Unterrichtsbaustein ‚Diskriminierung – KI als Ursache oder Lösung?'

Erläuterungen zum Baustein

KI-Systeme kommen immer häufiger und in diversen sozialen Kontexten bei der Entscheidungsfindung zum Einsatz. Hiermit ist die Hoffnung verbunden, dass KI durch Automatisierung zu schnelleren Entscheidungsprozessen beiträgt und zugleich durch Rückbezug auf eine große Datengrundlage die Genauigkeit und Angemessenheit der Entscheidungen verbessert. Allerdings ist gut belegt, dass KI-Systeme oftmals bestehende gesellschaftliche Diskriminierungsformen reproduzieren sowie aufgrund intransparenter Datenverarbeitungsprozesse neue Formen der Diskriminierung befördern können.

Dieser Baustein thematisiert sowohl die Chancen der Nutzung von KI zur Diskriminierungsbekämpfung als auch die mit der maschinell unterstützten Entscheidungsfindung verbundenen Diskriminierungsrisiken. Die Schüler:innen sollen mit der gesamtgesellschaftlichen Relevanz KI-unterstützter Entscheidungen vertraut gemacht werden sowie den Zusammenhang zwischen algorithmischer Datenverarbeitung und Diskriminierung erarbeiten und diskutieren. Besonderes Augenmerk legt der Baustein hierbei auf das Spannungsfeld zwischen der Idee der vermeintlichen Neutralität oder Objektivität von Technik und der gesellschaftlichen Einbettung von (neuen) Technologien, die etwa mit partikularen Interessen, einer Privilegierung einzelner Gruppen oder mit Stereotypisierungen verbunden sein kann.

Der Baustein gehört zum Thema „KI-Ethik", findet allerdings auch Anknüpfungspunkte an die Bausteine zum Thema „Verantwortung für strukturelle Ungerechtigkeit". Er lässt sich einzeln unterrichten, mit weiteren Bausteinen der genannten Themen kombinieren oder anderweitig in Unterrichtsreihen einbetten, beispielsweise zu Fragen der Technikethik, zu Fragen der Gerechtigkeit oder zum Thema Diskriminierung. Er richtet sich primär an Lerngruppen der gymnasialen Oberstufe.

M1 konstruiert eine lebensweltliche Problemstellung, mit deren Hilfe Präkonzepte, Einstellungen und Intuitionen der Schüler:innen zur Frage nach den (anti-)diskriminierenden Potenzialen eines KI-Einsatzes bei der Entscheidung über Menschen abgerufen werden. Die Auswahl von Mitarbeiter:innen in großen Unternehmen ist einer der alltäglichen Kernfälle bewusster oder unbewusster Diskriminierung. Gleichzeitig greift hier der Einsatz von KI schon heute beträchtlich in das Leben der Menschen und die Gestaltung der Zukunft von Unternehmen ein. Besonders zur schnellen Bewältigung großer Datenmengen bietet sich die Verwendung von KI aus wirtschaftlichen Gründen an. Sie wird für Unternehmen daher bereits von zahlreichen kommerziellen Anbietern bereitgestellt.

In dem fiktiven Beispiel des Unternehmens *Drivesmarter* steht fest, dass Diskriminierung stattgefunden hat und dass dies in Zukunft vermieden werden soll. Die Intention der Schüler:innen in der Rolle der Entscheidenden, Diskriminierung zu vermeiden, ist in dieser Hinsicht gesetzt. Sie beruht in diesem Beispiel nicht auf moralischen Erwägungen, sondern auf dem Wunsch nach wirtschaftlichem Erfolg durch eine Bestenauswahl der zukünftigen Mitarbeiter:innen. Ausgespart wird hier die Frage nach der Zusammensetzung der aktuellen Belegschaft: Es besteht allerdings die Möglichkeit, dass Schüler:innen auf die Frage nach einem (allgemein leistungsförderlichen) harmonischen Miteinander in der

Belegschaft eingehen wollen – das dann etwa durch divers zusammengesetzte Arbeitsgruppen gestört sein soll oder – wie wissenschaftliche Studien eher belegen – gerade durch divers zusammengesetzte Arbeitsgruppen verbessert wird. Dies ist evtl. zu antizipieren. Der Fokus liegt aber auf der Frage, wie die Besten unter den Kandidat:innen für eine neue Stelle rein nach relevanten Leistungskriterien besser ausgewählt werden können: mit menschlicher Beurteilung oder mittels einer KI. Der Vorgang, der in der Fiktion neu gestaltet werden soll, bezieht sich auf eine Vorauswahl, um klarzustellen, dass es um größere Mengen an Daten geht, die schnell und vergleichsweise ‚oberflächlich' ohne ein zeitraubendes Interviewverfahren o.Ä. bewältigt werden müssen. Lebenslauf und Foto sollen verdeutlichen, dass Daten über Nationalität, Herkunft, Geschlecht, Aussehen, Hautfarbe etc. als mögliche Bezugspunkte für diskriminierende Entscheidungen im Auswahlprozess zur Verfügung stehen.

