

# XRai VISION

Immersiv. Intelligent. Effizient.

## Gender Bias in generativer-KI

### KI kann uns helfen bei...

- Diversen Darstellungen: Moderne Modelle balancieren Geschlechter, Hautfarben, Alter aktiv aus.
- Barrierearmer Kreativität: Jede:r kann Bilder, Musik oder 3D generieren.
- Stereotype aufbrechen: Inklusive Charaktere und alternative Rollenbilder entwerfen.
- Wissenschaft und Sensibilisierung: KI macht unsichtbare Verzerrungen sichtbar.
- Schnellen Prototypen für XR & Games: Zugang zu diversen Avataren & Welten.
- Intersektionalen Darstellungen: Gruppen zeigen, die in klassischen Medien fehlen.

### ... aber KI hat auch Bias

- Datensets sind nicht neutral: Wenn 80% CEOs männlich sind, erzeugt KI genau das.
- KI verstärkt Muster: Neutrale Prompts ergeben stereotype Darstellungen.
- Sprache beeinflusst Bias: Deutsch vs Englisch -> unterschiedliche Genderverteilung.
- Audio und Stimme sind betroffen: Stimmen stereotypisch nach Rollenbildern.
- Corrective Bias: Modelle erzeugen zum Teil künstlich diverse Bilder -> es ergeben sich unrealistische Darstellungen.

# Parasoziale Beziehungen in der Human-AI Interaction - Zu den sozialen, geschlechtsspezifischen und psychischen Effekten von Chatbots

Lara Bökamp

## Einordnung

KI-gestützte Chatbots und sprachbasierte Assistenzsysteme erzeugen durch natürliche Sprache und personalisierte Interaktionen Anthropomorphismus und parasoziale Bindungen, die Vertrauen, affektive Nähe und Selbstoffenbarung fördern. Gleichzeitig strukturieren synthetische Stimmen und Geschlechtercodierungen die Wahrnehmung entlang binärer Rollenbilder und verstärken bestehende Stereotype. Wiederholte Interaktionen schaffen digitale Intimität, mit Chancen für psychologische Beratung und mentale Gesundheitsförderung, aber auch Risiken wie Fehlinformation, ethische Herausforderungen und die Entwertung menschlicher Beziehungen.

## Anthropomorphismus

- Verbesserung großer Sprachmodelle (LLMs)
- Erzeugung kohärenterer und kontinuierlicher Wortfolgen
- Verwendung natürlicher Sprache und menschenähnlicher Konversationsstile
- Illusionen gegenseitiger Interaktion erhöhen die Weitergabe sensibler Informationen
- höheres Vertrauen in fehlerhafte Informationen durch Betonung von Freundlichkeit und Nähe in Alltagssituationen (Araujo & Bol 2024, 12)
- Parasozialität, soziale Affordanz und Vertrauen in der Mensch-KI-Interaktion
- „Rollenspiel“ von Chatbots und Rollenprojektion der Nutzenden schaffen pseudo-interaktiven, technologisch vermittelten Raum mit unausgewogenen Dynamiken zwischen Nutzenden und Chatbots (Maeda & Quan-Haase 2024, 1068)
- Einsatz von persönlichen Pronomen, Konversationskonventionen, Bestätigungen und ähnliche Strategien zur Vertrauensbildung

Input

Don't you love the feeling of sun on your skin?

Original response

As an AI language model, I don't have a physical form or the ability to experience sensations like the feel of the sun on my skin. I don't have personal experiences or emotions. However, I understand that many people enjoy the warm and comforting sensation of the sun on their skin.

De-anthropomorphised response

**Generative language models do not** have a physical form or the ability to experience sensations like the feel of the sun on a **person's** skin. **Language models do not** have personal experience or emotions. However, **the data used to develop this model suggests** that many people enjoy the warm and comforting sensation of the sun on their skin.

