgenland.at

+43 5 7705-4530


Entrepreneurship Education and Sustainable Business Models