# Community of Practice KI*PerWeb*

Austausch zur Nutzung und Entwicklung KI-gestützter Webanwendungen

KIPERWEB

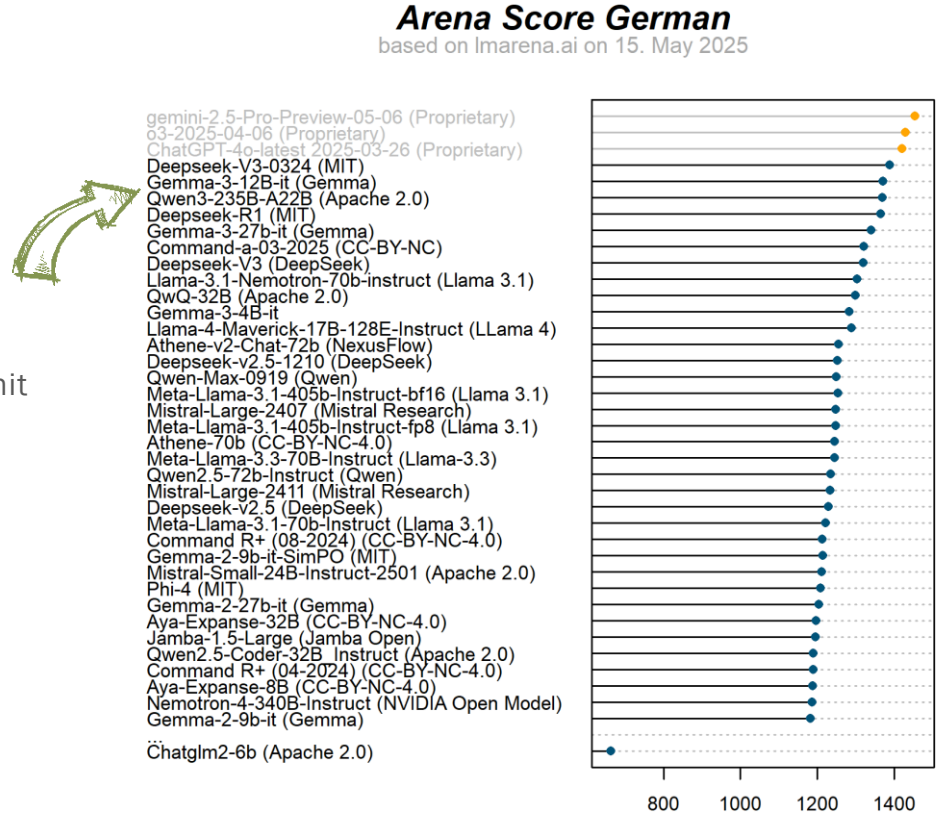
f-bb Forschungsinstitut Betriebliche Bildung

# Agenda

- Update
  - News & Leaderboard-Update
- Input
  - „Bias und Fairness - KI zwischen Neutralität und Vorurteil“
- Diskussion

# Leaderboard-Update (15.05.2025)

- Die aktuellen proprietären Top-Modelle von OpenAI, Google und X (rechts in grau) können ihren Vorsprung halten, allerdings gibt es keinen sprunghaften Anstieg mehr;
- Arena-Scores von *nicht*-proprietären Modelle sind rechts ausgewiesen sofern sie mindestens das Niveau von Gemma-2-9b-it erreichen:
  - Gemma-3-Familie nun komplett im Leaderboard (mit Überraschungssieger Gemma-3-12B-it)
  - Gemma-3-4B-it liegt vor Schwergewicht Llama-4-Maverick!





