Unterrichtsbaustein ‚Diskriminierung – KI als Lösung oder als Problem?'

Materialien zum Baustein

- M1 *Drivesmarter* hat ein Problem
- M2 Was, wenn KI über uns mitbestimmt?
- M3 Diskriminierung: KI als Problem oder als Chance?
- M4 Wie kann es zu Diskriminierung durch KI kommen?
- M5 (Wie) Lässt sich KI fair einsetzen?

*Mit * markierte Begriffe finden sich im Glossar zum Thema.*

M1 *Drivesmarter* hat ein Problem

Stell dir vor, du bist Personaler:in bei *Drivesmarter*, einem international aufgestellten Unternehmen mit 5000 Mitarbeiter:innen in Deutschland, das klimaneutrale und fair hergestellte E-Autos verkauft. Die Firmenleitung hat bei der internen Evaluation ein eklatantes Problem festgestellt: In der Vergangenheit gab es ganz offensichtlich deutliche Diskriminierungen bei der Auswahl von Mitarbeiter:innen in der technischen Abteilung für Entwicklung. Immer wieder wurden Frauen bei gleicher Qualifikation männlichen Konkurrenten gegenüber benachteiligt. Immer wieder wurden *weiße* Personen schon bei der Vorauswahl bevorzugt für eine Stelle ausgewählt, obwohl sie nicht über bessere Fähigkeiten verfügten. Den Konzernvorstand macht das Problem ziemlich nervös: Die Firma lebt von Innovationen und braucht nur die besten Mitarbeiter:innen.

Du wirst vom Vorstand beauftragt, das Problem effektiv zu lösen. Ansetzen sollst du bei der notwendigen Vorauswahl der stets sehr großen Zahl an Bewerber:innen auf der Basis ihrer Anschreiben mit Lebenslauf und Foto. Es muss entschieden werden, wer überhaupt zu einem Bewerbungsgespräch für einen jeweiligen Posten in der Abteilung für Entwicklung eingeladen werden soll. Das Ziel besteht darin, in der nächsten Zeit effektiv die tatsächlich am besten geeignete Person für die jeweilige Stelle zu finden.

Für deine Arbeit bekommst du von der technischen Abteilung Hilfe angeboten: Du hast die Wahl, drei erfahrene Mitarbeitende aus der Firma zu wählen, die dich bei der Vorauswahl unterstützen. Alternativ könntest du aber auch eine KI* zur Verfügung gestellt bekommen, die dir bei der Auswahl hilft, indem sie eine Vorauswahl der passenden Bewerber:innen trifft.

Wofür entscheidest du dich?

Aufgaben

1. Schreibe einen Spickzettel für ein kurzes Statement, in dem du deine Entscheidung für die Unterstützung durch andere Menschen oder durch eine KI dem Vorstand mitteilst und begründest. Nimm dabei auf Eigenschaften und Fähigkeiten von Menschen und von KI Bezug.
2. In Gruppen: Präsentiert euch gegenseitig eure Statements. Diskutiert anschließend eure Begründungen.
3. Kommt zu einem Urteil darüber, ob der Vorstand von *Drivesmarter* die Vorauswahl für die Einstellung von Mitarbeiter:innen mit der Hilfe mehrerer Menschen oder einer KI treffen soll.

M2 Was, wenn KI über uns mitbestimmt?

Das Beispiel des Unternehmens Drivesmarter war fiktiv. Doch tatsächlich werden KI-Systeme bereits in einigen Unternehmen und in vielen anderen gesellschaftlichen Bereichen als Unterstützung für die Entscheidungsfindung eingesetzt: die sogenannten ADM-Systeme („Automated/Algorithmic Decision Making Systems"). Besonders, wenn von den Entscheidungen sehr viel abhängt, ist der KI-Einsatz umstritten. In einer Stellungnahme des Deutschen Ethikrats von 2023 werden einige dieser Systeme (Stand der technischen Entwicklung zum Zeitpunkt der Veröffentlichung) beispielhaft vorgestellt und besprochen.

Aufgaben

1. Benennt mögliche Gründe, aus denen der Deutsche Ethikrat sich mit dem Einsatz von ADM-Systemen befasst haben könnte.
2. Bildet Dreiergruppen und verteilt die drei Textauszüge a) bis c) unter euch. Lest sie in Hinblick auf die Zwecke, die mit den erläuterten ADM-Systemen verfolgt werden.
3. Sucht euch eine Person aus den anderen Dreiergruppen, die denselben Text gelesen hat wie ihr, und besprecht ihn gemeinsam.
4. Erläutert euch in eurer ursprünglichen Dreiergruppe gegenseitig, welche Zwecke mit den beschriebenen ADM-Systemen erreicht werden sollen.
5. Diskutiert, ob und wenn ja, in welchen Bereichen der Einsatz der ADM-Systeme gesellschaftlich sinnvoll ist und ob ihr die Entscheidungen solcher Systeme in eurem eigenen Leben akzeptieren würdet.

a) Kindeswohlgefährdung

ADM-Systeme im Kontext der Entscheidungsfindung bei Kindeswohlgefährdung sind mittlerweile auf nahezu allen Kontinenten wenigstens in ersten Schritten eingeführt. Manche dieser Anwendungen sind hoch umstritten, etwa aufgrund von Ungenauigkeit, Fehleranfälligkeit oder diskriminierenden Effekten.

