Organisiert von:

PARIS LODRON UNIVERSITÄT SALZBURG

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

MENSCHZENTRIERTE KI-FORSCHUNG

Vorstellung vier inter- und transdisziplinärer Forschungsprojekte mit dem Fokus auf die Fairness von künstlicher Intelligenz

15.06.2021

TRANSPARENZ INTELLIGENTER WAHRNEHMUNGSSYSTEME

FairAlgos

Martin Kampel | TU Wien

Dieses Projekt wurde gefördert von FFG, Bridge - 878730

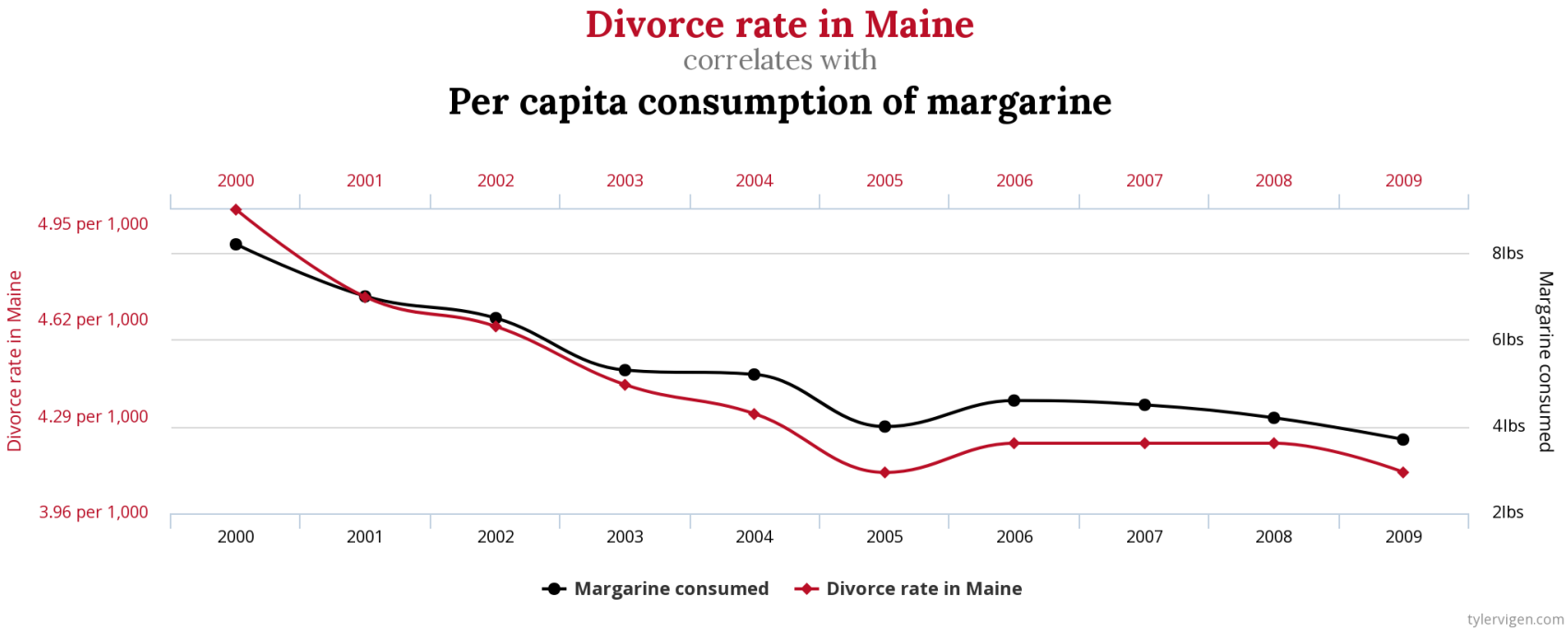
FAIRALGOS

FairAlgos: **Fair**ness, Bias and Transparency in Computer Vision **Algo**rithm**s**

- 2.5 Jahre Projektdauer, Start Mai 2020
- Faire Algorithmen in Visual Computing

- **Technik**: Computer Vision Lab, TU Wien
- **Unternehmen**: cogvis GmbH
- **Soziologie**: Wiener Zentrum für sozialwissenschaftliche Sicherheitsforschung

ALGOCARE – … LEARNING FROM DATA ?