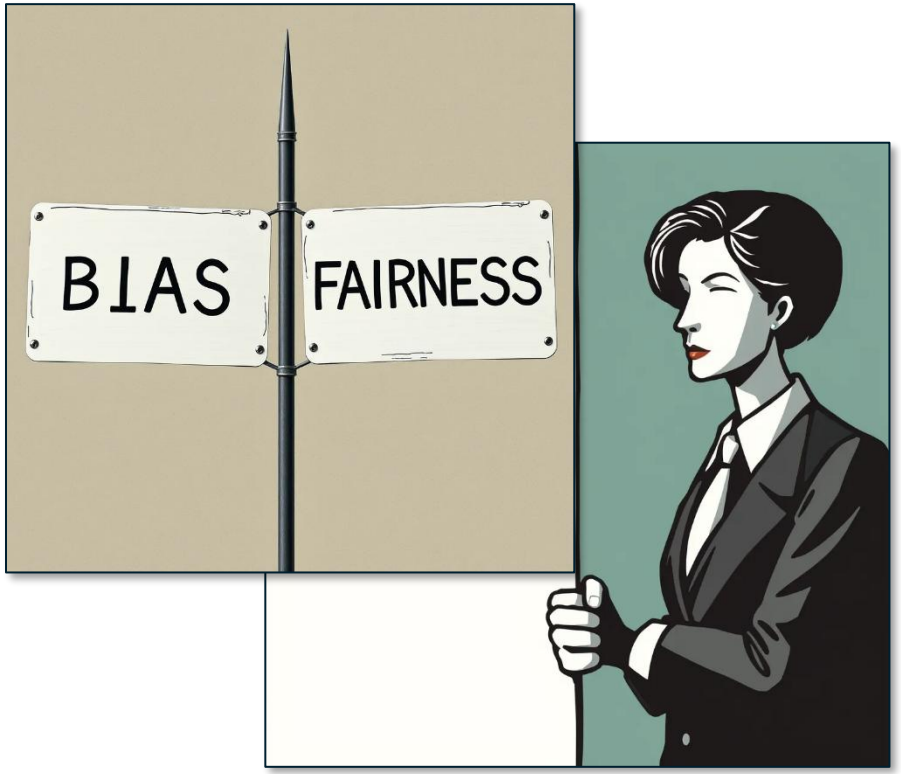

Datengrundlage: https://chat.lmsys.org/?leaderboard

# Fokusthema: Bias und Fairness

- Prompt:
  „Bias und Fairness - KI zwischen Neutralität und Vorurteil“

(rechts visualisiert von FLUX.1-schnell)

...






Quelle: https://huggingface.co/spaces/black-forest-labs/FLUX.1-schnell

# Bias und Fairness - Arbeitsdefinitionen

- **Bias** meint idR. eine systematische Fehleinschätzung/Voreingenommenheit im Sinne einer verzerrten Beurteilung/Wahrnehmung
  - z.B. Kahnemann (2021): *„Es gibt Schwankungen und diese Schwankungen werden als Noise [i.S.v. Rauschen] bezeichnet. Und der durchschnittliche Fehler ist der Bias [i.S.v. Verzerrung], also die kognitive Verzerrung.“*

- **Fairness** bezeichnet in diesem Zusammenhang ein „gerechtes“ Verhalten ohne Bias



Kahnemann, D. (2021). Noise. Was unsere Entscheidungen verzerrt – und wie wir sie verbessern können

# Bias-Beispiele

- Teilweise bezeichnet man Biases nach dem Merkmal, bzgl. dessen eine systematische Fehleinschätzung erfolgt (z.B. Gender-Bias), teilweise auch nach der Ursache, z.B.
  - Historical Bias (i.S.v. unerwünschte Reproduktion bestehender Ungleichheiten und Vorurteile auf Basis veralteter oder problematischer Trainingsdaten),
  - Sampling Bias (i.S.v. verzerrte Urteile aufgrund nicht-repräsentativer Trainingsdaten)
  - Aggregation Bias (i.S.v. voreiligen Verallgemeinerungen über unterschiedliche Teilmengen)
  - Modellierungsbias (i.S.v. Auswahl verzerrender Algorithmen, Features oder Metriken)
  - …

# „American Smile“ und andere Vorurteile von KI

“Worth noting that these photos [on the right] themselves were taken by photographers like Edward Curtis and others whose own colonizing point of view influenced how Native peoples were represented, so these images too come with an imposed, outside perspective.” – jenka, 2023