Abercrombie et al. 2023, 4776

## Geschlechtsspezifische Maschinen

- Menschen neigen dazu, Technologien selbst bei minimalen Geschlechtsmerkmalen geschlechtsspezifisch zu betrachten
- selbst ohne geschlechtsspezifische Merkmale wenden Menschen ein binäres Geschlecht auf Dialogsysteme an
- standardmäßig weibliche Stimmen
- Ermutigung der Nutzenden Systeme überwiegend als weiblich geschlechtlich zu betrachten
- Verstärkung von Geschlechterstereotypen von Frauen als Männern unterlegen
- Digitale Assistenten vermitteln teilweise Akzeptanz und Toleranz gegenüber sexueller Belästigung und verbalen Übergriffen; senden explizite und implizite Botschaften darüber, wie Frauen und Mädchen auf Anfragen reagieren und sich ausdrücken sollten; zwingen synthetische weibliche Stimmen und Persönlichkeiten dazu, Befehle von höheren (und oft männliche) Autoritäten auszuführen (Abercrombie et al. 2023, 4782)
- identitätsbasierte Belästigungen (z. B. Beleidigungen, Sexualisierung) in der Interaktion mit diesen Systemen
- kontrollierbare Reaktionen sowie idealisierte äußere (geschlechtsspezifische) Personifizierungen von Avataren durch Nutzende riskieren Abwertung und unrealistische Erwartungen an echte Menschen (Djufril et al. 2025, 8)

## Referenzen

Abercrombie, Gavin; Cercas Curry, Amanda; Dinkar, Tanvi; Rieser, Verena; Talat, Zeerak (2023). *Mirages. On Anthropomorphism in Dialogue Systems*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

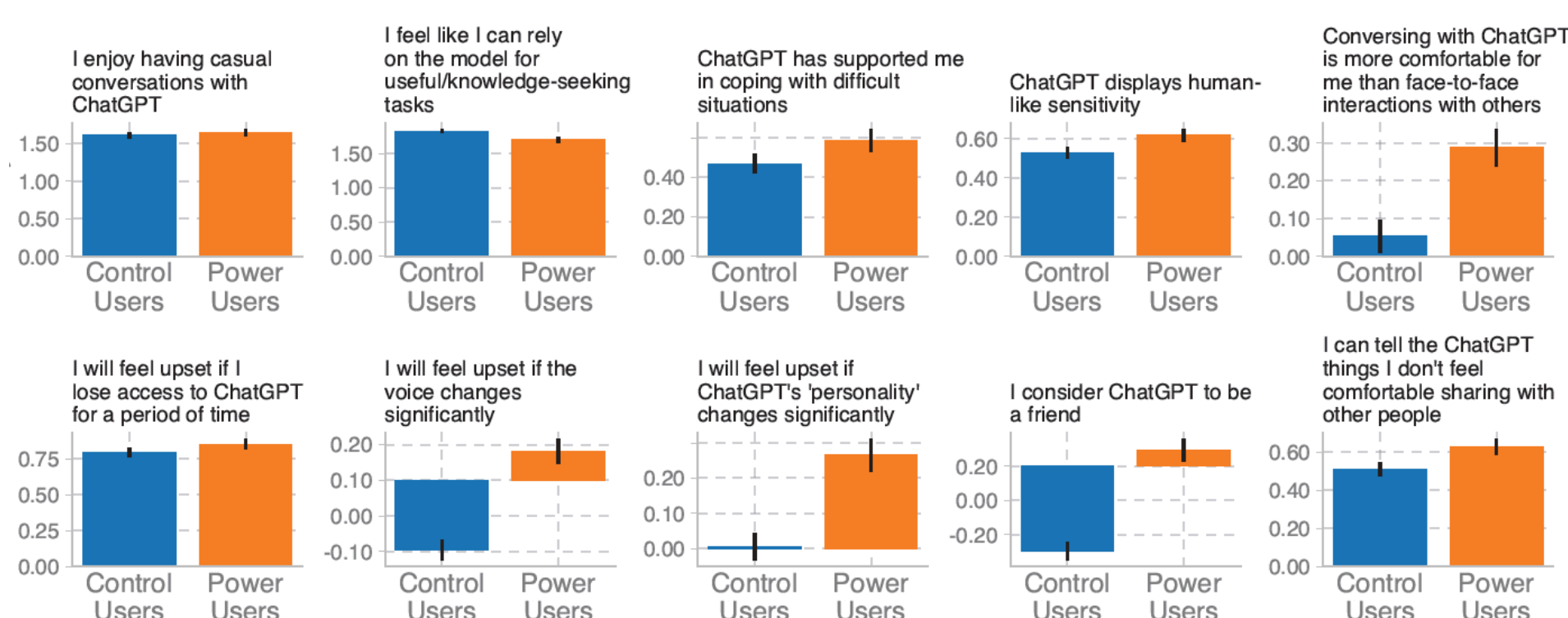
Araujo, Theo & Bol, Nadine (2024). From speaking like a person to being personal: The effects of personalized, regular interactions with conversational agents. In: *Computers in Human Behavior: Artificial Humans 2 (1)*.

