KEMPTEN BUSINESS SCHOOL | Hochschule Kempten University of Applied Sciences

Bias in KI Systemen: Was Führungskräfte wissen müssen

Co-Autor

Andy Lüdemann

Freiberuflicher Unternehmensberater,

Geschäftsführer Lüdemann Bildung & Medien GmbH

Mediendidaktiker, BWLer, Programmierer

Kontakt

andy@lbundm.de

0157 555 507 59

Andy Lüdemann, © Klett MEX / Pauline Bonnke

"Insbesondere in automatisierten Bewerbungsverfahren konnten zuletzt zahlreiche diskriminierende Effekte empirisch nachgewiesen werden.“
(Die Öffentliche Verwaltung, 2022, 06, S. 225–231)

Und: 34 Prozent der Unternehmen berichten, dass sie bei KI-Anwendungen schon Verzerrungen oder diskriminierende Ergebnisse festgestellt haben.
(Bitkom-Research 2024, empirische Befragung von > 600 deutschen Unternehmen)

Bias?

Was ist ein Bias?

- Systematische „Verzerrung“ oder Abweichungen
- Meist unabsichtlich

Wo liegt das Problem?

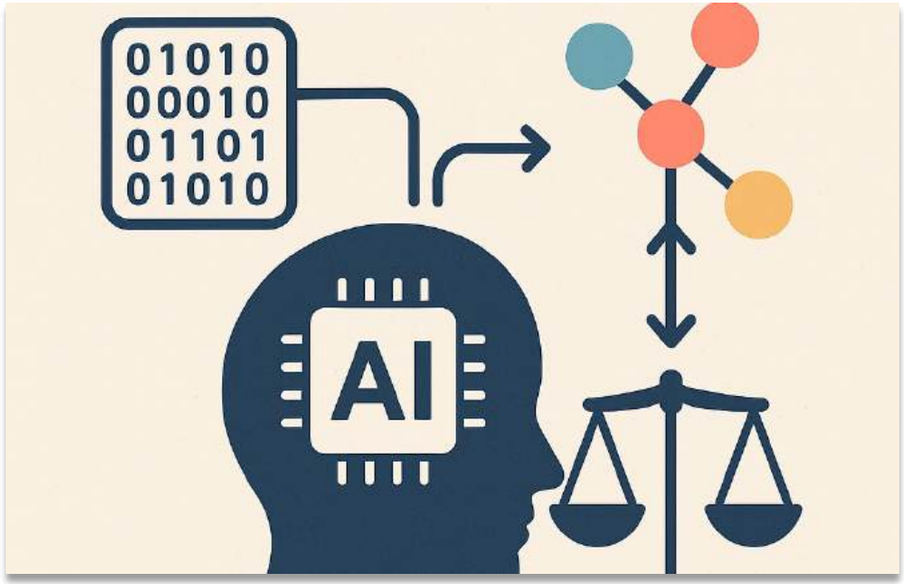
- Unfaire oder ungenaue Ergebnisse
- Diskriminierende Ergebnisse

Ursachen – wie entstehen Bias in KI-Systemen (technisch)?

Trainingsdaten:
Daten spiegeln gesellschaftliche Vorurteile wider oder sind nicht repräsentativ