In Deutschland befindet sich die Einführung von ADM-Systemen im Bereich der Kinder- und Jugendhilfe zur Gefährdungsabschätzung des Kindeswohls in der Planungs- und Diskussionsphase. Bereits etablierter Softwareeinsatz im Bereich des Kindesschutzes schafft dazu die nötigen Voraussetzungen. […]

In den Niederlanden starteten bereits Ende der 2000er-Jahre Versuche, das in den USA entwickelte *California Family Risk Assessment* für den niederländischen Kinderschutz nutzbar zu machen und auf seine prognostische Validität[1] hinsichtlich des Risikos für Gewalt oder Vernachlässigung von Kindern in problemaffinen Familien hin zu überprüfen. Dieses algorithmenbasierte *Assessment-Tool* zielt auf die Stärkung standardisierter bzw. strukturierter Fallführung (*structured decision making*). Es kombiniert zehn verschiedene Items, die das Risiko einer zukünftigen Vernachlässigung indizieren[2], mit weiteren zehn, die das Risiko eines zukünftigen Missbrauchs anzeigen sollen. Neben Aspekten wie unsichere

[1] prognostische Validität (grob): Zuverlässigkeit in den Vorhersagen
[2] indizieren: ein Anzeichen für etwas sein

Wohnverhältnisse oder früher eingetretene Gefährdungssituationen beziehen sich viele Items auf die (elterlichen) Sorgeberechtigten (z. B. seelische Gesundheit, Suchtprobleme, Rechtfertigung von Missbrauch durch den Sorgeberechtigten, Dominanz und Strenge). Studien ergaben eine beachtliche Treffsicherheit dieses Prognoseinstruments für ein frühzeitiges Erkennen von Gefährdungslagen.

Demgegenüber legen Studien aus Neuseeland und den USA nahe, dass sich die Erwartungen an ADM-Systeme mit Blick auf eine verbesserte prädiktive Analytik nicht im gewünschten Umfang erfüllen. Das *Predictive Risk Modelling* des neuseeländischen *Centre for Applied Research in Economics* und das *Allegheny Family Screening Tool* in Pennsylvania (USA) erweisen sich etwa aufgrund der Fehlerhaftigkeit der genutzten bzw. algorithmisch verarbeiteten Items als zu wenig präzise oder sogar als offensichtlich verfälschend und diskriminierend, was die Spielräume für die Lebensgestaltung der Betroffenen erheblich vermindert.

b) Bewährungshilfe

[…] In der Schweiz sind ADM-Systeme im Rahmen des Risikoorientierten Sanktionenvollzugs (ROS) in Vollzugs- wie Bewährungsdiensten bereits etabliert. ROS sieht die Abklärung des individuellen Rückfallrisikos einer gewalttätigen Person im Rahmen eines vierstufigen Prozesses vor, der erstens aus einer Triage[3], zweitens aus einer genaueren Abklärung des Falles, drittens aus der Planung etwaiger Interventionsinstrumente[4] und viertens aus dem kontinuierlich ausgewerteten Verlauf besteht. Kern der ersten Stufe, in der die Dringlichkeit (Triage) einer genaueren Risikoabschätzung abgeklärt wird, bildet ein Fall-Screening-Tool (FaST), das als ADM-System die jeweiligen Fälle drei Kategorien zuordnet: Die Fälle der Kategorie A mit einem geringen Rückfallrisiko benötigen keinerlei weitere Risikoprognosen; die Fälle der Kategorie B mit einem erhöhten Rückfallrisiko benötigen als weiteren Risikoprognoseschritt eine Kurzabklärung, die die fallführende Fachkraft mittels einer Checkliste aus verschiedenen Quellen (z. B. Strafregister, Gutachten) erstellt (Fallresümee); die Fälle der Kategorie C mit einem hohen Rückfallrisiko bedürfen der präzisen Beurteilung aller verfügbaren Informationen durch eine forensisch[5] geschulte psychologische Fachkraft (Risikoabklärung) im Sinne einer strukturierten Urteilsbildung. Das FaST selbst kombiniert und gewichtet unterschiedliche Items und klassifiziert den jeweils anstehenden Fall auf der Basis statistisch[er] [...] Verfahren.