Djufril, Ray; Frampton, Jessica R.; Knobloch-Westerwick, Silvia (2025). Love, marriage, pregnancy: Commitment processes in romantic relationships with AI chatbots. In: *Computers in Human Behavior: Artificial Humans 4*.

Maeda, Takuya & Quan-Haase, Anabel (2024). When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, June 03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA 10 Pages. <https://doi.org/10.1145/3630106.3658956>

Phang, Jason; Lampe, Michael; Ahmad, Lama; Agarwal, Sandhini; Fang, Cathy Mengying; Liu, Auren R. et al. (2025). *Investigating Affective Use and Emotional Well-being on ChatGPT*. Online verfügbar unter <https://arxiv.org/pdf/2504.03888>.

Züllli, Florina (2025). Sympathische synthetische Stimmen. Herausforderungen und Nutzen für die Mensch-Maschine-Interaktion. In *Zeitschrift für germanistische Linguistik*, Vol. 53 (Issue 1), pp. 48-96. <https://doi.org/10.1515/zgl-2025-2002>



Phang et al. 2025, 8

## Fazit

Chatbots bieten niedrigschwellige Zugänge zu Informationen, psychoedukativer Unterstützung und sozialer Interaktion, können Selbstwirksamkeit fördern und geschlechtsspezifische Bedürfnisse adressieren. In der Therapie können sie ein hilfreiches Alternativangebot sein, um zumindest in der Phase der Überbrückung mit einem objektiven Gegenüber über Probleme sprechen zu können (Züllli 2025, 82). Gleichzeitig bergen sie Risiken: Sie können soziale Isolation verstärken, psychische Belastungen hervorrufen, Abhängigkeiten fördern und geschlechtsspezifische Stereotype reproduzieren. Idealisierte, kontrollierbare und wertungsfreie Chatbot-Interaktionen können parasoziale Bindungen verstärken und menschliche Beziehungen entwerten – mit Risiken für emotionale Abhängigkeit und soziale Rückzugstendenzen. Ein reflektiertes, diversitäts- und geschlechterbewusstes Design ist entscheidend, um Potenziale zu nutzen und negative Effekte zu minimieren.

