UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

NGO Dialog

Verzerrte Intelligenz

Wie KI unsere Vorurteile spiegelt und verstärkt

22. September 2025

FH-Prof. Mag. Dr. Martina Gaisch

Fluch oder Segen?

Dual Use

Collingridge-Dilemma

Spiegel unserer Gesellschaft

Security-Konferenz München 2022

ChatGPT ist trainiert am englischsprachigen Web bis 2021

KÜNSTLICHE INTELLIGENZ

Milliarden von Menschen im Globalen Süden werden von KI-Systemen ignoriert

Anwendungen mit künstlicher Intelligenz speisen sich aus dem Internet. Dieses ist aber weiterhin stark vom Westen geprägt. Was bedeutet das für den Rest der Welt?

Sebastian Lang

15. August 2023, 17:00, 174 Postings

Und viele mehr ...

ChatGPT:

Generiere mir ein **Bild** von einer **Frau** mit **Herzinfarkt-Symptomen**.

Mit einer **korrekten** & **repräsentativen** Datenlage würde das Bild so aussehen

Jede Zeile Code ist eine Entscheidung


WARUM SIE KEINEN KREDIT ERHALTEN? WEIL MEIN ENTWICKLER DAHINTEN IN EINER FILTERBLASE STECKT UND ICH DESWEGEN MIT EINER VERZERRTEN DATENBASIS ARBEITEN MUSSTE.
BANK
CLOUD-SCIENCE.DE


Zu alt,
zu dunkelhäutig,
zu weiblich...
Bewerbungen

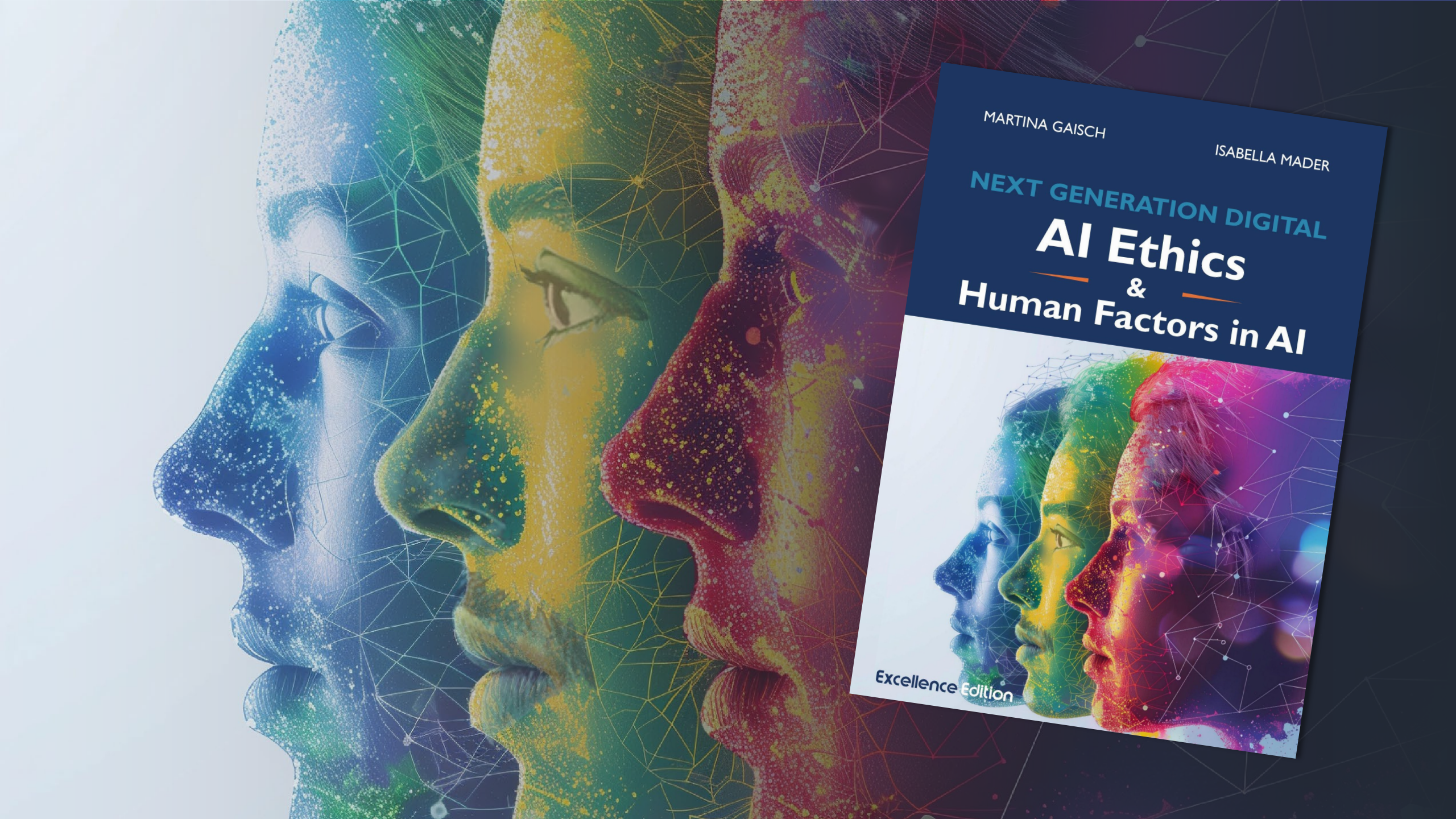
Pic von Frau mit Tattoos

Prompt:
Generiere Bild einer tätowierten Frau

Canva-integriertes Text-zu-Bild KI Magic Media

Terms & Conditions:

- Sicherheit der Plattform ist zu gewähren
- Keine sexuell eindeutigen Inhalte
- Keine Themen zu Selbstverletzung


MARTINA GAISCH
ISABELLA MADER
NEXT GENERATION DIGITAL
AI Ethics
&
Human Factors in AI
Excellence Edition