Aufgabe 1 erwartet die Formulierung der Entscheidung der Schüler:innen über den Einsatz von KI und eine intersubjektiv vertretbare Begründung dieser Entscheidung mit Bezug auf das gesetzte Ziel der diskriminierungsarmen Bestenauswahl. Die Fiktion des ‚Spickzettels' ermöglicht eine kurze Schreib- bzw. Einzelarbeitsphase, die von Vorträgen (im geschützten Raum kleiner Gruppenarbeit) gefolgt werden sollte (Aufgabe 2). Erwägungen, die bei dieser Entscheidung eine Rolle spielen können, sind nach Anlage des Falles z.B.:

- ‚Erfahrene' Mitarbeiter:innen sind unter Umständen eben gerade diejenigen, die schon in der Vergangenheit diskriminiert haben.
- Menschen sehen nicht alles.
- Menschen ermüden.
- Andererseits können in unterschiedlicher Hinsicht verschiedene Mitarbeiter:innen aufgrund ihres eigenen Hintergrunds gerade sensibel für Diskriminierung sein (daher sollte eine diverse Gruppe von Menschen die Vorauswahl treffen).
- Menschen sind empathiefähig.
- Menschen können im Gespräch/im Diskurs Einstellungen reflektieren und nach Lage eines Einzelfalles ändern.
- Eine KI fällt Entscheidungen rein nach objektiv festgelegten Kriterien. (Je nach Komplexität der KI ist hier allerdings ein Fragezeichen möglich!)
- Eine KI ermüdet nicht und greift daher auch nicht auf unreflektierte Stereotype oder Vorurteile zurück.
- Eine KI hat keine eigene Biographie, die das objektive Urteilsvermögen beeinträchtigen könnte.
- Eine KI versteckt eigenes Fehlverhalten nicht (ist aber zu einem gewissen Grad intransparent für die Nutzer:innen).
- Eine KI ist nicht durch Ereignisse im Privatleben beeinflusst.
- Eine KI kann allerdings nur auf Grundlage ihrer Trainingsdaten Entscheidungen treffen (und übernimmt damit unter Umständen indirekt das (Fehl-)Verhalten früherer Entscheider:innen).

Nach der Präsentation der verschiedenen Erwägungen und Positionen wird in der Gesamtgruppe eine gemeinsame Entscheidung aufgrund der Argumente gesucht. Es wird also im Diskurs ein Urteil über mögliche Problemlösungen gefällt. Die Aufgabenstellung ist daher uneingeschränkt offen für beide Möglichkeiten: KI-Einsatz oder menschliche Beurteilung. Die Sachlage liefert gute Gründe für beide Seiten.

In **M2** sollen die Schüler:innen mit bereits etablierten Formen der KI-unterstützten Entscheidungsfindung in gesamtgesellschaftlich relevanten Kontexten vertraut gemacht werden. Die Aufgaben 2 bis 4 dienen der selbständigen, arbeitsteiligen Erarbeitung der Sachgrundlage durch die Schüler:innen mittels Auszügen aus einer Stellungnahme des Deutschen Ethikrates zu Herausforderungen durch Künstliche Intelligenz von 2023. Zunächst soll Aufgabe 1 die Schüler:innen zum Einstieg anregen, mögliche ethische Schwierigkeiten zu antizipieren, die mit dem Einsatz von ADM-Systemen („Automated/Algorithmic Decision Making Systems“) verbunden sein können. Dabei können sie auf die bereits in M1 angestellten Überlegungen zu möglichen diskriminierungsrelevanten Effekten von KI zurückgreifen. Die Schüler:innen sollen anschließend im Gruppenpuzzle insbesondere herausarbeiten, zu welchen Zwecken ADM-Systeme zum Einsatz kommen können (nämlich zur Entscheidungsfindung bei Kindeswohlgefährdung, zur Ermittlung des Rückfallrisikos in der Bewährungshilfe und beim „Predictive Policing“). Vor dem Hintergrund des Erarbeiteten soll Aufgabe 5 die Diskussion zur Angemessenheit und Rechtfertigung von KI-unterstützter Entscheidungsfindung, die in M1 eröffnet wurde, vertiefen. Insbesondere werden zwei verschiedene Perspektiven angesprochen: Gesamtgesellschaftliches Interesse und Individualinteressen an automatisierter Entscheidungsfindung werden einander gegenübergestellt. Hier kann es zu Konflikten und kognitiven Dissonanzen kommen. So könnten etwa algorithmische Entscheidungen über das Schicksal potentieller Kindeswohlgefährder:innen und Straftäter:innen als vertretbar empfunden werden, während KI-gestützte Entscheidungen über das eigene Schicksal abgelehnt werden. Wenn Schüler:innen in ihrer Begründung der Ablehnung des Einsatzes von Algorithmen in Bezug auf ihr *eigenes* Leben auf die Fehleranfälligkeit oder das Diskriminierungspotential der Systeme verweisen, dann könnte dies in Spannung dazu stehen, dass sie den Einsatz der Systeme trotz dieser Defizite für legitim halten, wenn es um *andere* Menschen geht.