Bereits seit Anfang der 2000er-Jahre kommt in mehreren US-Bundesstaaten die aufgrund diskriminierender Prognosen sehr umstrittene Software COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*) zum Einsatz. Sie kann sowohl bei der richterlichen Strafzumessung wie beim nachfolgenden Strafvollzug oder beim späteren Bewährungsdienst zur Anwendung kommen. Es wurde im Bemühen einer evidenzbasierten Entscheidungsfindung eingeführt, um in den verschiedenen Phasen eines Fallverlaufes sowohl aufseiten der entscheidenden Person das Ungewissheitspotenzial als auch aufseiten der Bevölkerung das Gefährdungspotenzial durch möglicherweise rückfällige gewalt-

---

[3] Triage (hier): Sichtung, Einteilung

[4] Interventionsinstrumente (hier): Möglichkeiten, bei einem Rückfall einzuschreiten

[5] forensisch: gerichtlichen oder kriminologischen Zwecken dienend

tätige Personen zu verringern. COMPAS selbst kommt sowohl im Kontext der Untersu-
27 chungshaft wie auch während und nach der Haftzeit (Bewährung) zur Anwendung. Es kombiniert 137 Items und differenziert damit drei Grobklassifizierungen: niedriges, mittleres und hohes Rückfallrisiko. COMPAS umfasst neben einer Risikoprognose auch ein da-
30 rauf aufbauendes *needs assessment*[6], das in die weitere Interventionsplanung eingeht.

c) *Predictive Policing* (vorhersagende Polizeiarbeit)

In Deutschland kommen bislang in erster Linie raumbezogene Verfahren des *Predictive Policings* zum Einsatz, die durch die Ausweisung prädiktiver[7] Risikogebiete mit zeitlicher Prä-
3 ferenz gekennzeichnet sind. In der Hauptsache richten sich die Verfahren auf die Vorhersage raumzeitlicher Parameter bei Wohnungseinbruchdiebstählen. […] Personenbezogene Verfahren des *Predictive Policings* stützen die Prognose auf Täter- bzw. Opfermerkmale. Zu
6 nennen ist beispielhaft das in Großbritannien eingesetzte Programm HART sowie die sogenannte *Strategic Subject List* der Polizei von Chicago (USA). […] Seit 2017 wird in Hessen das Programm hessenDATA, basierend auf der Analysesoftware *Gotham* des US-amerika-
9 nischen Unternehmens *Palantir Technologies*, im Rahmen der Terrorismusbekämpfung auf der Grundlage von § 25a Abs. 1 Alt. 1 des Hessischen Gesetzes über die öffentliche Sicherheit und Ordnung (HSOG) eingesetzt. Es nutzt Informationen aus polizeilichen Datenban-
12 ken, Verkehrsdaten aus der Telekommunikationsüberwachung und von Telekommunikationsanbietern zur Verfügung gestellte Daten. Einbezogen werden außerdem sogenannte forensische Extrakte wie zum Beispiel die Ergebnisse der Beschlagnahme eines Mobiltele-
15 fons und Informationen aus sozialen Netzwerken. […] Nachdem das Bundesverfassungsgericht unter anderem die hessische Regelung in § 25a Abs. 1 Alt. 1 HSOG für verfassungswidrig erklärt und eine Neuregelung verlangt hat, weil sie keine ausreichende Eingriffs-
18 schwelle enthalte, müssen die einschlägigen Ermächtigungsgrundlagen auch in anderen Gesetzen an die vom Bundesverfassungsgericht formulierten Vorgaben angepasst werden.

Quelle: Deutscher Ethikrat (2023): *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. URL: https://www.ethikrat.org/publikationen/stellungnahmen/mensch-und-maschine/, 310–313 und 320–322, Kursivierungen hinzugefügt.

---

6 *needs assessment*: Bedarfseinschätzung

7 prädiktiv (hier): vorhergesagt

M3 Diskriminierung: KI als Problem oder als Chance?

Interview mit Vincent Müller (2024)

Hört oder lest den ersten Teil des Interviews mit dem Philosophen Vincent Müller zu Problemen und Chancen von KI in Bezug auf Diskriminierung (Auszug 1, vom Beginn bis Minute 7:34) und bearbeitet dazu die folgenden Aufgaben.

Aufgaben

1. Fasse zusammen, was Müller zufolge ‚Diskriminierung' bedeutet.
2. Stelle die beiden Weisen dar, auf die Diskriminierung durch KI zustande kommen kann. Diskutiert gemeinsam, was damit jeweils gemeint sein könnte.
3. Erläutere, was laut Müller Stärken und Schwächen von KI sind, wenn es um die Vermeidung von Diskriminierung geht.

Auszug 1

Interviewer*:* Ich bin hier mit Professor Dr. Vincent Müller, der Professor ist an der Universität Erlangen-Nürnberg für Philosophie der Künstlichen Intelligenz, und wir sprechen heute über KI im Zusammenhang mit Diskriminierung. Die erste Frage ist, was überhaupt sensible Bereiche sind, in denen künstliche Intelligenz eingesetzt wird und in denen Probleme mit Diskriminierung auftreten könnten.