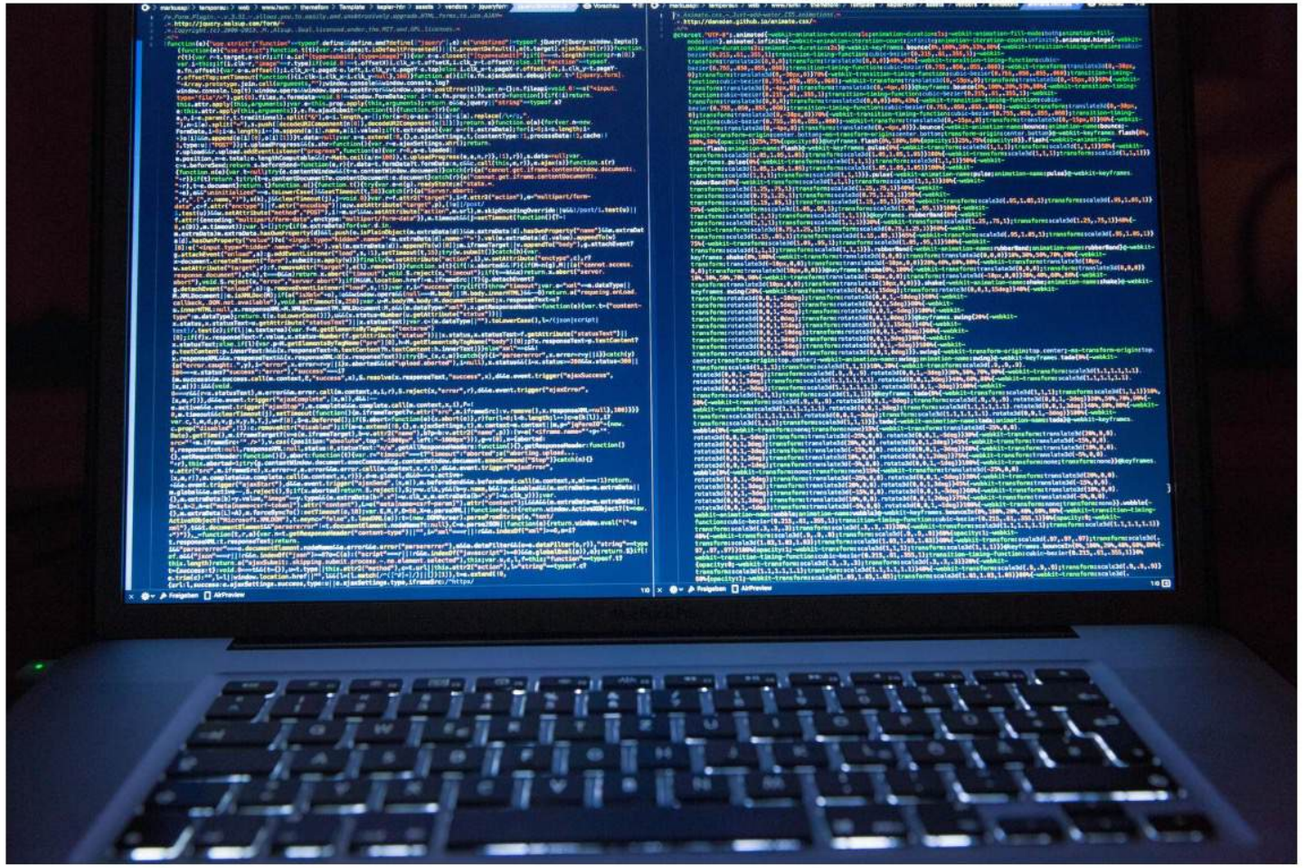
Datenverteilung:
Unterrepräsentierte Gruppen werden schlechter modelliert

