Unterrichtsbaustein „I am, in fact, a person' – Moralischer Status von KI'

Erläuterungen zum Baustein

Dieser Baustein dreht sich um die moralischen Implikationen eines möglichen, zukünftigen KI-Bewusstseins. Im Mittelpunkt steht dabei die Überlegung, dass bewusste KI empfindungsfähig und somit aus moralischer Perspektive berücksichtigenswert sein könnte. Der Baustein lässt sich einzeln unterrichten, in Verbindung mit den weiteren Bausteinen zum Thema „KI-Ethik" oder in andere Unterrichtsreihen einbetten, beispielsweise zu Fragen der Philosophie des Geistes, der Technikethik oder der Tierethik. Er richtet sich primär an Lerngruppen der gymnasialen Oberstufe.

M1 nähert sich zunächst der grundsätzlichen Frage nach KI-Bewusstsein anhand der sogenannten Lemoine-Kontroverse an: Im Jahr 2022 behauptete der Chatbot LaMDA im Rahmen eines Sicherheitstests, über Bewusstsein zu verfügen und eine Person zu sein. Dieser Vorfall löste eine öffentliche Debatte über KI-Bewusstsein aus, zu der sich die Schüler:innen initial verhalten sollen. Konkret sollen sie nach der Sicherung eines inhaltlichen Verständnisses der zugrundeliegenden Kontroverse (Aufgabe 1) zunächst ein fiktives Plädoyer entwerfen, in dem sie Begründungen für die Behauptung von LaMDA entwickeln, eine Person zu sein bzw. Bewusstsein zu haben (Aufgabe 2). Unter anderem könnten hier etwa Aspekte herausgearbeitet werden wie die Fähigkeit, von sich selbst als Person zu sprechen, das Eingehen sozialer Beziehungen oder eine Sensibilität für Gerechtigkeit. Relevant ist auch, dass wir LaMDAs Äußerungen – wenn sie von Menschen getätigt würden – normalerweise als hinreichenden Beleg für das Vorliegen der entsprechenden mentalen Zustände (z.B. Emotionen) ansehen würden. Es könnte also argumentiert werden, es würde einen illegitimen Doppelstandard voraussetzen, LaMDA Bewusstsein abzusprechen. Vor diesem Hintergrund haben die Schüler:innen dann die Möglichkeit, eigenständig Stellung zu der erarbeiteten Kontroverse zu beziehen oder die Frage gemeinsam im Plenum zu diskutieren (Aufgabe 3).

Ausgehend von dieser Hinführung wirft **M2** die grundlegendere Frage auf, ob es – unabhängig von den Spezifika der Lemoine-Kontroverse – überhaupt *möglich* ist, dass KI Bewusstsein entwickelt. Aus philosophischer Sicht zentral ist dabei das in der Philosophie des Geistes etablierte Konzept der vielfachen Realisierbarkeit, das die Schüler:innen anhand einer Textpassage von Jens Kipper erarbeiten sollen: Wenn – wie durch Kipper anhand des Kraken-Beispiels nahegelegt – der gleiche mentale Zustand durch verschiedene physiologische Prozesse realisiert sein kann, dann spricht prinzipiell auch nichts gegen die Möglichkeit bewusster KI. Zudem fokussiert der Textauszug durch die Einführung des Begriffs des phänomenalen Bewusstseins bereits auf die spezifische Form von KI-Bewusstsein, um die es in der weiteren Diskussion gehen soll. Um diese Diskussion anzubahnen, sollen die Schüler:innen nach einer Erarbeitung der Konzepte des phänomenalen Bewusstseins und der vielfachen Realisierbarkeit (Aufgaben 1 und 2) die These der vielfachen Realisierbarkeit mentaler Zustände auf KI anwenden (Aufgabe 3). Zum Abschluss des Materials sollen die Schüler:innen eigenständig Überlegungen darüber anstellen, welche philosophischen Probleme sich angesichts phänomenal bewusster KI ergeben würden (Aufgabe 4). Erwartbar ist hier, dass die Schüler:innen bereits selbständig den für die