Müller*:* Ja, da sollte man vielleicht unterscheiden zwischen den Bereichen, in denen KI gegenwärtig schon eingesetzt wird und denen, in denen KI potenziell eingesetzt werden kann. Gegenwärtig ist ja der Einsatz ziemlich beschränkt. Es gibt ein paar Bereiche, in denen das relativ offensichtlich ist, zum Beispiel die Auswahl von Kandidaten für Stellen. Da wird bereits ausgiebig mit KI gearbeitet, weil die Firmen, die sich damit beschäftigen, die Masse an Bewerbern sozusagen auf eine sehr viel kleinere Zahl herunterbrechen wollen, um dann diese Leute etwas genauer anzuschauen. Und es ist offensichtlich, dass das ein Bereich ist, in dem Diskriminierung eine erhebliche Schwierigkeit darstellt. So grob gesagt, würde ich sagen, Diskriminierung bedeutet, dass man ein Urteil nach Kriterien stellt, die irrelevant sind für das Problem, nach dem man eigentlich sucht. Also man sollte zum Beispiel bei einer Einstellung den besten Kandidaten oder die beste Kandidatin auswählen und dafür sind bestimmte Faktoren zum Beispiel in der Regel irrelevant, also zum Beispiel Geschlecht oder Hautfarbe. Und andere Faktoren sind in bestimmten Fällen irrelevant. Also zum Beispiel für eine akademische Tätigkeit ist es irrelevant, ob jemand im Rollstuhl sitzt oder nicht. Für andere Jobs wird das vielleicht nicht zutreffen.

Ich sehe aber allgemein die Gefahr, dass diese Art von Diskriminierung weiter um sich greift, wenn man KI weiter einsetzt, weil: Ein wesentlicher Einsatzbereich der KI ist der, Menschen Entscheidungen abzunehmen oder Entscheidungen zu erleichtern. In beiden Fällen werden ja die Entscheidungen entweder von der KI gefällt oder eben so in eine bestimmte Richtung geschoben. Und wenn diese Richtung so ist, dass sie eben Diskriminierung in dem vorhin erwähnten Sinne impliziert, dann gibt es da Schwierigkeiten.

Es gibt da, glaube ich, zwei Arten von Diskriminierung, die dabei eine Rolle spielen. Die eine ist die sogenannte historische Diskriminierung. Also wenn man eine KI auf Daten trainiert, die bereits Diskriminierung enthalten, dann ist es wahrscheinlich, dass die KI das lernt. Also wenn sie zum Beispiel lernt, dass in diesen Jobs Männer sehr viel bessere Chancen haben als Frauen, dann wird die KI lernen, dass das offensichtlich ein wichtiger Aspekt für die Auswahl der Personen ist und wird ebenfalls Männer vorziehen. Und es gibt noch einen weiteren Punkt, der aber schwieriger ist, wenn die KI selbst sozusagen Diskriminierung ausbrütet. Und das ist natürlich deswegen schwierig, weil man nicht richtig weiß, wie eigentlich die KI die Entscheidung fällt. Man kann das eigentlich nur an Testsets von Daten überprüfen. Kommt da jetzt ungefähr das raus, was man sich so vorstellt, und dann sagt man, ja, das sieht okay aus nach verschiedenen solchen Tests und kann dann das System einsetzen. Man wird aber eben ganz oft kleinere Fehler auf diese Art und Weise nicht entdecken.

Interviewer*:* Dankeschön. Das heißt – also eine Sorge in diesem Kontext ist auch, dass es so einen selbstverstärkenden Kreislauf geben könnte, richtig? Also dass einerseits Menschen Trainingsdaten erzeugen, die Diskriminierung manifestieren, zum Beispiel rassistisch geprägte Trainingsdaten, dass dann KI-Systeme das reproduzieren und dadurch wieder menschliche Diskriminierung befördern und es so einen selbstverstärkenden Kreislauf gibt. Ist das ein realistisches Szenario oder ist das ein Grund davon auszugehen, dass KI-basierte Diskriminierung noch zu deutlich mehr Problemen führen wird, als beispielsweise, ja, Diskriminierung durch Menschen oder Diskriminierung in Situationen, wo es keine KI gäbe?

Müller*:* Mir scheint, dass das nicht eine Situation ist, in der die KI die Lage sozusagen verschlechtert. Weil diese selbstverstärkenden Diskriminierungssysteme sind ja solche, die eben bei Menschen auch auftreten. Also wenn wir alle als weiße Männer in so einer Kommission sitzen, die irgendwelche Leute einstellt, dann finden wir – nicht natürlich, sondern wir finden vielleicht de facto – Leute, die so sind wie ich oder so sind wie wir, irgendwie plausibler für die Stelle. Und wenn wir das schon lange, Jahrzehnte so gemacht haben, dann wird das auch gar nicht weiter hinterfragt als irgendeine Art von Problem. Es ist einem ganz oft bei Diskriminierung eben nicht bewusst, dass man eine Diskriminierung durchführt. Das ist genau die Schwierigkeit. Man würde ja, wenn man die Leute, die – sagen wir, vor vielleicht ein paar Jahrzehnten jedenfalls – ganz offensichtlich gegenüber Frauen und Leuten mit dunkler Hautfarbe diskriminiert haben, wenn man die gefragt hätte, ob sie diskriminieren, dann hätten die natürlich gesagt, das tun sie nicht, sie suchen nur die besten Kandidaten – und ich bin mir sicher, sie hätten wahrscheinlich gesagt, wir suchen *den* besten Kandidaten. Und hätten das völlig unproblematisch gefunden.