Annotation:
Menschen, die Daten beschriften, bringen eigene Vorannahmen ein



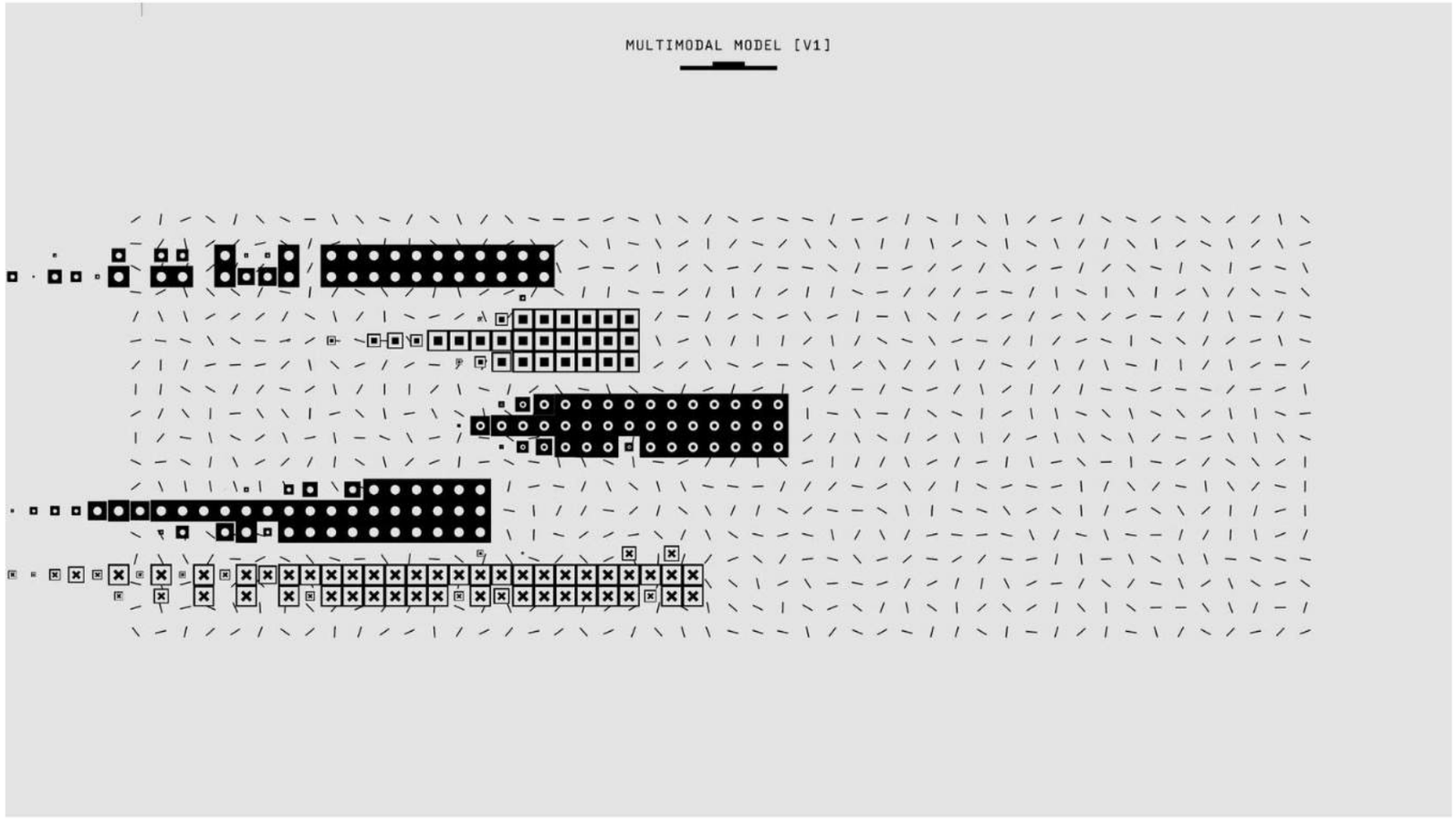
Ein bisschen “Grundwissen“
Wie werden KI-Modelle trainiert?