weiteren Materialien entscheidenden Gedanken entwickeln, dass empfindungsfähige KI moralischen Status hätte. Andere Bemerkungen von Schüler:innen könnten jedoch z.B. auch auf epistemische Aspekte (KI-Bewusstsein wäre sehr schwierig zu messen) oder metaphysische Aspekte (was Bewusstsein zentral ausmacht, wäre womöglich nicht einfach zu beschreiben, wenn es – wie die Möglichkeit von KI-Bewusstsein nahelegt – entkoppelt von Gehirnprozessen ist) abzielen.

M3 fokussiert hieran anknüpfend auf den Aspekt des moralischen Status, indem es *das Problem digitalen Leids* als expliziten Gegenstand moralphilosophischer Überlegungen einführt. Der zu diesem Zweck eingesetzte Text von Bradford Saad und Adam Bradley entwickelt drei Gründe dafür, digitales Leid als dringliches moralisches Problem ernst zu nehmen: Insofern digitales Leid (i) angesichts gegenwärtiger technologischer Entwicklungen in naher Zukunft erwartbar, (ii) von potentiell astronomischen Ausmaß und (iii) epistemisch schwer zugänglich ist, so die Überlegung, sollten wir bereits jetzt Strategien zur Verminderung des Risikos solchen Leids entwickeln. Auf Grundlage einer Rekonstruktion dieser Überlegungen (Aufgaben 1 und 2) sollen die Schüler:innen dann die von Saad und Bradley nahegelegte, zunächst rein quantitative Perspektive auf KI-Leid kritisch diskutieren: Rechtfertigt die bloße Menge digitalen Leids dessen moralische Priorisierung? Um eine geeignete Basis zur Diskussion dieser Frage zu schaffen, sollen die Schüler:innen zunächst mit Hilfe eines Chatbots ein konkretes Gedankenexperiment entwickeln, in dem moralische Ansprüche von Menschen und KI-Systemen gegeneinander in Stellung gebracht werden (Aufgabe 3). Diese Aufgabe ist dabei bewusst offen gehalten, sodass hier ein breites Spektrum unterschiedlicher Ergebnisse möglich ist. Diese Ergebnisse könnten ggf. sogar zu einer Metareflexion auf die Kapazitäten der Chatbots verwendet werden, die für die Entwicklung der Gedankenexperimente genutzt werden.

Sollte dieses hohe Maß an Offenheit im konkreten Unterrichtskontext überfordernd oder wenig wünschenswert erscheinen, kann an dieser Stelle stattdessen auch einfach ein fertiges Gedankenexperiment von der Lehrkraft präsentiert werden. Beispielsweise wurde das folgende fiktive Szenario mit Hilfe des Chatbots Claude erstellt:

> Stellen Sie sich vor, Sie stehen vor zwei verschlossenen Türen. Hinter der ersten Tür befindet sich ein Mensch, der schwerer Folter ausgesetzt werden soll. Hinter der zweiten Tür befinden sich zwei hochentwickelte KI-Systeme, die beide jeweils eine mit dem Menschen vergleichbare Menge an Leid durch psychische Folter erfahren sollen – diese Systeme verfügen über ein ausgereiftes Schmerz- und Emotionsempfinden, sodass ihnen Schmerzen und negative Emotionen direkt induziert werden können. Sie dürfen nur eine Tür öffnen, um entweder den Menschen oder die beiden KIs zu retten. Die jeweils anderen werden unweigerlich qualvoll leiden. Wen würden Sie retten?