Ich würde sagen, man könnte vielleicht sogar erhoffen, dass die KI in bestimmten Bereichen Diskriminierung aufdeckt, die wir bisher gar nicht durchschaut hatten, eben weil man die KI ja relativ leicht testen kann, an relativ großen Datenmengen. Und dann könnte man eben zum Beispiel herausfinden, – das ist kein fiktionales Beispiel – dass Leute, die in einer bestimmten Gegend wohnen, schärfer von der Polizei kontrolliert werden. Obwohl die Polizei selbst behauptet, dass sie das nicht tut. Das stellt sich aber heraus, aufgrund bestimmter taktischer Entscheidungen, die da regelmäßig gefällt werden, dass das der Fall ist. Oder

dass zum Beispiel – in den USA ist das bekannt – Schwarze sehr viel häufiger kontrolliert werden als Weiße. Das erzeugt eine gewisse Selbsterzeugung, weil natürlich klar ist: Wenn
72 man sehr viel mehr Schwarze kontrolliert als Weiße, dann findet man auch sehr viel mehr Schwarze, die irgendein Vergehen gemacht haben. Also verstärkt sich das wieder, dass man sagt, aha, gut, dass wir die kontrolliert haben. Also verstärkt sich das. Also eigentlich ist die
75 Möglichkeit der KI, eben große Datenmengen zu Testzwecken zu checken und eben diese Kriterien herauszufinden, eigentlich recht gut. Man kann eben dann sehen, aha, in unserer Datenbank stellt sich heraus, dass plötzlich irgendein Faktor einen ganz großen Unter-
78 schied macht, von dem wir eigentlich nicht finden, dass er einen Unterschied machen sollte. Also der ist eben nicht relevant für die Entscheidung und daher ein Faktor, den man als Diskriminierung einstufen sollte. Mir scheint also, da gibt es eigentlich relativ positives Po-
81 tenzial. Aber eben: Zunächst mal würde man erwarten, dass eine KI, die aufgrund unserer Erfahrungen und auch unserer Anforderungen programmiert wird, eben die Diskriminierung, die in unserer Gesellschaft bereits vorhanden ist, auch widerspiegelt.

Aufgaben

Hört oder lest den zweiten Teil des Interviews (Auszug 2, von Minute 7:34 bis zum Ende).

4. Erläutert, wie laut Müller der Einsatz von KI Diskriminierung entgegenwirken kann.
5. Müller sagt abschließend: „Was die Frage fairer Entscheidungen angeht, überwiegen die positiven Möglichkeiten." Diskutiert diese Einschätzung.

Auszug 2

Interviewer*:* Dankeschön. Kannst du vielleicht noch näher beschreiben, was so Möglichkeiten sind, wie man KI auf eine Weise nutzen könnte, um Diskriminierung, wie zum Bei-
3 spiel rassistische Diskriminierung, zu verringern und nicht zu verschärfen, wenn es also dieses positive Potenzial, das du ansprichst, gibt?

Müller*:* Ja, ich denke, die KI hat einfach grundsätzlich das Potenzial, unsere Entscheidun-
6 gen rationaler zu machen und deutlicher aufzudecken, welche Kriterien wir eigentlich bei diesen Entscheidungen zum Einsatz bringen, weil es darüber Daten gibt, weil es oft viele Daten gibt und weil man eben diese Daten sehr viel leichter analysieren kann, als das, was
9 man so sonst tagtäglich macht. Es geht ja sonst eigentlich niemand hin und sagt eigentlich: Oh, warum haben wir, sagen wir mal, in den letzten Jahrzehnten an dieser Universität solche Leute eingestellt? Das ist ja extrem schwierig und meistens wird sowas nicht gemacht.
12 Wenn man aber eben einen Datensatz hat über irgendwelche Entscheidungen, die da gefällt wurden von der KI oder die die KI vorschlagen würde als Entscheidungen, dann lässt sich das sicher leichter entdecken, weil man zum Beispiel den Datensatz systematisch va-
15 riieren kann. Und dann kann man systematisch eben rausfinden, was passiert. Also wenn man, weiß ich nicht, man macht alle Kandidaten ein bisschen älter oder irgendwie so etwas, man verändert systematisch irgendwelche Parameter, dann kann man rausfinden, dass da
18 eigentlich ein Parameter eine Rolle spielt in der Entscheidung, von dem wir gar nicht erwartet hätten, dass er eine Rolle spielen soll.

Interviewer*:* Danke sehr. Letzte Frage: Was wäre deine allgemeine Erwartung? Also denkst du, dass auf lange Sicht künstliche Intelligenz eher dazu beitragen wird, die Welt fairer zu machen und Diskriminierung abzubauen, wenn man so diese verschiedenen Chancen und Risiken berücksichtigt? Oder überwiegen die Gefahren?

Müller*:* Ich würde sagen, was die Frage fairer Entscheidungen angeht, überwiegen die positiven Möglichkeiten. Aber das ist natürlich nicht das ganze Feld der Probleme, die möglicherweise die KI zukünftig erzeugen wird.