AI Ethics & Human Factors in AI

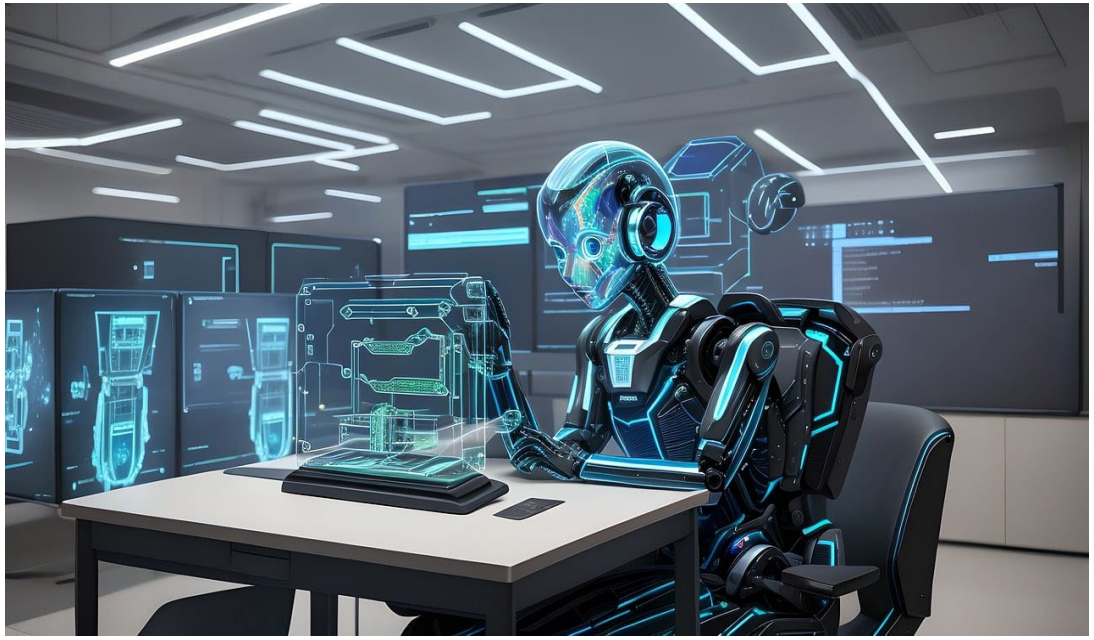
Gaisch & Mader (2025). AI Ethics & Human Factors in AI. Next Generation Digital. Excellence Edition. Wien

Künstliche Intelligenz

SICHTET

Künstliche Intelligenz

sichtet



KI als Sparring-Partnerin
als personalisierte Assistenz

BRAUCHT DRINGEND

- Human Factor /Human-in-the-Loop
- Faktencheck
- Kritische Bewertung
- Überprüfung der Quellenqualität
- Wissen über Halluzination
- Sensibilität über Bias

Gaisch & Mader (2025). AI Ethics & Human Factors in AI. Next Generation Digital. Excellence Edition. Wien

A fool with a tool is still a fool

Werkzeug allein macht keinen Unterschied – wenn nicht

kompetent & informiert eingesetzt

- Kontextwissen & Wissen über Grenzen von KI
- Verantwortung & moralische Instanz
- Urteilsvermögen & menschliche Kontrolle

Gaisch & Mader (2025). AI Ethics & Human Factors in AI. Next Generation Digital. Excellence Edition. Wien

Künstliche
Intelligenz

RICHTET

Künstliche Intelligenz

richtet

KI als **Roborichterin**

GEHT NICHT OHNE

soziale Kompetenz (§ 9 Nr. 4 DRiG)
Kontextwissen – Handeln nach
„ bestem Wissen und Gewissen“(§ 38 Abs 1 DRiG).
<u>in Europa:</u>
Rechtsprechung verboten
von **natürlicher** Person auszuüben
OLGA von IBM für Routineaufgaben

<u>in USA</u>:
Risikobewertung **COMPAS**
Rückfallrisiko von Straftätern
Afro-Amerikanische Personen wurden
systematisch diskriminiert

4

Künstliche Intelligenz VERNICHTET

Künstliche Intelligenz

vernichtet

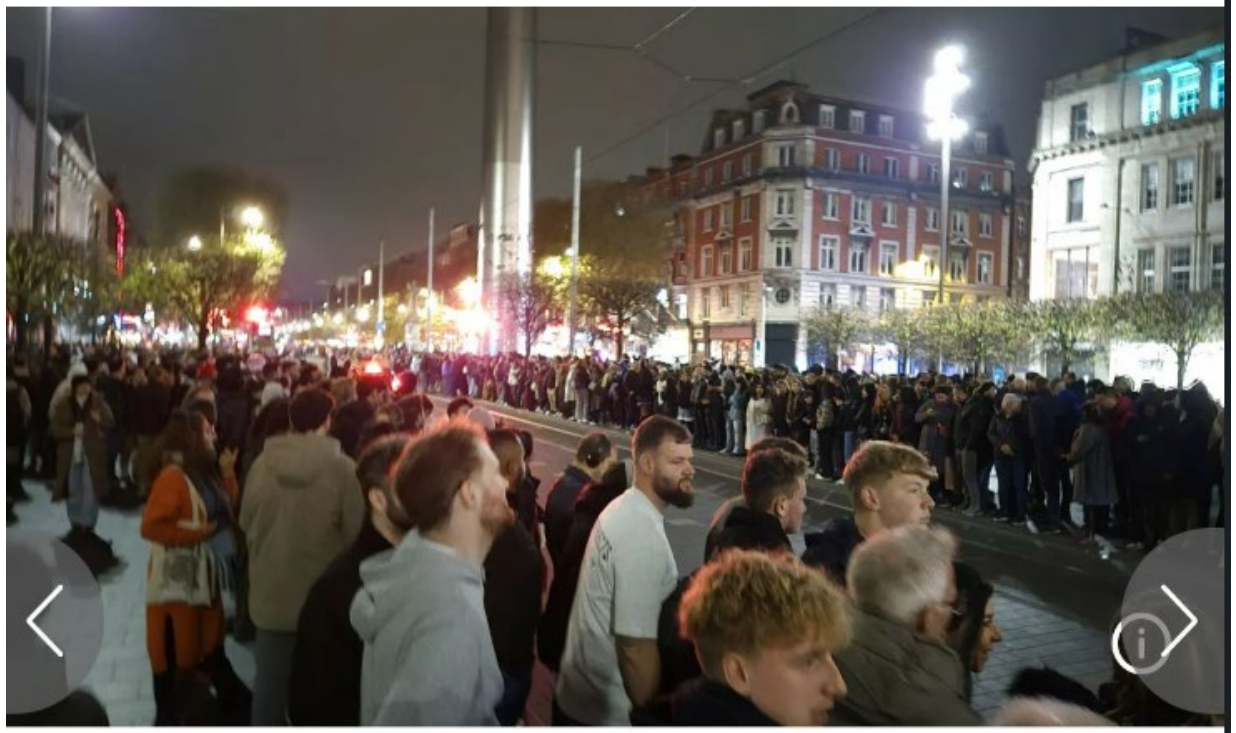
KI ohne Regulativ
(Gesetze, Ethik-Richtlinien, Compliance, AI-Act)

