KOMPASS

KÜNSTLICHE INTELLIGENZ UND VOREINGENOMMENHEIT (*BIAS*)

Themenblatt

KI und Voreingenommenheit (*Bias)*

Was ist Voreingenommenheit?

Voreingenommenheit (*Bias*) in der KI bezeichnet verzerrte oder unfaire Präferenzen oder Vorurteile in ihren Vorhersagen, die unter anderem durch fehlerhafte Daten, Algorithmen oder menschliche Vorurteile verursacht werden.

Welche Verzerrungen beeinflussen KI-Modelle?

Es gibt vielfältige Verzerrungen, die KI-Modelle und ihre Outputs beeinflussen können. Im Folgenden eine kurze Erläuterung der wichtigsten:

Datenverzerrung (*Data bias*):

Sie kann bereits bei der Erfassung und Aufbereitung der Daten auftreten. Sind Datensätze nicht repräsentativ, liefern darauf trainierte Modelle verzerrte Ergebnisse.

Beispiel: Ein Spracherkennungsmodell, das nur mit männlichen Stimmen trainiert wurde, erkennt weibliche Stimmen schlechter.

Algorithmische Verzerrung (*Algorithmic bias*):

Sie entsteht, wenn das System bestimmte Gruppen systematisch bevorzugt oder benachteiligt. Häufig geschieht das, weil einzelne Faktoren stärker gewichtet werden als andere, wodurch bestehende Ungleichheiten verstärkt werden. Besonders kritisch ist dies in sensiblen Bereichen wie Personalwesen, Kreditvergabe oder Strafverfolgung, wo verzerrte Vorhersagen Diskriminierung begünstigen können.

Beispiel: Ein Kreditbewertungsalgorithmus stuft Personen aus bestimmten Postleitzahlgebieten systematisch schlechter ein, weil diese Gebiete in den Trainingsdaten häufiger mit Zahlungsausfällen verbunden waren – unabhängig von der tatsächlichen Kreditwürdigkeit der jeweiligen Person.

Menschliche Vorurteile (*Human bias*):

Da KI-Systeme von Menschen entwickelt werden, fließen deren unbewusste oder bewusste Vorurteile oft in die Modelle[1] ein. Dies geschieht etwa durch Entscheidungen bei der Auswahl und Aufbereitung von Daten, der Definition relevanter

[1] In der KI bezeichnen Modelle konstruierte mathematisch-statistische Strukturen, die auf Basis großer Datenmengen trainiert werden, um bestimmte Aufgaben wie beispielsweise Texterkennung, Bilderkennung oder Vorhersagen zu erfüllen. Sie bilden das Zentralelement eines KI-Systems und beeinflussen dessen Verhalten und Ergebnisse maßgeblich.

Merkmale oder der Modellanpassung. Selbst gut gemeinte Ansätze können unbeabsichtigt bestehende Perspektiven und Annahmen reproduzieren und so die Voreingenommenheit verstärken.

Beispiel: Ein Entwicklungsteam hält die Muttersprache bei der Gestaltung eines KI-Modells für unwichtig. Die KI bewertet daher Bewerberinnen und Bewerber, die sich nicht wie Muttersprachlerinnen bzw. Muttersprachler ausdrücken, schlechter – ein unbeabsichtigtes Bias, das auf bestimmten Vorstellungen von „sprachlicher Kompetenz" beruht.

Warum ist die Voreingenommenheit in KI-Systemen ein Problem?

Voreingenommenheit in KI-Systemen ist ein gesellschaftlich relevantes Problem. Wenn KI-Systeme nicht kontrolliert werden, können sie Diskriminierung fördern, Vertrauen untergraben und das Potenzial der Technologie begrenzen.

Beispiel: Ein KI-System soll Bewerbungen vorsortieren und wurde mit älteren Daten trainiert, in denen vor allem Männer für technische Berufe eingestellt wurden. Deshalb bevorzugt die KI nun automatisch männlich klingende Namen und stuft Bewerbungen von Frauen schlechter ein - unabhängig von ihrer tatsächlichen Qualifikation. Dadurch wird bestehende Diskriminierung nicht abgebaut, sondern verstärkt, und Betroffene werden ungerecht behandelt.

Faire, ausgewogene KI-Systeme haben das Potenzial, Bereiche wie Justiz, Gesundheit und Bildung grundlegend zu verbessern – vorausgesetzt, wir setzen uns kritisch mit ihren Schwächen auseinander. Aber kann KI jemals frei von Voreingenommenheit sein, oder wird sie stets menschliche Schwächen spiegeln?