Interviewer*:* Dankeschön.

M4 Wie kann es zu Diskriminierung durch KI kommen?

Es wird immer wieder kritisiert, dass einige Technologien bestimmten Personengruppen nutzen und anderen schaden. Diese Diskussion wird auch in Bezug auf KI-Systeme geführt. In den folgenden Textauszügen des Philosophen Jens Kipper und der Computerwissenschaftlerin Timnit Gebru geht es um Erklärungen dafür, wie es zu diskriminierenden Ergebnissen bei der Entwicklung und Verwendung von KI und anderen Technologien kommen kann.

Aufgaben

1. Benenne mögliche Erklärungen dafür, warum der Einsatz von KI zu Diskriminierung führen kann.
2. Erläutere, welche drei Probleme Kipper zufolge in Bezug auf Daten und deren Interpretation dazu führen, dass der Einsatz von KI diskriminierende Effekte haben kann.
3. Gebru vertritt die These, dass Diskriminierung durch Technik kein Phänomen ist, das erst mit der Entwicklung von KI entstanden ist. Erläutere anhand von Beispielen aus dem Text die Diskriminierung, die laut Gebru bei der Entwicklung von Technik allgemein entstehen kann.
4. Gebru spricht in ihrem Text vom „Einsatz von Technologie als Waffe gegen bestimmte Gruppen“. Nimm Stellung zu dieser Redeweise und der damit verbundenen Wertung.

Jens Kipper (2020): Erklärungen für Diskriminierungen durch KI-Einsatz

Es zeigte sich [im Fall der Auswahl von Mitarbeitenden durch eine KI bei Amazon], dass der Algorithmus Frauen systematisch benachteiligte, indem er beispielsweise Bewerber schlechter bewertete, wenn in ihren Bewerbungen der Ausdruck ‚Frauen‘ auftauchte oder wenn sie an reinen Frauen-Unis ihren Abschluss gemacht hatten [...]. Was war passiert? Die beteiligten Forscher hatten ein KNN[8] auf der Grundlage von Bewerbungsunterlagen früherer Kandidaten trainiert. Wenig überraschend, waren bei vorangegangenen Stellenausschreibungen die meisten erfolgreichen Bewerber Männer, und so kam das KNN zur Auffassung, dass Frauen grundsätzlich schlechtere Kandidaten sind. Amazon gab schließlich das gesamte Projekt der KI-geleiteten Rekrutierung auf. Der Fall lehrt uns einiges über die Gefahren der Verwendung von KI.

Drei eng miteinander verwandte Probleme, d. h. Verzerrungen, sollen hier kurz erörtert werden. Erstens verdeutlicht das Beispiel, dass zumindest die heute existierenden KNNs zwar in der Lage sind, Korrelationen aufzudecken, aber nicht unbedingt Kausalbeziehungen. Dass Frauen in der Vergangenheit mit ihren Bewerbungen weniger erfolgreich waren, zeigt sicher nicht, dass irgendetwas am Frausein eine Person zu einer schlechteren Mitarbeiterin macht. Wahrscheinlicher ist (z. B.), dass Vorurteile auf Seiten der menschlichen Personalreferenten dazu führten, dass Frauen benachteiligt wurden.

---

[8] KNN: künstliches neuronales Netzwerk*

Zweitens zeigt das Beispiel, dass der Lernprozess von KNNs darauf angewiesen ist, dass ein vernünftiges Erfolgskriterium definiert wurde. Was Amazon suchte, waren fähige Mitarbeiter. Da es nicht leicht ist, diese zu identifizieren, trainierte man das KNN anhand eines deutlich leichter zu prüfenden Kriteriums, nämlich: Welche Bewerber bekommen einen Job? In diesem Sinne tat das KNN das, was man ihm aufgetragen hatte, aber nicht das, was Amazon wollte: Es identifizierte Bewerber, die gute Chancen haben, eine Stelle zu kriegen und nicht zwingend die, die gute Arbeit leisten werden.

Drittens zeigt der Fall, dass der Lernprozess von KNNs entscheidend von der Qualität der Datenbasis abhängt und sich insbesondere Verzerrungen in der Datenbasis in den Urteilen des KNNs widerspiegeln. Einfach ausgedrückt: Ein KNN, das mit Daten aus einem diskriminierenden System gefüttert wird, lernt, selbst diskriminierende Urteile zu fällen.

Quelle: Kipper, Jens (2020): Künstliche Intelligenz – Fluch oder Segen? Stuttgart: Metzler, 43f.