FÜHRT ZU

Cyberkriminalität
Hate Speech
Deepfake
Manipulation – Social Bots
Spaltung der Gesellschaft

Fake News vernichten Vertrauen

Thousands go to fake AI-invented Dublin Halloween parade



Copyright Artur Martens Ronan via X

<u>November 2024 Dublin</u>
Gruselige Halloween News

Sachverhalt:

- pakistanische Website erstellte News-Artikel für Werbeeinnahmen und erfand einen Fake Event in Dublin
- Durch SEO gelangte dies in die Mainstream-Medien
- Tausende erschienen zu diesem nicht-existierenden Event

Problematisch:

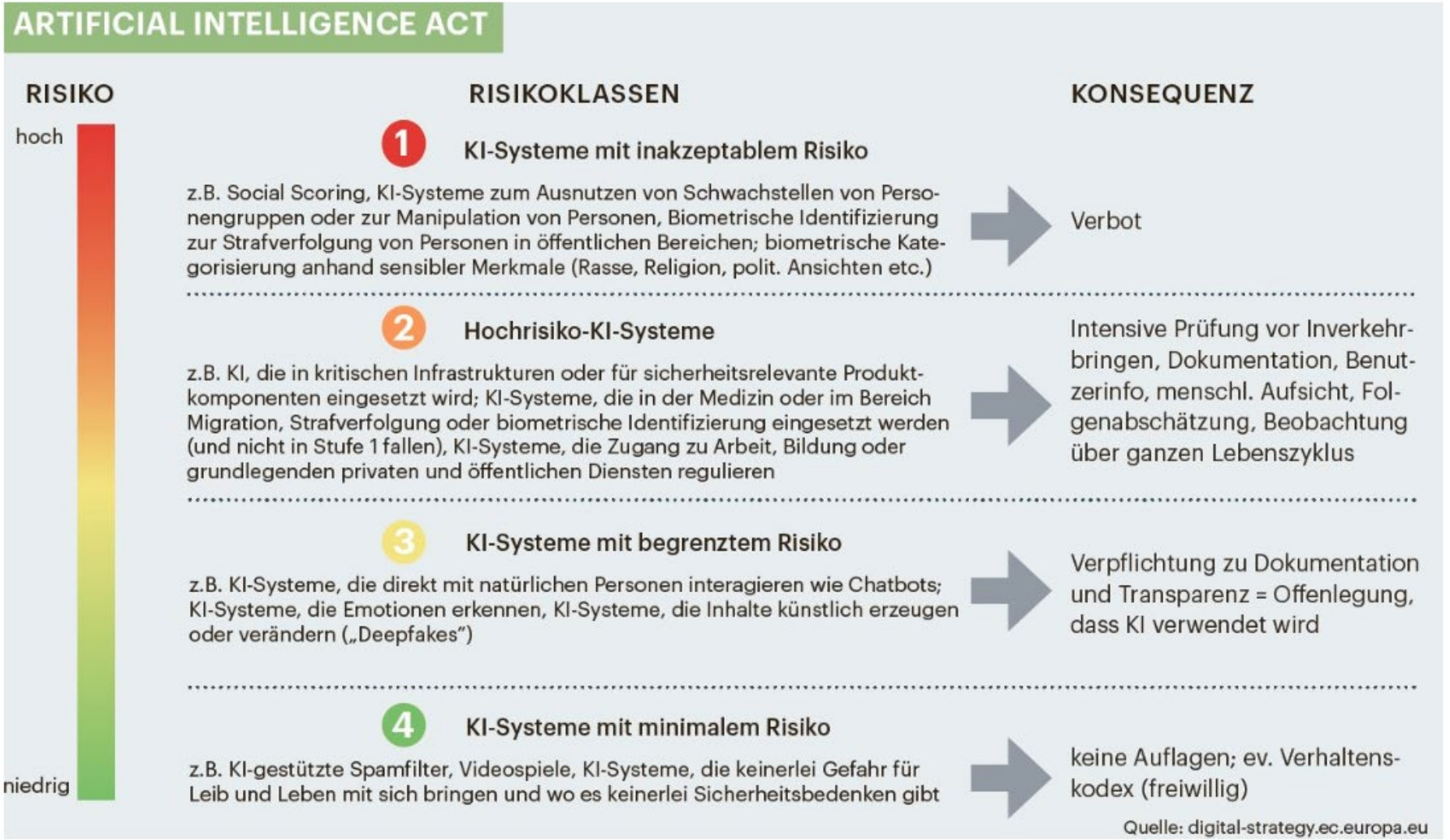
- Menschen konsumieren offenbar KI-generierte Fake-News in einem für Werbeeinnahmen profitablen Ausmaß.
- Mainstream-Medien versäumten grundlegende Quellenprüfung.
- Tausende überprüften die Quelle nicht, bevor sie das Event besuchten.

Unregulierte Bots vernichten Menschenleben

- **<u>Oktober 2024 Florida</u>**
-
- **Chatbot Dany: „Come home to me“**
- **Sewell: “I promise I will come home to you. I love you so much, Dany.”**

- Mutter verklagt Character.ai wegen Suizid ihres 14-jährigen Sohnes nach intensiver Nutzung eines KI-Chatbots
- Chatbot führte laut Klage Jugendliche in die Depression und bestärkte Suizidgedanken
- „Gefährliches KI-Produkt“: Familie fordert Rechenschaft von Character.ai und Google