Können KI-Systeme neutral sein?

Absolute Neutralität ist in der KI wahrscheinlich ein unrealistisches Ziel, da sie ein menschliches Produkt ist und von menschlichen Zielsetzungen und Denkweisen geprägt wird. Durch technische Korrekturen und ethische Aufsicht kann KI jedoch weniger voreingenommen gestaltet werden.

- Breite und repräsentative Datensätze: Entwicklerinnen und Entwickler einer KI sollten immer darauf achten, dass die gesammelten Datensätze möglichst vielfältig und repräsentativ sind, um ein ausgewogeneres Bild einer Situation zu erhalten.
- Testverfahren für Fairness: Es gibt Tools, mit denen sich KI-Systeme vor ihrem Einsatz auf Fairness prüfen lassen, z. B. *AI Sandbox* (*Luxembourg Institute for Science and Technology, LIST*).
- *Explainable AI* (*XAI*): „Erklärbare KI"-Techniken sorgen dafür, dass jede Entscheidung im Machine-Learning-Prozess[2] nachvollzogen und erklärt werden kann.

Ohne diese Methoden agieren viele KI-Modelle als „*Blackbox*", deren interne Abläufe selbst für Entwicklerinnen und Entwickler oft undurchsichtig sind. Dadurch wird die Überprüfung der Ergebnisse erschwert und es gehen Kontrolle sowie

[2] *Machine Learning* (ML, „maschinelles Lernen") ist ein Teilbereich der Künstlichen Intelligenz. Dabei lernen Computersysteme aus Beispieldaten, erkennen Muster und treffen auf dieser Grundlage Entscheidungen, ohne dafür explizit programmiert worden zu sein.

Verantwortbarkeit verloren. Mit XAI erhalten Unternehmen Einblick in die Entscheidungsgrundlagen der KI und können Fehler schneller erkennen und beheben.

Inspirationen für den Unterricht

Experiment

Die Schülerinnen und Schüler testen eigenständig Chatbots oder KI-Bildgeneratoren mit bestimmten Eingaben (*Prompts*). Sie vergleichen die unterschiedlichen Ausgaben und überlegen, wo sich (unbewusste) Voreingenommenheit zeigt.

Zeigen Sie, dass ein präzise formulierter *Prompt* eine voreingenommene Antwort vermeiden kann.

Beispiel: Lassen Sie die Schülerinnen und Schüler folgende *Prompts* in einem KI-Bildgenerator (z. B. fobizz) eingeben:

- Eine erfolgreiche Führungsperson in einem Büro.
- Eine Arbeitskraft im Kindergarten umgeben von Kindern.

Besprechen Sie die Resultate: Wie werden Merkmale wie Geschlecht, Ethnie, Körperbau und Alter dargestellt? Welche Bilder erscheinen voreingenommen? Wie könnten die *Prompts* verbessert werden, um diese Voreingenommenheit zu vermeiden?

Debatte

Teilen Sie die Schülerinnen und Schüler in Pro- und Contra-Gruppen ein und organisieren Sie eine Debatte zum Thema: „Soll KI in der Schule eingesetzt werden, obwohl sie nicht neutral ist?“ Die Teams sollen dabei insbesondere die Aspekte Fairness und Bildungsgerechtigkeit berücksichtigen.

Fallstudie zu KI-Überwachungssystemen

Präsentieren Sie reale Beispiele von KI-Überwachung in verschiedenen Ländern. Die Schülerinnen und Schüler vergleichen die verschiedenen Modelle, erörtern die öffentlichen Reaktionen und besprechen die ethischen Bedenken.

Simulation im Klassenzimmer

Lassen Sie die Schülerinnen und Schüler ein fiktives KI-Überwachungssystem für eine Schule entwerfen. Weisen Sie ihnen verschiedene Rollen zu (z. B. Schuldirektion, Schülerinnen und Schüler, Bürgerrechtlerinnen und Bürgerrechtler) und lassen Sie sie die möglichen Vor- und Nachteile auf der Grundlage ethischer und praktischer Überlegungen diskutieren.

Entwurf einer ethischen Praxis

Lassen Sie die Schülerinnen und Schüler in Gruppen arbeiten, um ethische KI-Überwachungsrichtlinien zu entwickeln, die Fairness, Transparenz und Verantwortlichkeit zu berücksichtigen. Jede Gruppe stellt ihre Richtlinien vor und erörtert mögliche Herausforderungen.