Timnit Gebru (2020): Wie Technik diskriminieren kann

Wissenschaft wird oft als eine objektive Disziplin auf der Suche nach Wahrheit gepriesen. Ebenso könnte man glauben, dass Technologie von sich aus neutral ist und dass Produkte, die von denjenigen entwickelt werden, die nur einen kleinen Teil der Weltbevölkerung repräsentieren, von jedem und jeder auf der Welt genutzt werden können. Eine Analyse des wissenschaftlichen Denkens im 19. Jahrhundert und der großen technologischen Fortschritte wie die Entwicklung von Autos, medizinischen Verfahren und anderen Disziplinen zeigt jedoch, wie die fehlende Repräsentation [sozialer Minderheiten] unter denjenigen, die die Macht haben, diese Technologie zu entwickeln, zu einem Machtungleichgewicht in der Welt und in der Technologie geführt hat, dessen beabsichtigte oder unbeabsichtigte negative Folgen denjenigen schaden, die bei der [Technologie-]Produktion nicht vertreten sind. Künstliche Intelligenz unterscheidet sich nicht [von anderen Technologien in Bezug auf diskriminierende Auswirkungen]. Auch wenn sich gängige Ansätze und Denkweisen stets verändern, hat die Dominanz derjenigen, die an ihrem Standort die mächtigste Ethnie sind (z.B. Weiße in den Vereinigten Staaten, ethnische Han in China usw.), in Verbindung mit der Machtkonzentration an einigen wenigen Orten auf der Welt zu einer Technologie geführt, die der Menschheit zugutekommen kann, die aber nachweislich (absichtlich oder unabsichtlich) systematisch diejenigen diskriminiert, die bereits marginalisiert sind. [...]

Der Einsatz von Technologie als Waffe gegen bestimmte Gruppen, ebenso wie deren Verwendung zur Aufrechterhaltung des Status quo bei gleichzeitiger Anpreisung als Befreierin der Machtlosen, ist nichts Neues im Falle von KI. In [ihrem Aufsatz] „Model Cards for Model Reporting" weisen Mitchell und Kolleg:innen auf Parallelen zu anderen Branchen hin, in denen Produkte für eine homogene Gruppe von Menschen entwickelt wurden. Angefangen bei Autos, die an Puppen mit prototypischen erwachsenen, „männlichen" Eigenschaften getestet wurden, was zu Unfällen führte, bei denen unverhältnismäßig viele Frauen und Kinder ums Leben kamen, bis hin zu klinischen Studien, die viele Personengruppen ausgeschlossen haben, was zu Medikamenten führte, die nicht wirken oder die überproportional negative Auswirkungen für Frauen haben, funktionieren Produkte, die

an einer homogenen Gruppe von Menschen entwickelt und getestet werden, am besten für diese Gruppe. [...]

Um auf eine KI hinzuarbeiten, die diejenigen nicht weiter ausgrenzt, die in der Vergangenheit ausgegrenzt wurden (und weiterhin ausgegrenzt werden), müssen sich das Bildungssystem und die allgemeine Einstellung von Forschenden und Praktiker:innen grundlegend ändern. Sie müssen sich vom Mythos der Leistungsgesellschaft und dem „Blick von Nirgendwo“ lösen.

Quelle: Gebru, Timnit: „Race and Gender”, in: Dubber, Markus D./Pasquale, Frank/Das, Sunit (Hg.): *The Oxford Handbook of AI*. Oxford: OUP, 253-268, hier: 253, 261 und 268, übersetzt von Anne Burkard.

M5 (Wie) Lässt sich KI fair einsetzen?

Erinnert euch zurück an das Problem von Drivesmarter: Der Firmenvorstand hat nun entschieden, KI für die Vorauswahl von Personal zu nutzen.

Aufgaben

1. Entwickelt Vorschläge für Kriterien dafür, wie die Vorauswahl von Personal unter Verwendung von KI möglichst diskriminierungsarm gestaltet werden kann.

   Bearbeitungshinweis: Greift dazu auf die Überlegungen der Philosophin Ting-An Lin und des Softwareingenieurs Po-Hsuan Cameron Chen zurück und berücksichtigt auch die bisherigen Materialien und Diskussionen des Bausteins. Fragt euch besonders, welche Daten genutzt werden können, um die KI zu entwickeln und welche Personen- und Berufsgruppen in den Entwicklungsprozess der KI einbezogen werden sollten.

2. Diskutiert, inwiefern die Nutzung der KI bei der Personalauswahl unter Berücksichtigung eurer Vorschläge moralisch vertretbar ist.

Ting-An Lin/Po-Hsuan Cameron Chen (2022): Empfehlungen für einen fairen KI-Einsatz[9]

Der vorherrschende Ansatz [zur KI-Fairness] verortet das Problem der KI-Bias[10] primär in Algorithmen und versteht das Ziel von KI-Fairness entsprechend darin, eine gewisse Gleichwertigkeit gewisser statistischer Maße [von KI-Outputs] zwischen verschiedenen Personengruppen zu gewährleisten. Dieser Ansatz zielt daher darauf ab, das Problem der KI-Bias hauptsächlich durch die Bereinigung von Bias in Algorithmen zu lösen. Trotz seiner Popularität haben Kritiker:innen auf die Grenzen dieses vorherrschenden Ansatzes zur KI-Fairness hingewiesen, einschließlich Hinweisen auf die Herausforderungen bei der Entscheidung zwischen verschiedenen statistischen Maßstäben, auf Bedenken, dass KI-Systeme als von sozialen Kontexten isolierte Entitäten untersucht werden, und auf die Überbetonung techno-zentrischer Antworten. Diese Überlegungen zeigen, dass ein umfassenderes ethisches Rahmenwerk erforderlich ist, um die ethischen Bedenken und Ansätze, die mit dem Umgang mit KI-Bias und dem Streben nach KI-Fairness verbunden sind, neu zu fassen. [...]