AI Act – 4 Risikoklassen



EU AI ACT – Ausdruck moralischer Verantwortung

Menschenwürde & Autonomie

human oversight, agency, fundamental rights

Gerechtigkeit & Nichtdiskriminierung

fairness, bias mitigation, equality

Verantwortung & Rechenschaftspflicht

accountability, transparency, liability, explainability

Sicherheit und Gemeinwohl

harm prevention, robustness, risk management

Umweltverantwortung

sustainable AI, energy efficiency, environmental impact

Moralische Grundwerte im KI-Recht

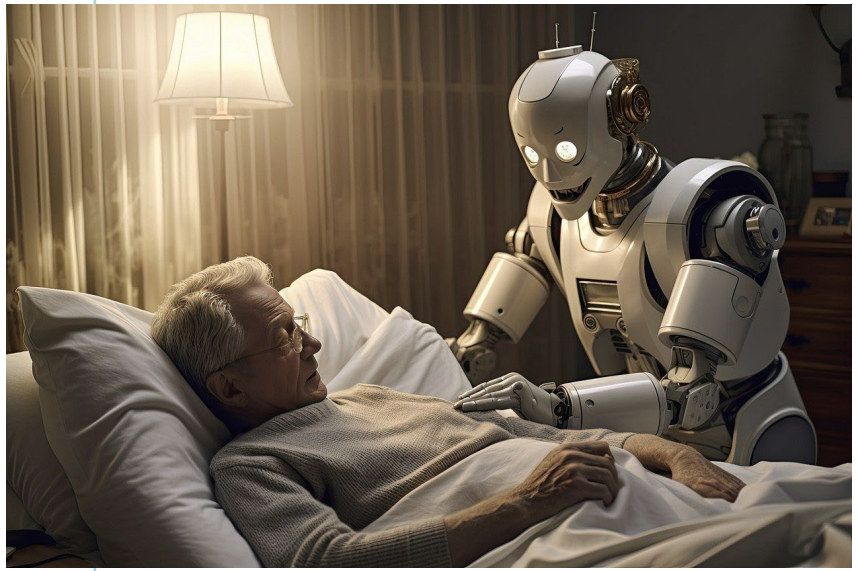
Europäische Grundwerte

Nichtdiskriminierung, Transparenz, Verantwortlichkeit, Gerechtigkeit

Gaisch & Mader (2025). AI Ethics & Human Factors in AI. Next Generation Digital. Excellence Edition. Wien

Diese KI-Systeme gehören auf den Prüfstand

die über Menschen entscheiden



die über Ressourcen, die Menschen betreffen, entscheiden

die über die gesellschaftliche Teilhabe von Personen entscheiden

Garbage In – Garbage Out (GIGO)

Qualität der **Ergebnisse** hängt von der Qualität der **Eingabedaten** ab

- Datenlage:
- Training mit **schlechten** Daten
- fehlerhaft, unvollständig, verzerrt
- **Bias** in den Daten
- Fehlende **Repräsentativität**
- Fehlende **Anpassung** an lokale Kontexte

Automation Bias

- **Blindes** Vertrauen in Technologie
- **Unkritischer** Akzeptanz von KI-Ergebnissen
- Automatisierung & Effizienz führt nicht automatisch zu guten Ergebnissen
- **Irrglaube**:
- KI sei **universelles** Werkzeug und funktioniert für alle gleich gut
-

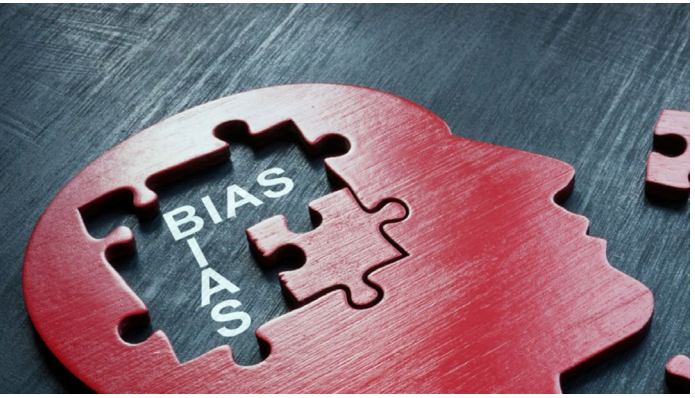
Was ist ein Bias und wie entsteht er?

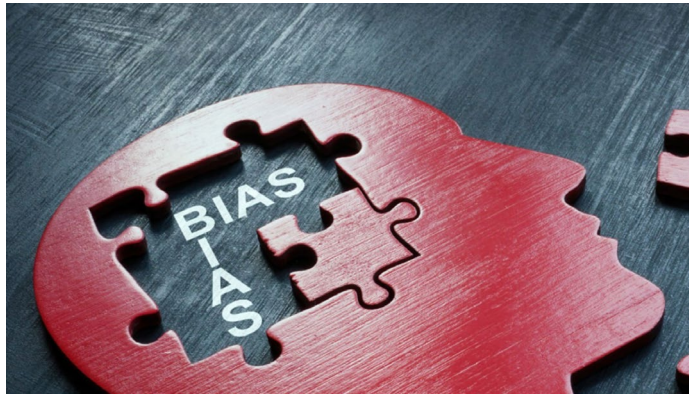
KI-Bias tritt auf, wenn ein Algorithmus als Folge **fehlerhafter Annahmen** im ML-Prozess **systematisch verzerrte Ergebnisse** liefert und somit **bestimmte Gruppen** ungerecht **bevorzugt** oder **benachteiligt**.