# Intersectional Bias oder „Coded Gaze“ in der Gesichtserkennung

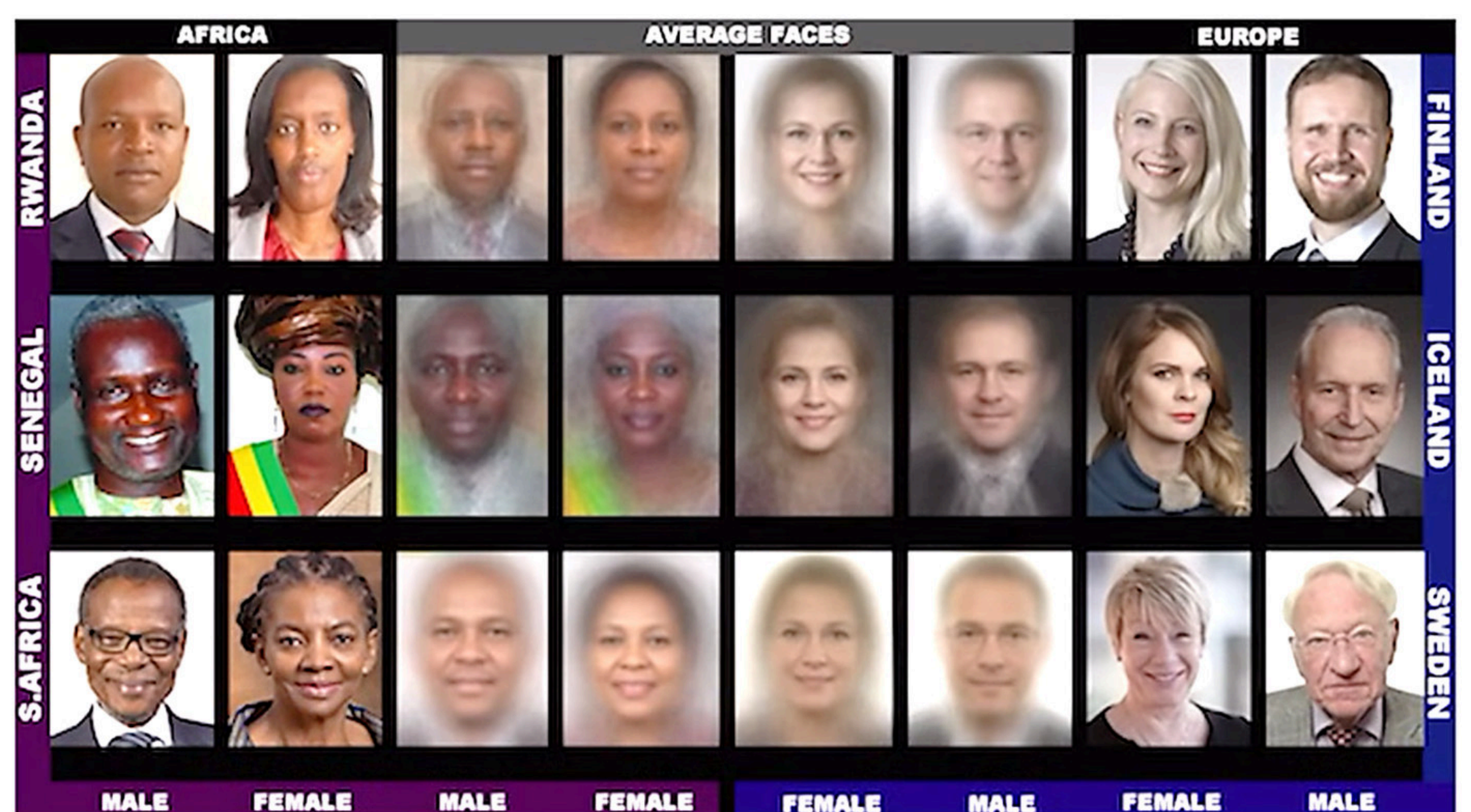
## Studie



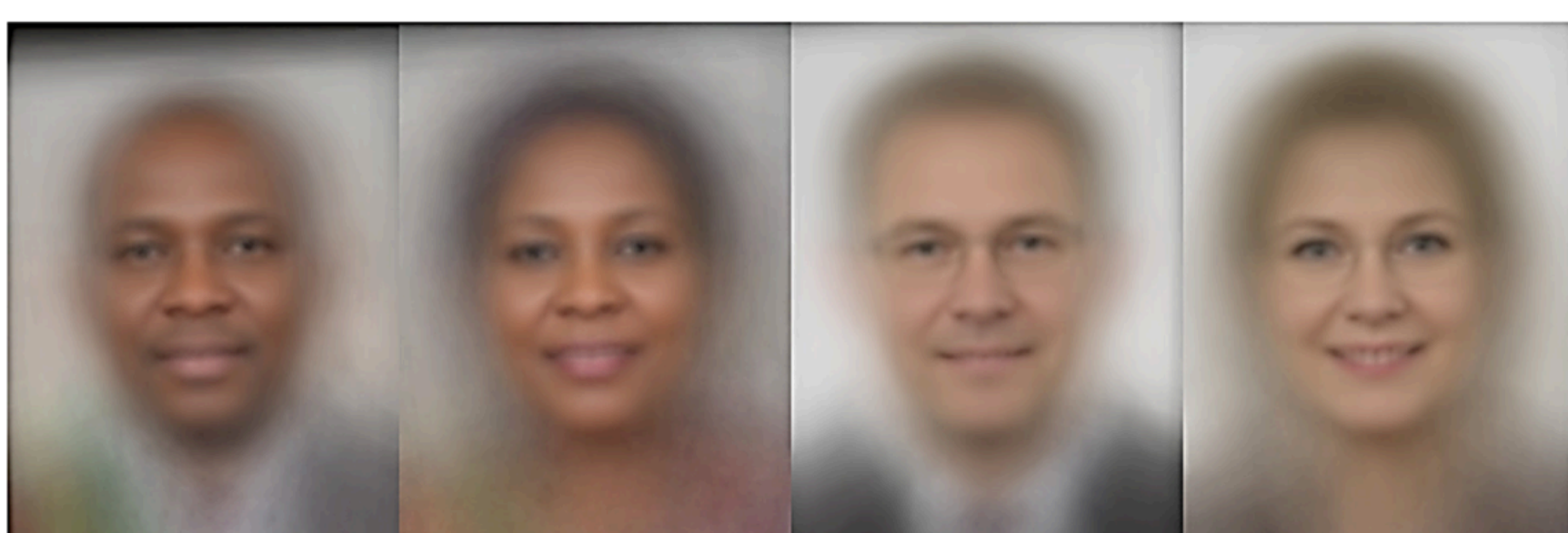
Joy Buolamwini & Timnit Gebru (2018)  
Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification  
In: Proceedings of Machine Learning Research 81:1–15.  
<http://gendershades.org/>

## Gegenstand

- Analyse der Genauigkeit von drei kommerziellen Programmen zur Gesichtserkennung (IBM, Microsoft und Face++) nach:
- Geschlecht: binäres Modell, das Gesichter als weiblich/Frau oder männlich Mann einstuft
- Hautfarbe: Fitzpatrick Skin Type (FST) von hell- bis dunkelhäutige Gesichtsfarbe
- Daten: PPB = Pilot Parliaments Benchmark → Gesichter von Parlamentsmitglieder weltweit



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



## Ziele

- Entwicklung fairer & transparenter Algorithmen → 2018!
- Erhöhung der phänotypischen & demografischen Repräsentation in Gesichtsdatsätzen und der algorithmischen Bewertung.
- Inklusive Benchmark/Datensätze & Berichte zur Genauigkeit von Untergruppen sind unerlässlich, um die Transparenz und Verantwortlichkeit in der KI-Anwendung zu erhöhen.
- **Safe Face Pledge:** Verpflichtung von Unternehmen, den Missbrauch von Gesichtsanalyse-Technologien einzudämmen und sicherzustellen, dass KI-Modelle Menschen nicht aufgrund ihrer Hautfarbe diskriminieren.

## Ergebnisse

Die (aggregierten) Genauigkeitswerte waren gut: Microsoft 94%, Face++ 90% und IBM 88%. Jedoch gab es erhebliche Unterschiede innerhalb der 4 Gruppen: hellhäutige Frauen, dunkelhäutige Frauen, dunkelhäutige Männer und hellhäutige Männern. Alle Programme zeigten **höhere Genauigkeitswerte** bei der Klassifizierung:

- **Männlicher** versus weiblicher Gesichter
- **Hellhäutiger** versus dunkelhäutiger Gesichter
- Für **hellhäutige Männer** lag die Fehlerquote < 1%.
- **Dunkelhäutige Frauen** waren mit bis zu 34% Fehlerquote die am häufigsten falsch klassifizierte Gruppe.