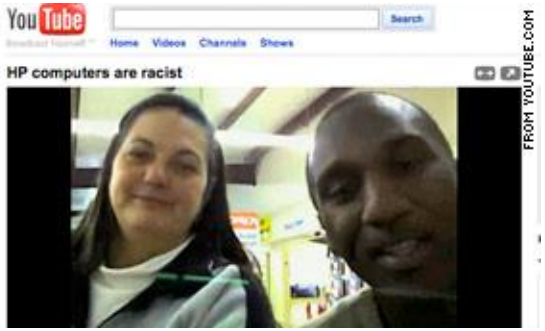
Quelle: http://www.tylervigen.com/

FAIRALGOS

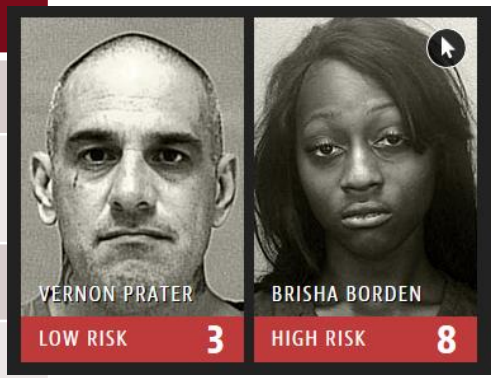
Computer Vision als „**Driving Application der Künstlichen Intelligenz**“

Zahlreiche **Beispiele** für unfaire, intransparente oder rassistische Algorithmen:

- Asiatische Gesichter wurden nicht korrekt detektiert
  Nikon Coolpix S630: „Did someone blink?“
- HP Face Tracking Software
  Kein Tracking oder Detektion von „black faces“
- Northpoint predicting crimes



	Vernon Prater	Brisha Borden
Arrest reason	Depth	Depth
Criminal record	• 3 armed robberies • 1 attempted armed robbery	4 juvenile justice data
Subsequent offenses	1 grand depth	-
Bewertung	HIGH RISC: 3	HIGH RISC: 8



FAIRALGOS – XAI IN COMPUTER VISON

- Deep Learning kommt häufig für die Klassifizierung von Bildern zum Einsatz
- dabei werden große Datenbanken verwendet

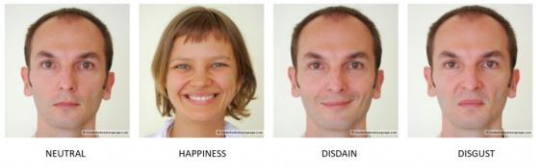
- **Bias in Daten führt zu Bias in Modellen → Potential zur Diskriminierung**

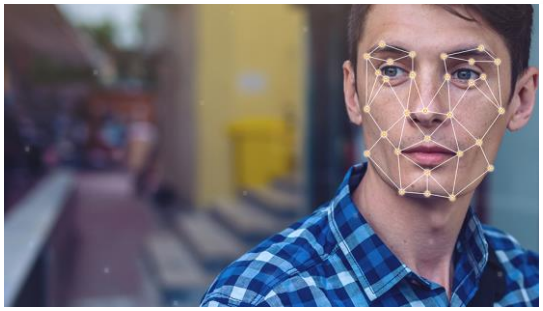
- wird z.B. ein Modell hauptsächlich mit Personen mit heller Hautfarbe trainiert, kann das zu schlechten Ergebnissen bei Personen anderer Herkunft führen.

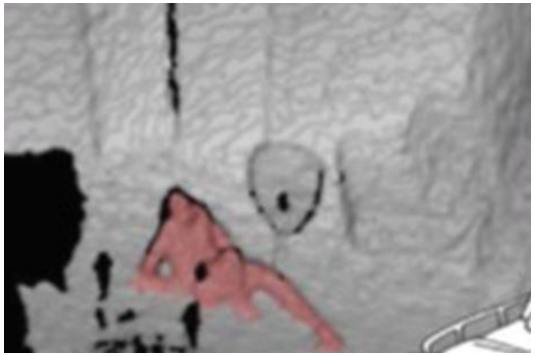
- 2015: Google Photos hat ein Foto von einem Mann mit dunkler Hautfarbe und seine Freunde mit dem Tag „Gorillas“ markiert.

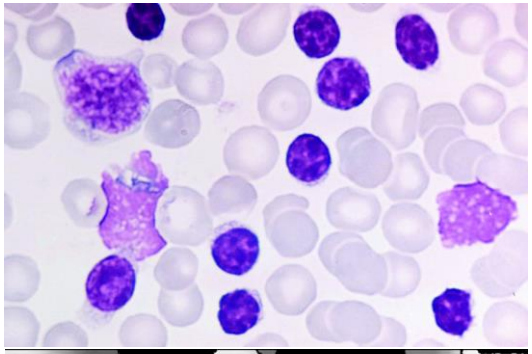
FAIRALGOS – NEEDED IN …

- AI basierte Auswertungen von medizinische Bilddaten (Diagnostik/Therapieplanung, …)
- Gesichtserkennung (Zutrittssysteme bis Anwendungen am Handy)
- Verhaltensanalyse (Surveillance, Active Assisted Living, etc.)
- Bildbasierte Erkennung von Geschlecht / Alter / ethnischer Zugehörigkeit / Analyse von Gesichtsausdrücken / Emotionen, …