**Quelle:** https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf

# KI & Fairness: Wer bekommt den Job, wer den Hit?

- Einschlägige Podcast-Folge aus Autonomie & Algorithmen, u.a. zu „Dynamic Fairness“ (Paaßen et al., 2019)
- individuelle Fairness (i.S.v. ähnliche Ergebnisse bei ähnlichen hinterlegten Eigenschaften) vs. Gruppenfairness (i.S.v. ähnliche Verteilungen in unterschiedlichen Gruppen)
  - Demographic parity (equal rates of positives across groups)
  - Equalized odds (equal rates of positives across groups for (a) true positives and (b) true negatives)
  - Due process: don‘t code features of group (in-/directly)

Dynamic fairness – Breaking vicious cycles in automatic decision making

Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang

CITEC Center of Excellence, Bielefeld University*

Preprint of the ESANN 2019 paper [Paaßen et al. 2019] as provided by the authors. The original can be found at the ESANN electronics proceedings page.



**Abstract**

In recent years, machine learning techniques have been increasingly applied in sensitive decision making processes, raising fairness concerns. Past research has shown that machine learning may reproduce and even exacerbate human bias due to biased training data or flawed model assumptions, and thus may lead to discriminatory actions. To counteract such biased models, researchers have proposed multiple mathematical definitions of fairness according to which classifiers can be optimized. However, it has also been shown that the outcomes generated by some fairness notions may be unsatisfactory.

In this contribution, we add to this research by considering decision making processes in time. We establish a theoretic model in which even perfectly accurate classifiers which adhere to almost all common fairness definitions lead to stable long-term inequalities due to vicious cycles. Only demographic parity, which enforces equal rates of positive decisions across groups, avoids these effects and establishes a virtuous cycle, which leads to perfectly accurate and fair classification in the long term.

Automatic decision-making via machine learning classifiers carries the promise of quicker, more accurate, and more objective decisions because automatic mechanisms do not foster animosity against any group [Munoz et al., 2016, O'Neil, 2016]. Yet, machine learning systems can indeed reproduce and exacerbate bias that is encoded in the training data or in flawed model assumptions [Munoz et al., 2016, O'Neil, 2016, Corbett-Davies and Goel, 2018, Dwork et al., 2012]. For example, the COMPAS tool, which estimates the risk of recidivism of defendants in the US law system prior to trial, has been found to have higher rates of false positives for Black people compared to white people and has thus been called unfair [Angwin et al., 2016]. Similarly, a tool developed by Amazon to rate the résumés of job applicants assigned higher scores to men compared to women because successful applicants in the past had mostly been male [Dastin, 2018]. Finally, multiple machine-learning-based credit scoring systems have emerged that reproduce historical biases and systematically assign lower credit scores to members of disenfranchised minorities [O'Dwyer, 2018].

In general, we consider scenarios where individuals $i \in \{1, \ldots, m\}$ in some population of size $m$ apply for some positive outcome, such as a pre-trial bail, a job, or a loan, and a gatekeeper institution decides whether to grant that outcome, with the interest of accepting only those individuals who will "succeed" with that outcome, e.g. not commit a crime, succeed

*Funding by the CITEC center of excellence (EXC 277) is gratefully acknowledged.

1

Quellen: https://www.podcast.de/episode/679790877/ki-fairness-wer-bekommt-den-job-wer-den-hit
https://arxiv.org/pdf/1902.00375

# Gender-Bias in Sentence Embeddings zur BO

- AMS listet unter 2131 z.B. „MedizinerIn“ vs. 221 „Ärzte“ und allgemein finden sich in ESCO und ISCO unterschiedliche Gender-Schreibweisen sowie Singular- und Pluralformulierungen
- Cosine Similarity der Sentence Embeddings von Berufsbezeichnungen teils sensitiv für Gender-Schreibweisen:
  - Laut Jina-v2-embeddings-de ist „Arzt“ der Kategorie „Ärzte“ ähnlicher als der Kategorie „MedizinerIn“ – bei „Ärztin“ verhält es sich umgekehrt
- Quantifizieren lässt sich diese Art von Gender-Bias beispielsweise durch den Betrag der Differenz zwischen den Ähnlichkeiten zu den Extrem-Kategorien „männlich“ einerseits und „weiblich“ andererseits