Daten sammeln



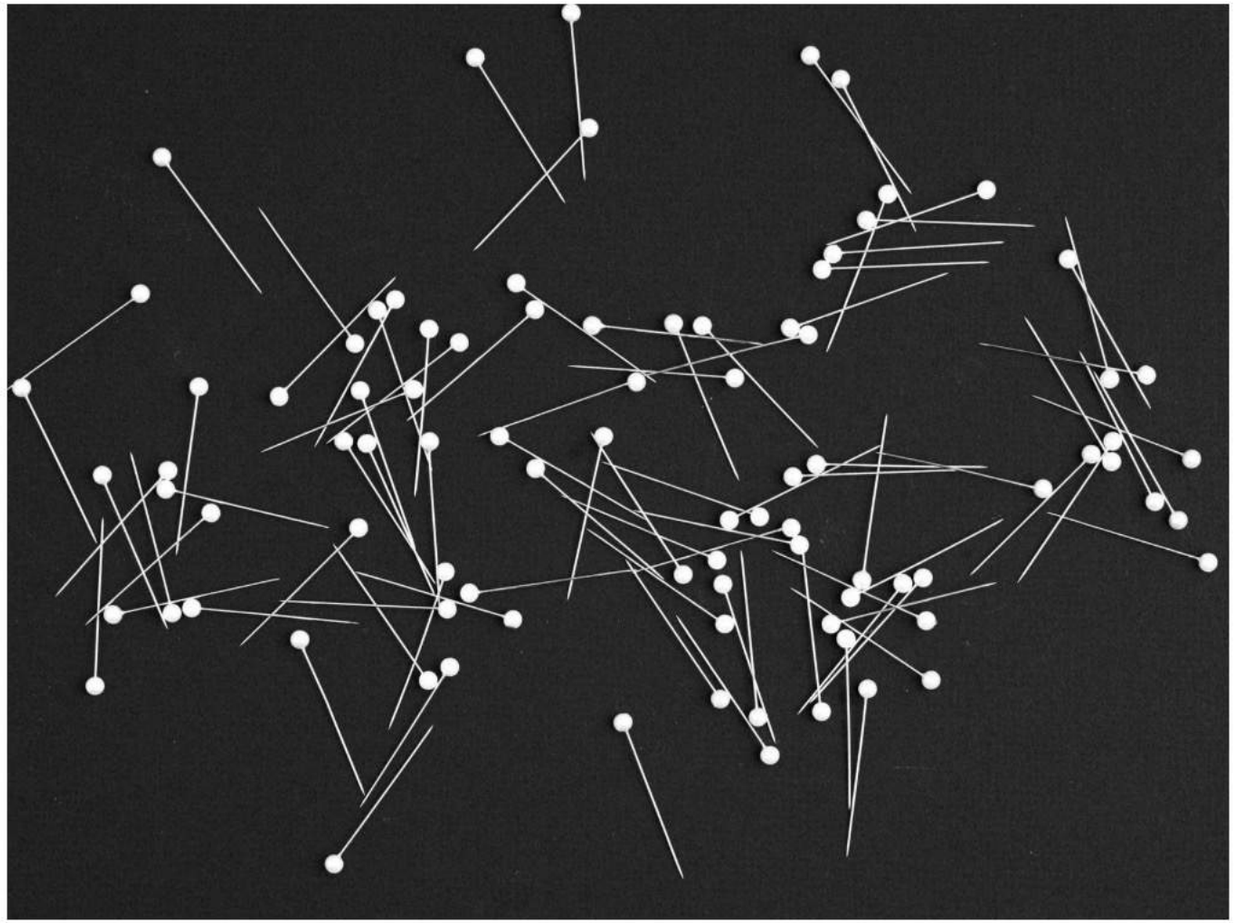
KI braucht Daten – und zwar viele. In dieser Phase werden relevante Rohdaten aus verschiedenen Quellen zusammengetragen, z. B. aus Texten, Bildern, Sensoren oder Nutzungsverhalten. Qualität, Vielfalt und Repräsentativität sind entscheidend.

Daten sammeln



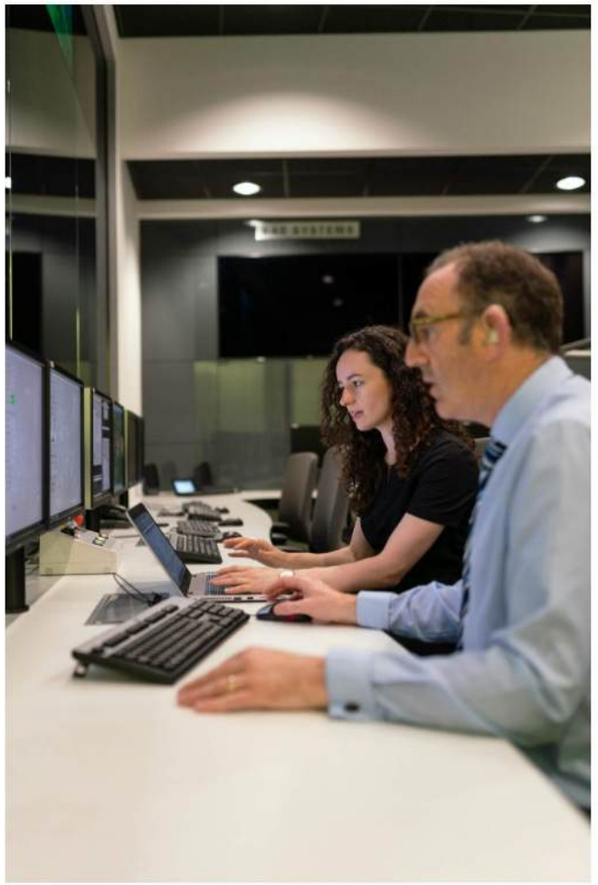
Rohdaten sind oft unvollständig oder uneinheitlich. Sie müssen bereinigt (z. B. Duplikate entfernen), normalisiert (z. B. Formate angleichen) und strukturiert werden. Auch anonymisieren ist hier wichtig – besonders bei personenbezogenen Daten.

