Mögliche Antworten zur Diskussionsaufgabe

Welches Problem wird in den jeweiligen Szenarien beschrieben?

In manchen Situationen weichen die Vorhersagen von KI-Verfahren von der Erwartung ab bzw. benachteiligen Individuen oder bestimmte Personengruppen.

- **Szenario 1** beschreibt dieses Phänomen bei der Bewerberauswahl. Frauen erhalten deutlich seltener ein Jobangebot der angesprochenen Firma als Männer.
- **Szenario 2** beschreibt, dass kommerzielle Gesichtserkennungssysteme zwar bei weißen Personen sehr gut funktionieren, gerade schwarze Frauen jedoch nur schlecht erkannt werden.
- **Szenario 3** beschreibt den Fall, dass die Abschlussnoten der Schülerinnen und Schüler, die über Studienplätze und Ausbildungschancen entscheiden, deutlich von den Bewertungen durch die Lehrkräfte abwichen und dabei eigentlich nicht relevante Merkmale wie der Schulbezirk eine Rolle spielten.

Welche Ursache(n) könnten die problematischen Entscheidungen der KI-Systeme in den beschriebenen Szenarien haben? / Erläutere unter Einbeziehung der verwendeten Daten, warum es zu diesen Problemen kommen konnte.
Ursache für diese Entscheidungen sind oftmals Datensätze, die gewisse Verzerrungen (Bias) aufweisen.

- **Szenario 1:** So wurden für das Bewerberscreening vor allem die historischen Daten von Bewerberinnen und Bewerbern herangezogen. Wurden in der Vergangenheit aber vor allem männliche Bewerber eingestellt, wird diese Entscheidung durch das Modell verfestigt, da es gelernt hat, dass Bewerbungen von Frauen schlechter seien (Historischer Bias).
- **Szenario 2:** Die Bilddatenbanken, die für das Training der Gesichtserkennungssoftware verwendet wurden, weisen vermutlich einen deutlich höheren Anteil an Bildern von weißen Männern als von schwarzen Frauen auf. Die Ursache ist in diesem Fall also in der Auswahl der Daten zu suchen, die von der realen Verteilung der Nutzer abweicht (Selection Bias).
- **Szenario 3:** Im dritten Fall wurden Merkmale berücksichtigt und ein Zusammenhang zwischen dem Schulbezirk und der Abschlussnote hergestellt. Auch wenn ein solcher Zusammenhang prinzipiell bestehen kann (Korrelation), resultiert aus der Zugehörigkeit zu einem Schulbezirk noch lange nicht die Schulnote (keine Kausalität). Der Unterschied zwischen Korrelation und Kausalität wird auch im folgenden Cartoon verdeutlicht.

Korrelation vs. Kausalität (CC-BY-SA Haubert, Seegerer, Albrecht)

Die Verwendung verzerrter Daten (Daten, die die Realität nur bedingt widerspiegeln), kann dabei sowohl absichtlich als auch unabsichtlich bzw. unbewusst erfolgen.

Basierend auf deinen Erfahrungen mit dem Training von Modellen: Wie könnte man diese Probleme lösen?

Es gibt verschiedene Möglichkeiten, diese Probleme zu adressieren oder gar zu lösen. Ein Hauptaugenmerk sollte auf einer geeigneten Auswahl der Trainings- aber auch der Testdaten liegen. Je besser Trainingsdaten die Realität wiedergeben, desto zuverlässiger sollte das Modell in der Praxis funktionieren. Weiterhin denkbar wäre, das System in weiteren möglichst vielfältigen Situationen und mit unterschiedlichen Daten zu testen. Außerdem sollte den Ergebnissen des Modells nicht blind getraut, sondern die Ergebnisse kritisch hinterfragt werden.

Konkret für die verschiedenen Szenarien wäre Folgendes denkbar:

- **Szenario 1:** Weglassen persönlicher Informationen, die nicht für die Einstellung relevant sind (wie Alter, Foto oder Geschlecht). Alternativ können Manche Merkmale der eingestellten Personen austauschen/variieren (bspw. Hobbies zufällig ersetzen, um so für ein “Rauschen” zu sorgen); nicht ausschließlich auf die Entscheidung des Modells verlassen, sondern manuell prüfen.
- **Szenario 2:** Mehr Trainingsdaten von schwarzen Frauen verwenden; Abgleich der Personengruppe der Nutzerinnen und Nutzer mit der Personengruppe, die auf den Trainingsbildern abgebildet ist.
- **Szenario 3:** Lehrkräfte in die Entscheidung einbinden; nicht persönliche Merkmale wie den Schulbezirk entfernen und nur persönliche Merkmale berücksichtigen.

Wichtig: Die konkreten Ideen stellen nur eine Auswahl der vorhandenen Möglichkeiten dar! Die Schülerinnen und Schüler können noch viele weitere sinnvolle Ideen entwickeln.

Welche Regeln sollten für einen fairen Einsatz von KI-Systemen gelten?

Hier sind eine Vielzahl an Möglichkeiten denkbar. So könnte man Regeln vorschlagen, die Ergebnisse nicht automatisch zu akzeptieren, sondern stets einen menschlichen Entscheider zurate zu ziehen, wenn das System für Entscheidungen mit großer Tragweite herangezogen wird. Außerdem könnte man Regeln zu den verwendeten Datensätzen vorschlagen, sodass ein ausgeglichener Datensatz Voraussetzung für das Modell würde.