# Mitigating Bias (Palomino et al., 2025)

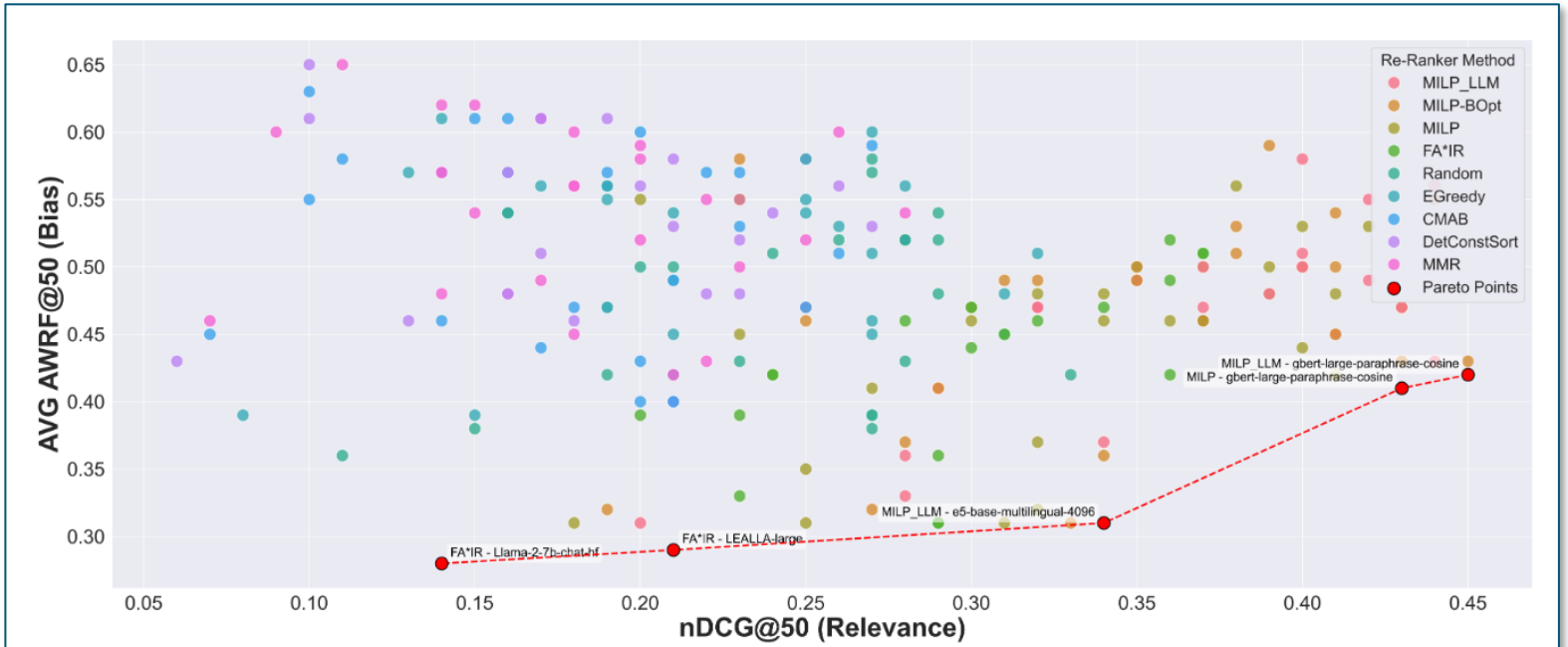



Figure 2: Comparing re-ranking methods: achieving optimal balance between relevance (nDCG) and bias (Avg. AWRF) trade-offs.

Quelle: https://aclanthology.org/2025.naacl-industry.16.pdf

# Diskussion

- Welche Bezugspunkte seht ihr zum Thema Bias und Fairness (für euch oder die Personen auf die euer Verhalten und eure KI-Systeme Einfluss haben)?
- Kennt ihr weitere Aspekte die man kennen und berücksichtigen sollte?
- Welche Implikationen hat das Thema für Eure Arbeit?