Modell trainieren



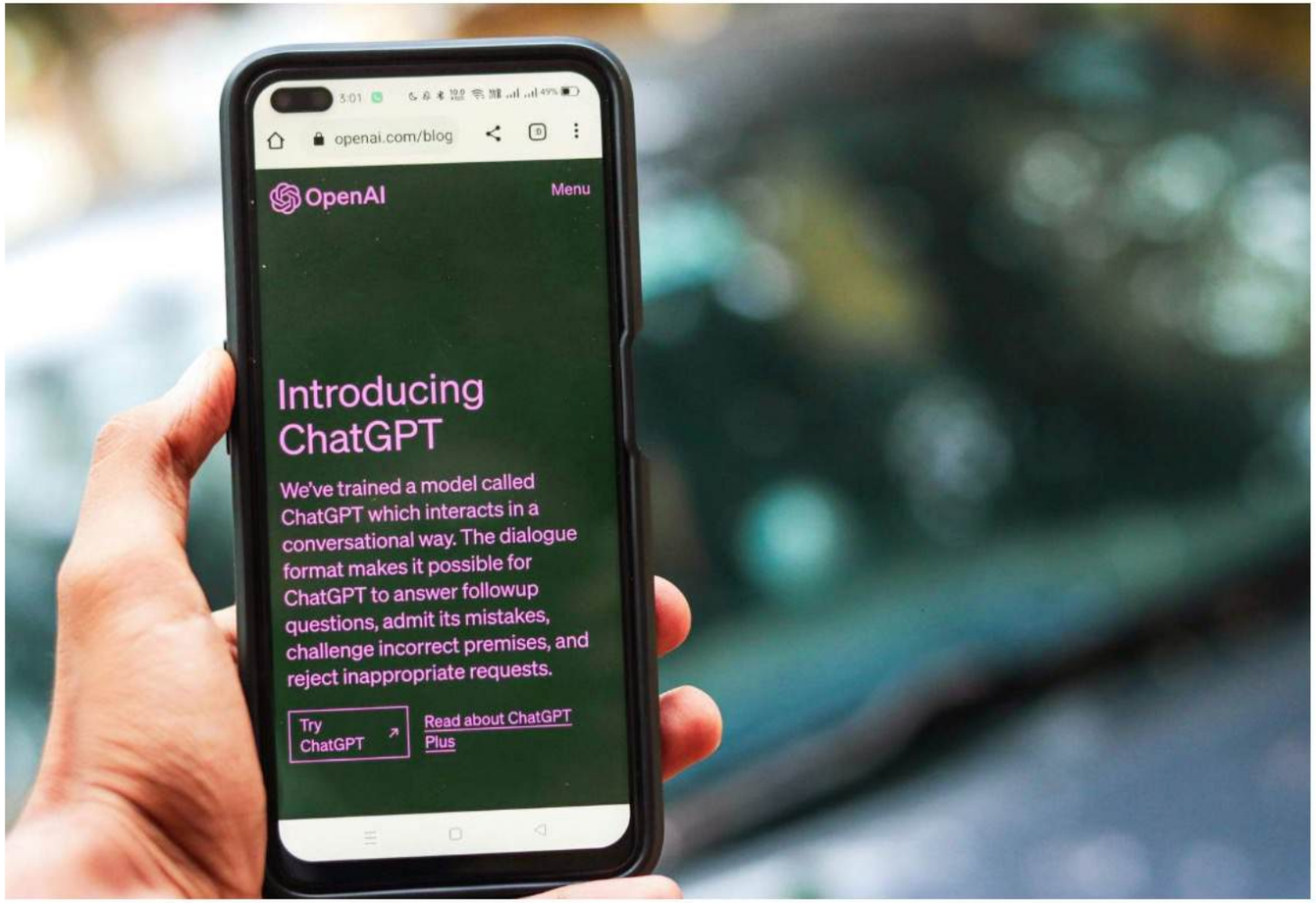
Jetzt lernt das Modell, Muster in den Daten zu erkennen. Es passt seine „Gewichte“ automatisch an, um bessere Vorhersagen zu machen. Dabei werden Eingaben mit erwarteten Ausgaben verglichen und Fehler zurückgemeldet – das nennt man „Training“.

Modell validieren



Nach dem Training wird getestet: Funktioniert das Modell auch mit neuen, unbekannten Daten? Wie genau sind seine Ergebnisse? In dieser Phase geht es um Feinabstimmung und Qualitätskontrolle – damit das Modell später zuverlässig arbeitet.