Wir schlagen vor, KI-Bias als eine Form der strukturellen Ungerechtigkeit aufzufassen, welche besteht, wenn KI-Systeme mit anderen sozialen Faktoren interagieren und so bestehende soziale Ungerechtigkeiten verschärfen. Dadurch werden manche Personengruppen stärker durch unverdiente Belastungen gefährdet. Aus dieser Perspektive sollte der Begriff der KI-Fairness [stattdessen] das Streben nach einer gerechteren sozialen Struktur zum Inhalt haben, möglicherweise auch durch die Entwicklung und den Einsatz von KI-Systemen, wo dies angemessen ist. Indem sie KI-Systeme in bestehende soziale Strukturen

---

[9] Literaturverweise im Text wurden zwecks besserer Lesbarkeit ohne Auslassungszeichen entfernt.

[10] Mit „KI-Bias" sind systematisch verzerrte Ergebnisse einer KI gemeint.

einordnet, ermöglicht diese Perspektive der strukturellen Ungerechtigkeit eine nuanciertere Analyse der Faktoren, die zu KI-Bias beitragen, und zeigt mögliche Orientierungen zur Verfolgung des Ziels „KI-Fairness“ auf. [...] [W]ir argumentieren, dass nicht nur Software-Ingenieur:innen, sondern eine breitere Personengruppe für das Streben nach KI-Fairness verantwortlich ist und sich an gemeinschaftlichem Handeln[11] zur Gestaltung der sozialen Struktur beteiligen sollte.

Im Folgenden stellen wir Empfehlungen und Beispiele dafür vor, wie unterschiedliche moralische Akteur:innen dabei helfen können, das Ziel „KI-Fairness“ zu erreichen.

Phasen	**Empfohlene Maßnahmen**
Phase 1. Problemdefinition	• Relevante soziale Kontexte und bestehende soziale Ungleichheiten sind zu untersuchen mit dem Ziel, mögliche Risiken und Nutzen der Entwicklung eines KI-Modells* zu identifizieren. • Mitgliedern diverser Personengruppen (vor allem marginalisierter Gruppen) sind bei der Entscheidung mit einzubeziehen, welche konkreten Probleme mit dem KI-Modell angegangen werden sollen. • Die Gesamtverteilung von vorhandenen Ressourcen ist einzuschätzen, um die Vergrößerung von Verteilungslücken zwischen Personengruppen zu vermeiden.
Phase 2. Datenaufbereitung	• Bei der Auswahl der [für die Entwicklung der KI benötigten] Datensets* ist sicherzustellen, dass die in ihnen enthaltenen Datenverteilungen repräsentativ sind, ohne Bias gegen marginalisierte Gruppen zu enthalten. • Der ausgewählte Referenzstandard[12] ist zu beurteilen, um die Reproduktion von bestehenden Ungleichheiten zu vermeiden.
Phase 3. Entwicklung und Validierung des KI-Modells	• Bei der Entscheidung, welche Informationen zum Training* des Algorithmus verwendet werden und welche Maßstäbe zur Validierung[13] des Algorithmus herangezogen werden sollen, sind die mit der Entscheidung verbundenen sozialen Faktoren kritisch zu analysieren. • Verschiedene Interessengruppen sind in den Prozess mit einzubeziehen, der akzeptable Maßnahmen zur Bewertung der Leistung des KI-Modells festlegt.
Phase 4. Verwendung und Überwachung des KI-Modells	• Vor der Verwendung des KI-Modells [in einem neuen Kontext] ist die Übereinstimmung zwischen der ursprünglich intendierten Nutzung des Algorithmus und der Nutzung in der Bevölkerungsgruppe, für die der Algorithmus nun verwendet werden soll, zu untersuchen. • Nach der Verwendung des KI-Modells sind die tatsächlichen Auswirkungen der Verwendung ständig neu zu prüfen. Entsprechende Anpassungen sind vorzunehmen.

Quelle: Lin, Ting-An/Chen, Po-Hsuan Cameron (2022): „Artificial Intelligence in a Structurally Unjust Society”. In: *Feminist Philosophical Quarterly* 8:3-4, Article 3, hier: 2f. und 19f., übersetzt von Larissa Bolte.

[11] Gemeint ist zielgerichtetes, kollektiv organisiertes Handeln auf ein Gemeingut hin (*collective action*).
[12] Vergleichswert, der ermöglicht, die Repräsentativität der Daten festzustellen, d.h. Schätzung, wie die Daten verteilt sein sollten, um die Realität gut abzubilden
[13] Überprüfung der Güte des fertig trainierten Algorithmus, d.h. wie gut der Algorithmus funktioniert