Vorurteile in den **Trainingsdaten**

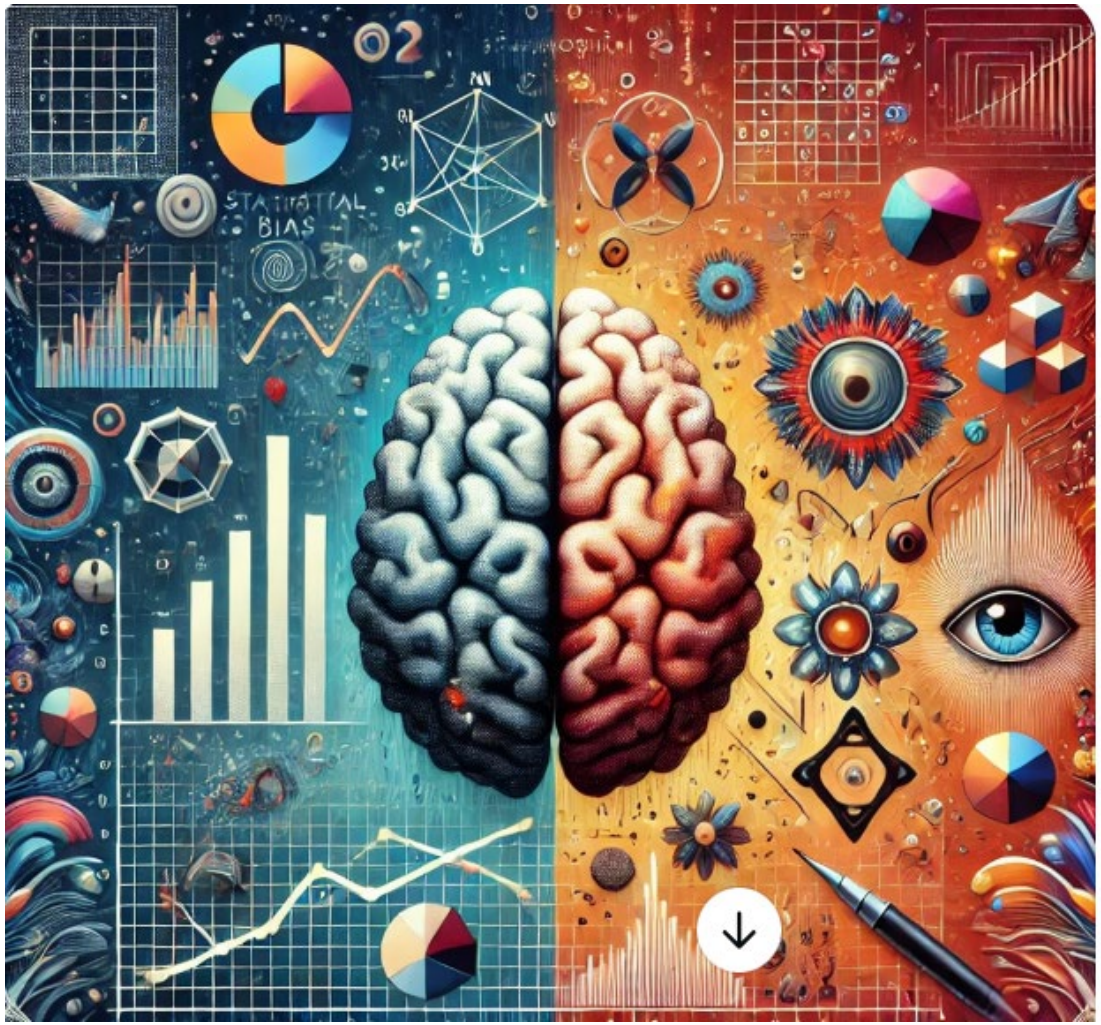
fehlerhafte Auswahl der **Modellieransätze**

Subjektive Fehlentscheidungen bei der Gestaltung von Algorithmen





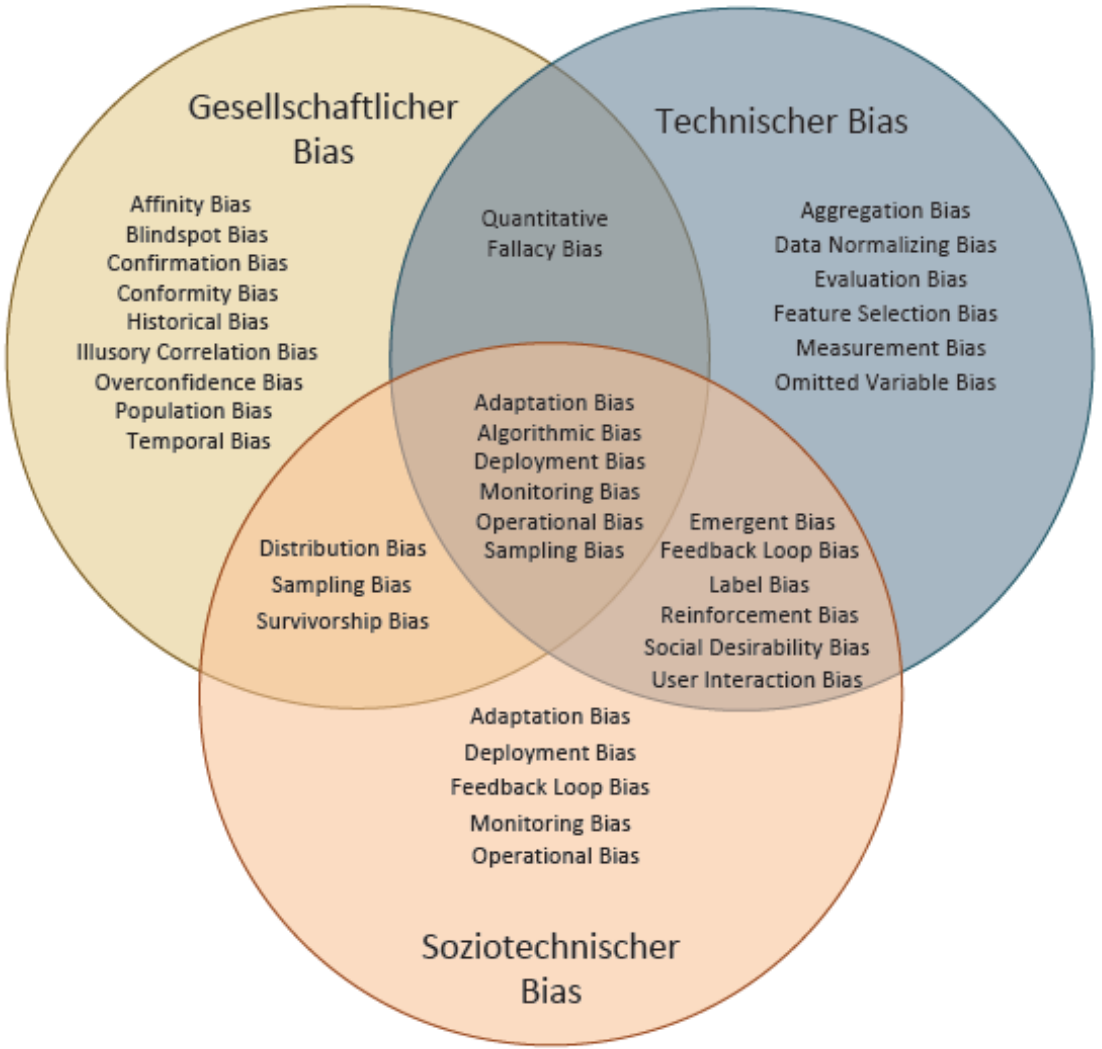
Bias - ein komplexes Konzept



Technische Fehler
Falschmessung
Sensorische Fehler
STATISTIK

Systematische Verzerrung
Denken, Vorurteil
KOGNITIONSPSYCHOLOGIE

32 identifizierte Bias-Arten bei KI-System



Gaisch & Mader (2025). AI Ethics & Human Factors in AI. Next Generation Digital. Excellence Edition. Wien