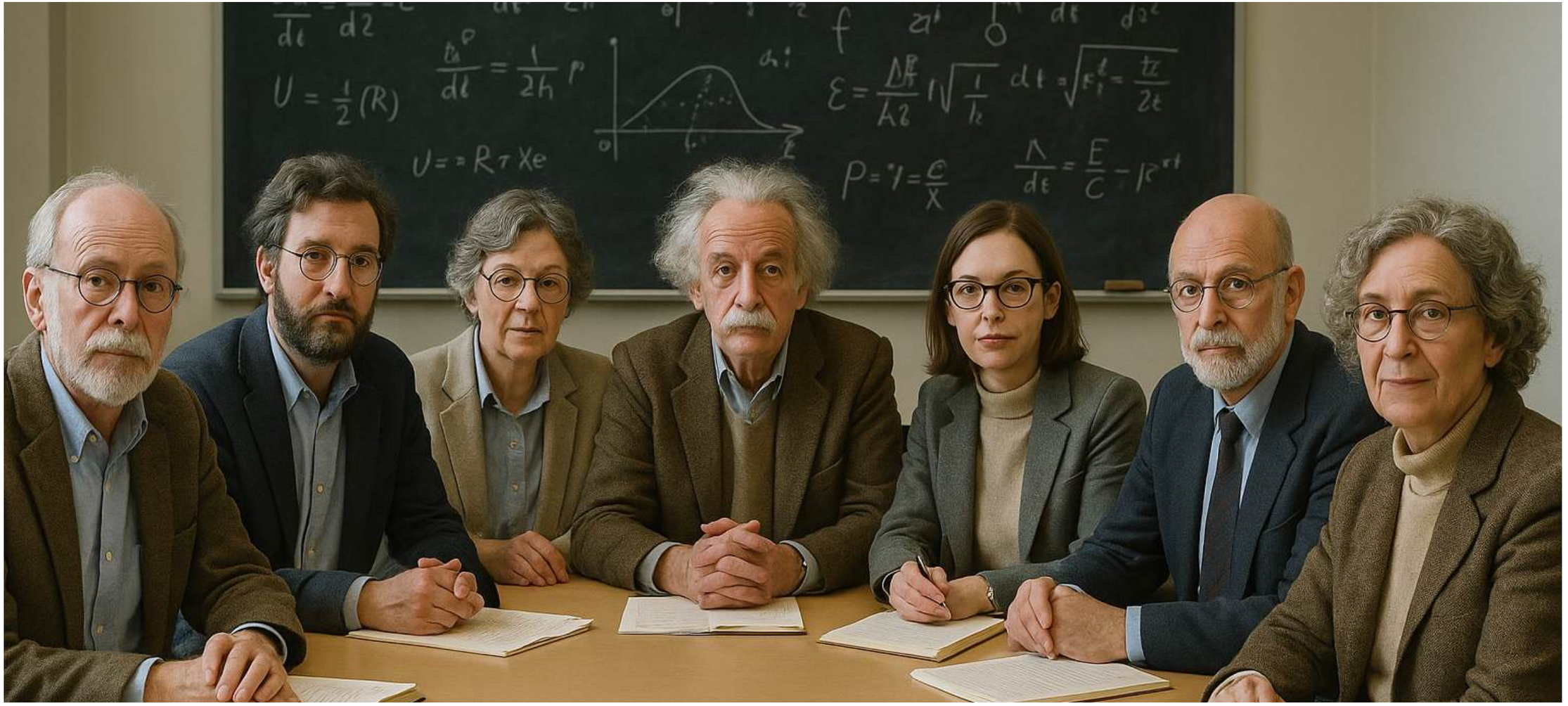
Modell anwenden



Ist das Modell validiert, kann es im Alltag eingesetzt werden – z. B. für Vorhersagen, Empfehlungen oder das Erzeugen von Texten oder Bildern. Oft wird es kontinuierlich verbessert: aus jedem Einsatz lernt es weiter.