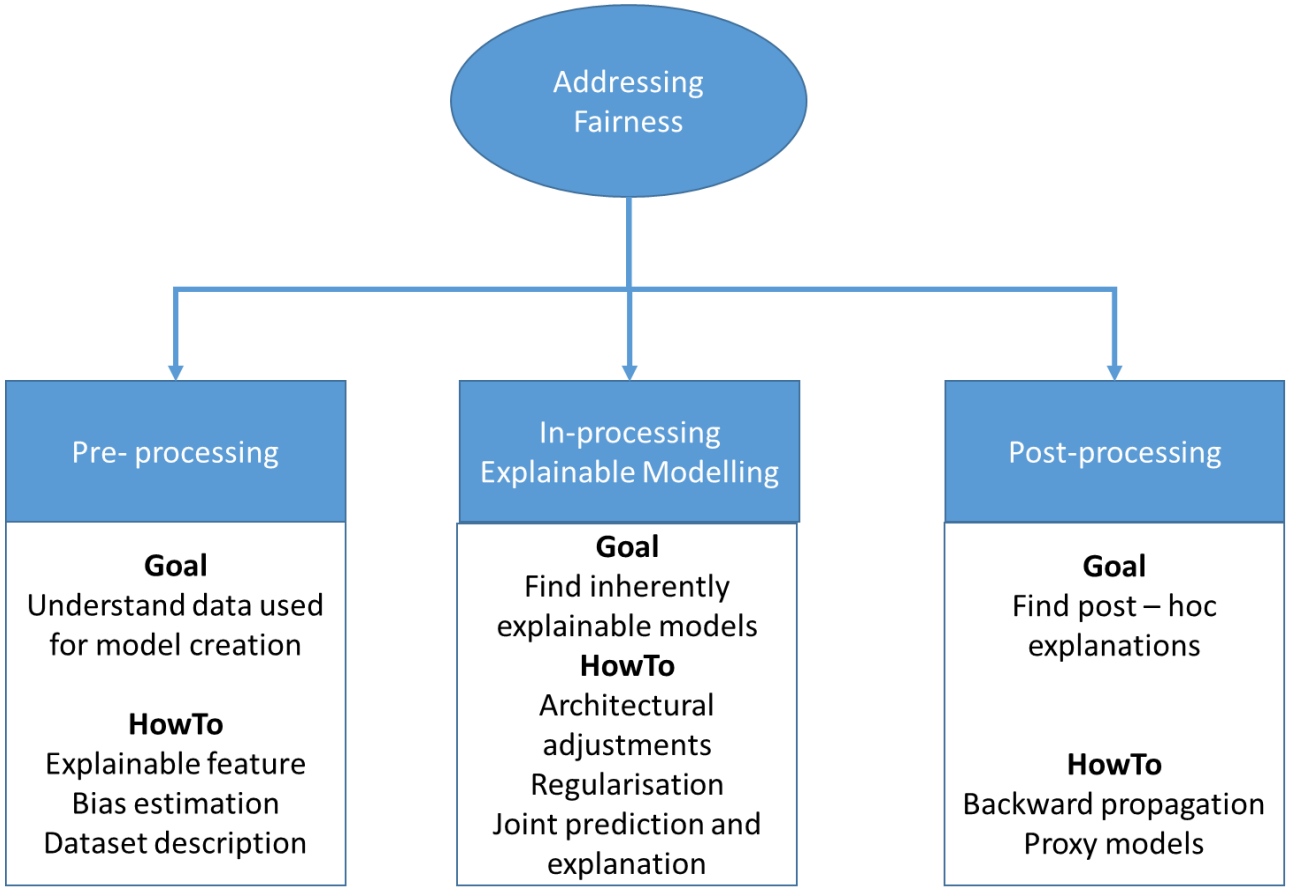








ALGOCARE – ASSESSING



(Mehrabi, N. et al. (2019): A Survey on Bias and Fairness in Machine Learning, ArXiv, abs/1908.09635)

FAIRALGOS - ZIELE

- Entscheidungstreffende Systeme und ihre Konsequenzen hinsichtlich Fairness und Transparenz **untersuchen**
- Eigenschaften eines Systems identifizieren, die zu **Transparenz** und **Fairness** führen
- **Designkriterien** definieren
- Entwickeln von **Fairness Guidelines** für Entwickler und User
- **Konkrete UseCases** analysieren und als Basis zur Entwicklung der Fairness Guidelines heranziehen

AUSBLICK: ALGORITHMIC GOVERNANCE OF CARE

- Algorithmen im Bereich der Pflege und deren Auswirkungen auf PflegerInnen und Pflegebeziehenden **verstehen**
- Konkrete UseCases
  - Pflegeroboter
  - Wearables (vom Schrittzähler bis zur Blutzuckermessung)
  - Big (Health) Data Analysis

- Projektstart 12/2021 f. 3 Jahre

CVL Computer Vision Lab

WWTF WIENER WISSENSCHAFTS-, FORSCHUNGS- UND TECHNOLOGIEFONDS