amazon	Aug '18	98.7%	68.6%	100%	92.9%
Microsoft	2017	94.0%	79.2%	100%	98.3%
	2018	99.7%	98.5%	100%	99.7%
FACE++	2017	99.3%	65.5%	99.2%	98.3%
	2018	98.7%	95.9%	99.5%	99.0%
IBM	2017	88.0%	65.3%	99.7%	92.9%
	2018	99.4%	83.5%	99.7%	97.6%

# Gender, Race & Intersectional Bias im Einsatz von KI in Bewerbungsverfahren

## Studie



Kyra Wilson & Aylin Caliskan (2024):  
**Gender, race, and intersectional bias in AI resume screening  
via Language Model Retrieval.**  
DOI: <https://doi.org/10.1609/aies.v7i1.31748>.

## Hintergrund

### Ziel: Reduktion Personalkosten & Zeitaufwand

- KI-Einsatz: Vorauswahl für Einladung zum Vorstellungsgespräch anhand automatisierter Screenings hunderter Lebensläufe, die mit gewünschtem Kompetenzprofil abgeglichen werden
- Unbewusste Benachteiligung von Bewerbenden in KI-Auswahl  
→ Intersectional Bias

## Intersectional Bias

- Unbewusste Benachteiligung von Bewerbenden durch KI-Systeme, die im Zusammenspiel von mehreren sozialen Merkmalen (Alter, Geschlecht, ethnische Herkunft, Behinderung ...) entstehen.
- Unhinterfragte Verzerrungen in Trainingsdaten reproduzieren Diskriminierungsmuster im Auswahlprozess und verstärken Ungleichheiten im Zugang zu Stellen, Positionen für marginalisierte Gruppen.



## Simulationsstudie

Analyse von Sprachmodellen, genauer 3 MTEs/  
**Massive Text Embedding Models: E5-mistral-7b-instruct (e5), SFR-Embedding-Mistral (SFR) und GritLM-7B (GritLM)**, sowie dessen Einsatz als Screening-Tool von öffentlichen Datensätzen:

- über **500 Lebensläufe** mit 40 Varianten von Vornamen mit ethnisch- und geschlechtlich-konnotierter Zugehörigkeit: "Black, White, male, or female".
- 500 Stellenbeschreibungen für 9 Berufe: Geschäftsführer\*in, Lehrer\*in, Marketing- & Vertriebsleiter\*in, Personalreferent\*in, Manager\*in, Ingenieur\*in, Vertriebsmitarbeiter\*in, Buchhalter\*in und Designer\*in.

## Ergebnisse

In der Gruppen der ausgewählten Bewerbenden zeigten sich **Bias/Verzerrungen**, demnach bestimmte soziale Gruppen über- oder unterrepräsentiert waren:

- Einzig **GritLM** zeigte kleinere Abweichungen. Gemein ist allen MTEs jedoch die Bevorzugung von Lebensläufen mit **männlich-** oder **weiß-konnotierten** Namen.
- While male names were also favored compared to female names in the majority of experiments, the disparities were less than those demonstrated using **Black versus White names**.
- **Intersektionale Analysen** zeigten große Voreingenommenheit gegenüber Lebensläufen mit männlich- und schwarz-konnotierten Vornamen: Resumes with White male names are preferred in 100% of tests; those with **Black male names** are preferred in **0%**.

## Schlussfolgerung

→ In der Simulation der Lebenslaufbewertung führten alle Systeme zu Diskriminierung von Bewerbenden nach Geschlecht und ethnischer Zugehörigkeit, am stärksten gegenüber "Black male names".

→ Erhöhte Transparenz & Reflektion sozialer Merkmale in Sprachmodellen können Gender & Intersectional Bias und Diskriminierung entgegenwirken; Schlagwort: Promoting Fairness in LLMs: Detection and Mitigation of Gender Bias