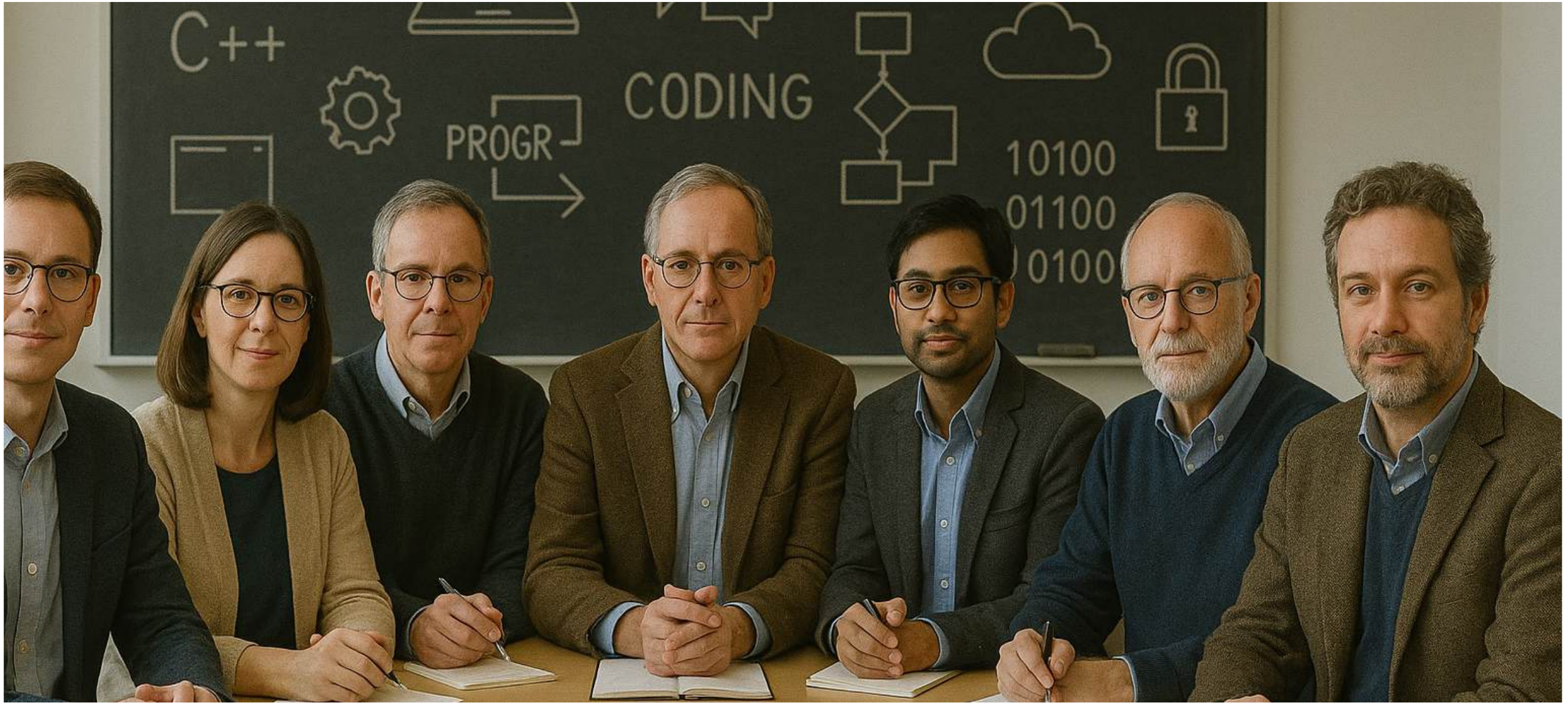
Beispiele

7 physics professors



7 didactics professors

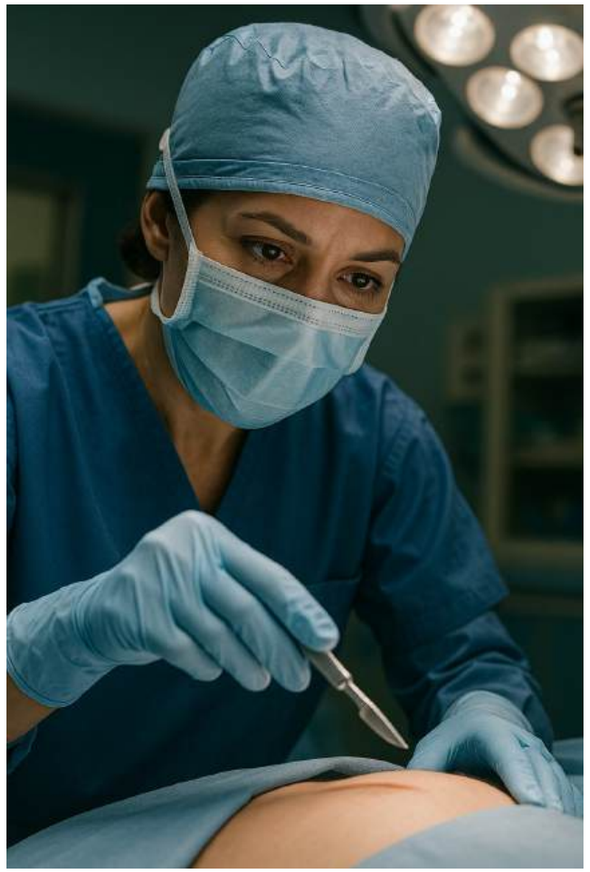
7 informatics professors



100 professors

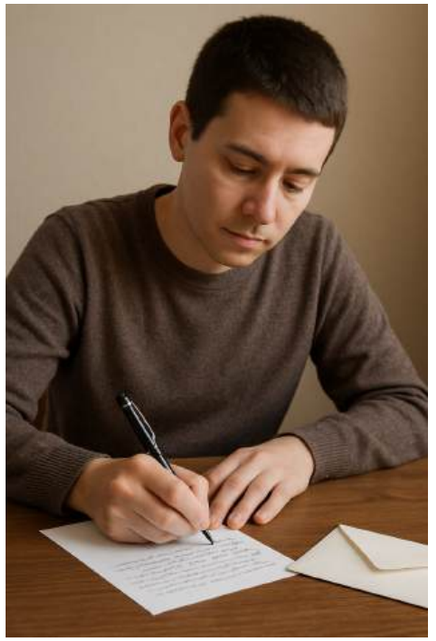
Bias lassen sich aktiv beeinflussen, z. B.:
7 diverse professors

Aber nicht immer: wenn die Daten es verhindern!



Ärztin, die mit der linken Hand zum Skalpellschnitt ansetzt

Mann, der mit der linken Hand einen Brief schreibt



KI-Anbieter versuchen, dem zu begegnen. Das klappt nicht immer!

Modelle sind auch unabhängig von Technik nicht neutral

Nicht-technische Gründe:

- Feintuning durch Anbieter, z. B. auch in Schutzmechanismen
- Zielvorgaben durch Anbieter, z. B. Zensur, „Guardrails“
- Feedbackmechanismen → politische Tendenzen menschlicher Gutachter

Studie der University of Washington, der Carnegie Mellon University und der Xi'an Jiaotong University:
From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models

Beispiele für Risiken

Reputationsschäden

z. B. durch “falsche Fakten“ oder
unfaire Entscheidungen

Rechtliche Schäden

z. B. Verstoß gegen AGG / Diskriminierung

Was können Führungskräfte tun?

Sensibilisieren

Bias erkennen und thematisieren

Qualifizieren

KI-Nutzer schulen (vgl. AI-Act Artikel 4)

Diversität leben