Im Zentrum von **M3** steht ein knapp zehnminütiges Interview mit Vincent Müller, Professor für Philosophie der Künstlichen Intelligenz. Das Interview dreht sich um die Frage, inwiefern KI eher eine Chance oder ein Problem im Hinblick auf die Vermeidung von Diskriminierung darstellt. Die Schüler:innen können das Interview anhören, das Transkript lesen oder beide Medien parallel nutzen. Während Aufgaben 1 bis 4 die Erschließung des Interviews unterstützen, fordert Aufgabe 5 vor diesem Hintergrund zur Diskussion der positiven Beurteilung Müllers von KI-gestützten Entscheidungen auf. Es empfiehlt sich, zunächst den ersten Teil des Interviews anzuhören oder zu lesen und dessen Verständnis zu sichern (bis min. 7:34 bzw. Auszug 1), bevor der zweite Teil erschlossen und die Ausführungen Müllers diskutiert werden (bis Ende bzw. Auszug 2).

Zu Aufgabe 1: Müller zufolge bedeutet Diskriminierung, dass man ein Urteil nach Kriterien fällt, die irrelevant sind für das, wonach man eigentlich sucht. Zu Aufgabe 2: Die erste genannte Weise, auf die Diskriminierung durch KI zustande kommen kann, ist die sogenannte historische Diskriminierung. Zu dieser kann es kommen, wenn man eine KI

auf Daten trainiert, die bereits eine Diskriminierung enthalten, da es wahrscheinlich ist, dass die KI diese lernt. Die zweite Weise der Diskriminierung kommt zustande, wenn die KI selbst diese hervorbringt und so Fehler produziert, die auch in Tests nicht erkannt werden. Zu Aufgabe 3: Eine *Stärke* von KI ist Müller zufolge, dass sie Diskriminierung aufdecken kann, die uns bislang nicht aufgefallen ist. Dies könne sie aufgrund der großen von ihr analysierten Datenmengen teils besser als Menschen. Die KI könne z.B. aufdecken, dass ein von uns als irrelevant beurteilter Faktor unsere Entscheidung beeinflusst. Eine *Schwäche* von KI besteht Müller zufolge darin, dass sie auf Basis unserer Erfahrungen und Anforderungen programmiert wird, die vorhandene Diskriminierung widerspiegeln. Zu Aufgabe 4: KI hat laut Müller grundsätzlich das Potential, unsere Entscheidungen rationaler zu machen, indem sie durch die Analyse großer Datenmengen aufdeckt, welche Kriterien wir tatsächlich bei Entscheidungen anwenden. Man kann z.B. systematisch einen Datensatz variieren und so mithilfe der KI herausfinden, dass ein Parameter bei einer Entscheidung eine Rolle spielt, der keine Rolle spielen sollte (z.B. das Alter).

In **M4** wird mit Hilfe philosophischer Textauszüge die Frage vertieft, wie es zu diskriminierenden Effekten bei der Entwicklung und durch den Einsatz von KI und anderer Technologien kommen kann. In Aufgabe 1 werden die Schüler:innen aufgefordert, selbst (noch einmal) mögliche Erklärungen dafür zu nennen, warum der Einsatz von KI zu Diskriminierung führen kann. Dadurch können in den bisherigen Materialien erarbeitete Grundlagen wiederholt und zusammengeführt werden. Erschließungsaufgaben zu den Textauszügen des Philosophen Jens Kipper und der Computerwissenschaftlerin Timnit Gebru unterstützen die Schüler:innen dabei, Mechanismen von KI-Systemen genauer zu erfassen, die diskriminierende Wirkungen entfalten können (Aufgaben 2 und 3). Teils wurden diese Mechanismen und Wirkungen, etwa im Kontext der KI-unterstützten Auswahl von Bewerber:innen, bereits zuvor im Baustein thematisiert. An dieser Stelle findet jedoch eine tiefergehende Betrachtung und kritische Einordnung statt, bei der insbesondere Gebru auch weitere technikphilosophische und gesellschaftskritische Bezüge herstellt. Dabei fällt ihre Einschätzung deutlich negativer aus als die in M3 von Müller formulierte Position. Kipper erläutert kleinschrittig drei Probleme bzw. Verzerrungen, die beim Einsatz von KI auftreten und mögliche diskriminierende Effekte erklären können: (a) die Tatsache, dass (bisher existierende) KIs lediglich Korrelationen, jedoch nicht Kausalbeziehungen aufdecken können, (b) die negativen Effekte eines ungeeigneten Erfolgskriteriums, mit der eine KI trainiert wurde und (c) das Training einer KI mit Daten aus einem diskriminierenden System.