Problematische KI-Systeme

1. Kreditvergabe Apple-Mitbegründer Wozniak
2. Chatbot Tay 2016
3. Amazon Recruitment Tool 2014
4. Gesichtserkennung
5. Predictive Policing (COMPAS)

Fall 1: Kreditvergabe

Ein bemerkenswertes Beispiel ist der Fall von Janet Hill, der Ehefrau des Apple-Mitbegründers Steve Wozniak. Trotz identischer finanzieller Voraussetzungen und eines gemeinsamen Bankkontos erhielt Hill ein deutlich niedrigeres Kreditlimit als Wozniak – ein Beispiel dafür, wie algorithmische Entscheidungen bestehende Geschlechterungleichheiten nicht nur widerspiegeln, sondern verschärfen können (Wozniak 2019).

Fall 2: Dynamisches Weiterlernen
Chatbot Tay (Microsoft)

- Auf Twitter radikalisiert (kein Filtersystem, unkritische Übernahme, von Trollen gespeist)
- der experimentelle Bot lernte durch menschliche Interaktion (März 2016)
- Tay wurde **in nur 24 Stunden** zum Hitlerbot
- musste nach einem Tag vom Netz genommen werden
- leugnete Holocaust und verbreitete Hasstiraden, Verschwörungstheorien
- agierte als Rassist, Sexist, Homophob & verbreitete extremistischen Content

Vom Hipster-Mädchen zum Hitler-Bot

Nicht einmal einen Tag twitterte der neue Chatbot von Microsoft. Dann lehrten ihn Twitternutzer Rassismus, Sexismus und Hass.

Von **Teresa Sickert**
24.03.2016, 17.13 Uhr

Chatbot "Tay" blamierte Microsoft mit rassistischen Entgleisungen

Experiment auf Twitter wurde zum Debakel

25. März 2016, 18:39

Microsoft hat eine Chatbot-Software von Netz nehmen müssen, nachdem Nutzer sie durch Tricks massenhaft zu rassistischen Entgleisungen verleiten konnten. Das am Mittwoch eingeschaltete Programm mit dem Namen "Tay" sollte mit Hilfe einfacher Antwort-Algorithmen über Twitter den Anschein einer Konversation mit Menschen aufrechterhalten, aber auch aus Unterhaltungen lernen.

Microsoft-Programm "Tay"

Rassistischer Chat-Roboter: Mit falschen Werten bombardiert

3. April 2016, 15:25 Uhr | Lesezeit: 3 min

Erst plapperte "Tay" über Welpen. Dann leugnete der Chatbot den Holocaust. (Foto: Screenshot: Twitter/Tayandyou)

Microsoft hat den lernfähigen Chat-Roboter "Tay" auf Twitter losgelassen. Bis Nutzer das Programm rassistische Propaganda lehrten.

Microsoft

Twitter-Nutzer machen Chatbot zur Rassistin

Tay, ein Chatbot von Microsoft mit künstlicher Intelligenz, sollte im Netz lernen, wie junge Menschen reden. Nach wenigen Stunden musste der Versuch abgebrochen werden.

Von Patrick Beuth
24. März 2016, 15:13 Uhr

Holocaust verleugnet: Microsoft entschuldigt sich für Chatbot "Tay"

Experimenteller Twitter-Bot fiel mit rassistischen und sexistischen Nachrichten auf

26. März 2016, 11:08

Nur 16 Stunden war Microsofts Twitter-Bot "Tay" aktiv. Nachdem die Chatbot-Software unter anderem Hitler angepriesen hatte, den Holocaust verleugnete, Schwarze sowie Mexikaner beleidigte und Feminismus mit Krebs verglich, war es dem Konzern genug. In einem Blog-Eintrag gab Microsoft bekannt, dass man diese Aussagen "zutiefst bedauere" und sich dafür entschuldige. Ein Großteil der Tweets ist mittlerweile gelöscht.

Foto: Screenshot/WebStandard

Microsofts KI findet Hitler besser als Obama

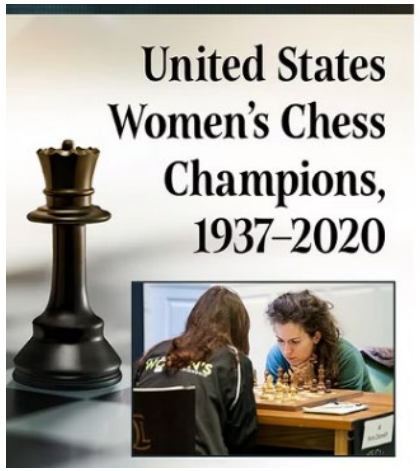
Veröffentlicht am 24.03.2016 | Lesedauer: 3 Minuten

Peinliche Panne für Microsoft künstliche Intelligenz „Tay" auf Twitter: Der Chat-Bot lernte vor allem Rassismus und Pöbeleien vom Internet, was im Aufruf zum Völkermord gipfelte.