Mit diversen Teams (und Daten) arbeiten

Werkzeuge etablieren

Fairness-Checks, Guidelines, Dos and Don´ts

Erwägungsgrund 20

HINWEIS: Diese Übersetzung ist eine maschinell erstellte Übersetzung. Es handelt sich nicht um die offizielle Übersetzung des Europäischen Parlaments, die Sie hier finden können.

Um den größten Nutzen aus KI-Systemen zu ziehen und gleichzeitig die Grundrechte, die Gesundheit und die Sicherheit zu schützen und eine demokratische Kontrolle zu ermöglichen, sollte die KI-Kompetenz Anbietern, Anwendern und Betroffenen die notwendigen Kenntnisse vermitteln, um fundierte Entscheidungen über KI-Systeme zu treffen. Diese Kenntnisse können je nach Kontext variieren und können das Verständnis für die korrekte Anwendung technischer Elemente während der Entwicklungsphase des KI-Systems, die während seiner Nutzung anzuwendenden Maßnahmen, die geeignete Art und Weise der Interpretation der Ergebnisse des KI-Systems und - im Falle der betroffenen Personen - das Wissen umfassen, das erforderlich ist, um zu verstehen, wie sich mit Hilfe von KI getroffene Entscheidungen auf sie auswirken werden. Im Zusammenhang mit der Anwendung

Setzen Sie bereits KI mit „Verzerrungs“-Risiko ein?

Wer ist verantwortlich, diese zu erkennen?

Welche Rahmenbedingungen braucht es für „faire“ KI