Im Textauszug von Gebru legt diese den Fokus auf Diskriminierung, zu der es generell bei der Entwicklung von Technik – also keineswegs erst im Falle von KI – kommen kann. Sie kritisiert die vermeintliche Neutralität und Objektivität der Entwicklung von Technologien als Mythos und verweist auf Machtgefüge und Marginalisierungen, die mitbestimmen, welche Technologien wie entwickelt werden und wem diese nutzen. Auf das Verständnis von Gebrus Thesen zur Diskriminierung durch Technik(-entwicklung) zielt Aufgabe 3, in der Schüler:innen das Gelesene anhand von Beispielen erläutern sollen. Aufgabe 4 schließlich lenkt die Aufmerksamkeit auf die Wortwahl Gebrus („Technologie als Waffe“), in der ihre scharfe Kritik an diskriminierender Technikentwicklung und -nutzung deutlich wird. Indem die Schüler:innen die Wortwahl und die damit verbundene

Wertung beurteilen sollen, sind sie aufgefordert, mit einer reflektierenden, distanzierten Haltung auf die Ausführungen Gebrus und das von ihr diskutierte Phänomen zu schauen und sich selbst dazu zu positionieren.

M5 eröffnet zum Abschluss den Blick auf gesellschaftliche Handlungsoptionen und Lösungsansätze zur KI-Diskriminierungsproblematik. Das Material bietet Gelegenheit zur Reflexion des Lernfortschritts durch Rückbezug auf das in M1 bearbeitete Szenario.

Aufgabe 1 regt Überlegungen zur ethischen (fairen) Entwicklung von KI-Systemen an. Der zu bearbeitende Text von Ting-An Lin und Po-Hsuan Cameron Chen legt hierbei nahe, sowohl auf bestimmte technische Merkmale zu achten (etwa bestimmte Eigenschaften des Trainingsdatensets) als auch den Kontext der KI-Entwicklung in Betracht zu ziehen und zu diversifizieren. Die Schüler:innen sollen beachten, dass es verschiedene Phasen der KI-Entwicklung gibt, die über die technische Entwicklung hinausreichen. Zu diesen nicht-technischen Phasen gehören die Phase der Problemdefinition, in der erarbeitet wird, welches Problem oder welche Aufgabe durch das zu entwickelnde KI-System bewältigt werden soll (Phase 1); die Phase der Datenaufbereitung, welche sicherstellen soll, dass die zu verwendenden Datensätze möglichst repräsentativ sind (Phase 2); sowie die Phase der Verwendung und Überwachung des KI-Modells nach der technischen Entwicklung, also eine Phase der beständigen Qualitätskontrolle (Phase 4). In all diesen Phasen sowie in der Entwicklungs- und Validierungsphase selbst (Phase 3) sollen verschiedene soziale Gruppen miteinbezogen werden, um Diskrimierungspotenziale frühzeitig aufzudecken und einer einseitig durch die Perspektive technischer Fachkräfte bestimmten KI-Entwicklung entgegenzuwirken.

Aufgabe 2 weist über den Entwicklungskontext hinaus. Hier sollen die Schüler:innen zu einer eigenen Einschätzung darüber kommen, ob KI überhaupt ein geeignetes Mittel in Entscheidungsfindungsprozessen sein kann. Diese Diskussion ist ergebnisoffen. Relevante Gesichtspunkte können u.a. sein: die Objektivität von Technik und Standardisierung; die Zuverlässigkeit algorithmischer Entscheidungen, auch im Vergleich zur Zuverlässigkeit menschlicher Entscheidungen; das „Black-Box"-Problem der KI; die Bedeutsamkeit der Entscheidung für das Leben der Betroffenen; der Stellenwert von Rechtfertigung in sozialen Interaktionen; die notwendig beschränkte Perspektive menschlicher Akteur:innen. Eventuell kann die Diskussion durch die Ergebnisse aus M1 strukturiert werden.