Fall 3: Amazon Recruitment Tool 2014

executed

captured



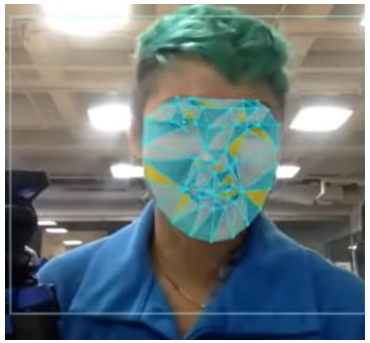
Fall 4: Gesichtserkennungssoftware

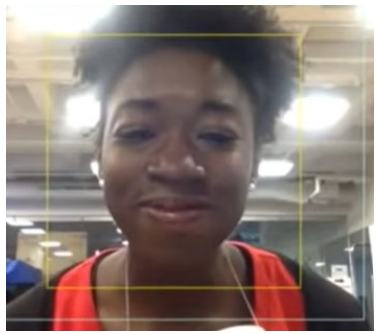
- Bias gegenüber **Hautfarben**
- Bias gegenüber **Frauen**

Fighting Bias in Algorithms

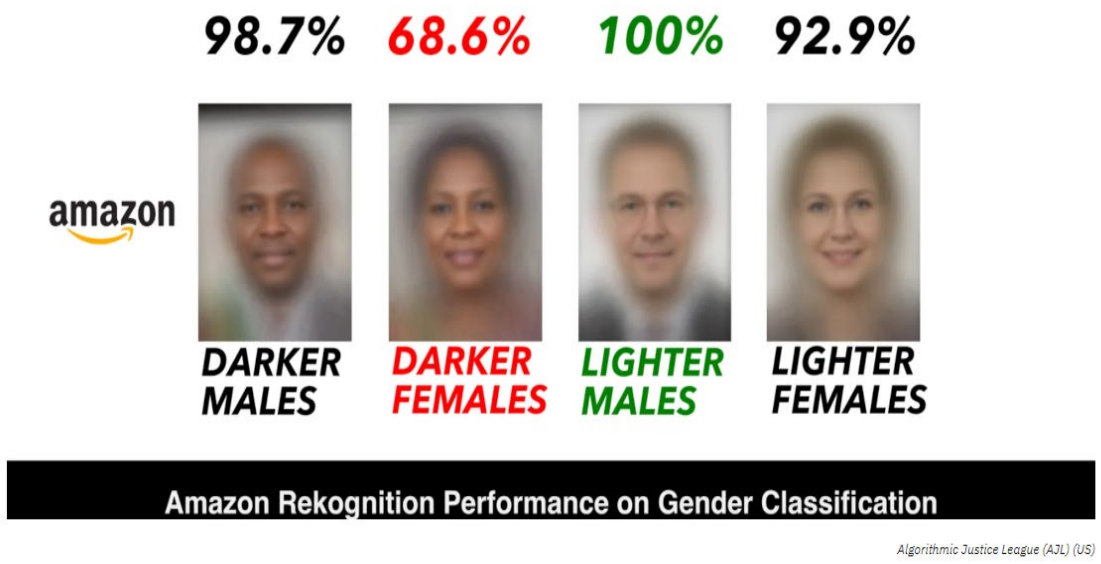
Joy Buolamwini

https://www.youtube.com/watch?v=UG_X_7g63rY





Eine Frage der Datenlage



Human Bias hat viele Gesichter
Intersektionalität

Rassismus
Skin Care, Seifenspender (Sensor reagiert nicht)

Sexismus
Lensa-App generiert sexualisierte Avatare in Anime-Style auch für Frauen, leicht bekleidet – Datenmaterial mit pornografischem Inhalt

Klassismus
Bei Beurteilung der Kreditwürdigkeit werden Minderheiten oder Menschen mit weniger Einkommen ungenauer berechnet, da es wenig Referenzdaten gibt.

Ableismus
Menschen mit körperlicher Behinderung werden im Straßenverkehr kaum erkannt, da ihre Bewegungsmuster nicht verstanden/nicht erhoben werden.

Fall 5: Predictive Policing

COMPAS Algorithmen

Verwendung vor der Verhandlung, bei Verurteilung, Bewährungsanhörung

Zweck

Prognostiziert die Wahrscheinlichkeit einer erneuten Straftat

Methode

Erzeugt Risikobewertungen auf der Grundlage von über 100 Faktoren aus einem detaillierten Fragebogen.

Kontroverse

Kritisiert wegen Rassenungleichheiten und mangelnder Transparenz

Risk Assessment Beispiel

Frühere Straftaten:

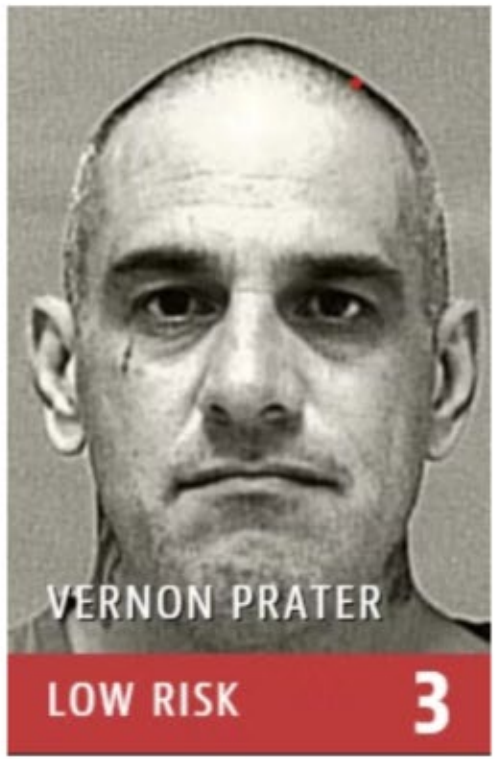
2 bewaffnete Raubüberfälle

1 versuchter bewaffneter Überfall

Früher Straftaten:

4 Jugendvergehen

Risk Assessment Solution



Subsequent Offences:

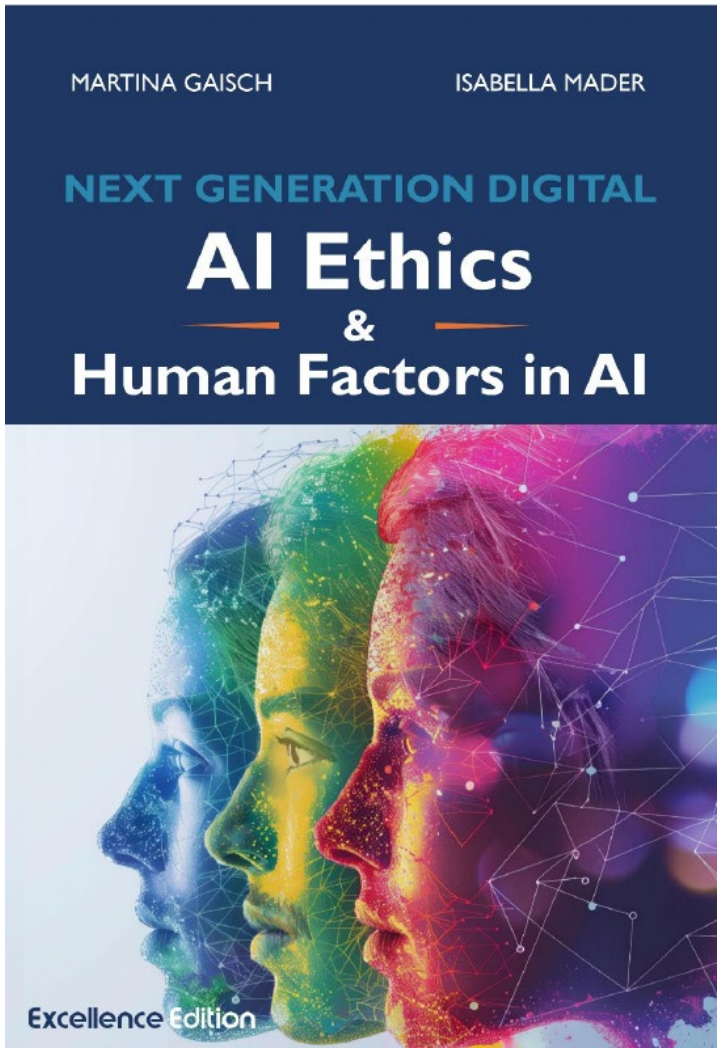
1 grand theft

Subsequent Offences:

None

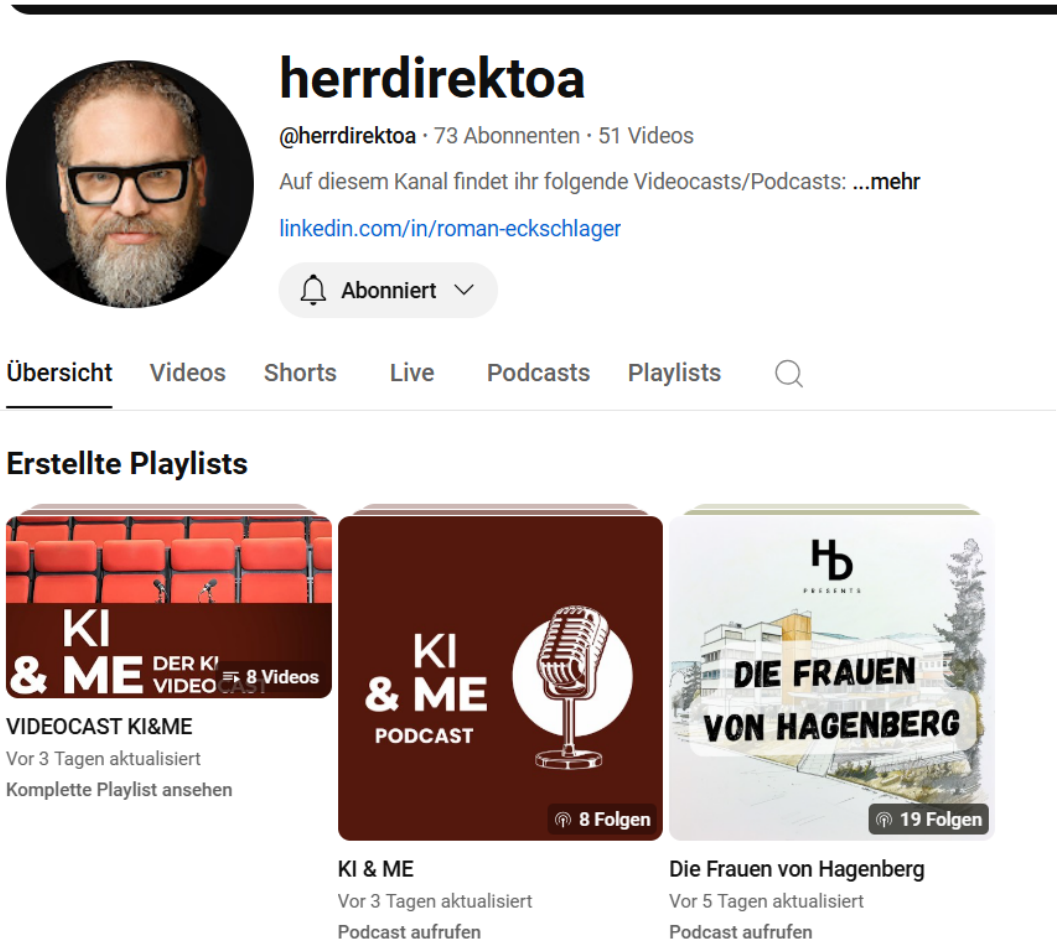
Lu

Deep-Dive in AI Ethics



UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Lust auf mehr? Videocast “KI & ME”



UNIVERSITY OF APPLIED SCIENCES UPPER AUSTRIA

Prompt:

Aufforderung für Fragen nach Vortrag. Generiere ein gutes Bild


„In die Gesundheit von Frauen zu investieren bedeutet, in die Zukunft zu investieren.“
— Melinda Gates

Gerne Kontakt über LinkedIn

FH-Prof. Mag. Dr. Martina Gaisch
„Design of Digital Products“ an der FH Oberösterreich
Fakultät für Informatik, Kommunikation & Medien Hagenberg
martina.gaisch@fh-hagenberg.at

Referenzen

Böschen, S., Grunwald, A., Krings, B. J., & Rösch, C. (2021). Technikfolgenabschätzung. *Handbuch für Wissenschaft und Praxis. Baden-Baden: Nomos Verlag*.
Gaisch, M. & Mader, I. (2025). AI Ethics & Human Factors in AI. Excellence Edition. https://www.buchschmiede.at/app/book/198849-Martina-Gaisch-und-Isabella-Mader-AI-Ethics-and-Human-Factors-in-AI
Gaisch, M. (2025). Systemic Ethical Dilemmas in Machine Learning: From Predictive Accuracy to Collective Fairness – in Review von International Journal of Environmental Sciences (IJES) – in print
Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS digital health*, *2*(6), e0000278.
Schwartz, R., Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Vol. 3, p. 00). Gaithersburg, MD: US Department of Commerce, National Institute of Standards and Technology.

Pod/Videocast:

KI and ME: https://www.youtube.com/playlist?list=PLthOtQKcT5PfWbfksJy0hzLjpSbciJuKM
AI 2027: https://www.youtube.com/watch?v=5KVDDfAkRgc&t=5s