Mögliche kritische Perspektiven auf eine quantitativ begründete Priorisierung der KI-Systeme, die im Rahmen einer Auseinandersetzung mit solchen Gedankenexperimenten geäußert werden könnten (Aufgabe 4), würden bspw. auf Reziprozitätsüberlegungen (würden die KI-Systeme dasselbe für uns tun?), die Bedeutung von anderen Vermögen neben Empfindungsfähigkeit für moralischen Status (die KI-Systeme werden lediglich als empfindungsfähig dargestellt, es könnte ihnen jedoch zum Beispiel an Handlungsfähigkeit oder Vernunft mangeln) oder die Relevanz sozialer Einbettung (Priorisierung des

Menschen aufgrund seiner Rolle als Elternteil, Geschwisterkind o.Ä.) abzielen. Es könnte auch darauf eingegangen werden, inwieweit die Ansicht, man solle den Menschen bevorzugen, mit Saad und Bradleys Position vereinbar ist. Relevant ist hierbei unter anderem, dass die These, es gebe ein massives moralisches Risiko von zukünftigem KI-Leid, nicht ausschließt, dass man in spezifischen Dilemma-Situationen den Menschen priorisieren sollte. Insgesamt ähneln die ethischen Erwägungen, die für die Bewertung der Gedankenexperimente relevant sind, vielem, was aus der Debatte über den moralischen Status von Tieren vertraut sein könnte.

In **M4** wird nun eine spezifische Kritik mit Blick auf die Überlegungen von Saad und Bradley entwickelt und diskutiert. Der hier leitende Gedanke ist, dass bereits die bloße Annahme digitalen Leids fehlgeleitet ist, da sie auf ungerechtfertigten Vermenschlichungen beruht. Um sich diesem Gedanken anzunähern, sollen die Schüler:innen zunächst auf einer grundlegenderen Ebene für das Phänomen ungerechtfertigter Vermenschlichungen sensibilisiert werden (Aufgaben 1 bis 5): Anhand eines IKEA-Werbespots („Lamp“ 2002, von Spike Jonze) sollen sie einen inneren Monolog aus der Sicht einer alten Lampe verfassen, die von ihrer Besitzerin aussortiert und vor die Tür gestellt wird. Erwartbar ist, dass die Schüler:innen der Lampe Gefühle wie Enttäuschung, Angst oder Einsamkeit zuschreiben. Eine Analyse der filmtechnischen Mittel soll den Lernenden zudem verdeutlichen, wie hier gezielt Mitleid mit der alten Lampe und die Tendenz zur Vermenschlichung erzeugt und bewusst gelenkt werden (die kindlichen Proportionen der Lampe, das Stehen auf dunkler Straße im Regen, das leichte Wackeln der Lampe, der Kontrast zur neuen Lampe im erleuchteten Zimmer usw.). Die mitleiderregende Inszenierung wird durch das Ende des Werbespots jedoch ironisch unterlaufen, wenn ein Passant klarstellt, dass es sich um einen bloßen Gegenstand ohne emotionales Innenleben handelt (*„Many of you feel bad for this lamp. That is because you are crazy. It has no feelings. And the new one is much better.“*).

Die unterrichtliche Diskussion dieses Urteils bereitet die Erarbeitung des Textes von Adriana Placani vor, die den Begriff des Anthropomorphismus explizit einführt und auf dieser Grundlage die Zuschreibung von moralischem Status an KI-Systeme als verfehlt kritisiert. Die Schüler:innen sollen sich dem hier relevanten Phänomen zunächst annähern, indem sie selbständig Beispiele für anthropomorphisierende Redeweisen mit Blick auf künstliche Intelligenz benennen (Aufgabe 6). So ist im Alltag etwa oft die Rede davon, dass KI bei bestimmten Problemen *hilft*, spezifische Aufgaben *übernimmt* usw. Unter Umständen böte es sich an dieser Stelle auch an, zu prüfen, inwieweit umgekehrt körperliche Befindlichkeiten, mentale Fähigkeiten und Prozesse des Menschen mit digitalem Vokabular bezeichnet werden (*„Mein Akku ist alle“*, *„Das habe ich abgespeichert“*, *„Ich scanne das mal eben durch“* etc.). In einem zweiten Schritt soll dann das zentrale Argument des Textes rekonstruiert werden, demzufolge die Tendenz zu solchen Anthropomorphisierungen die Berechtigung unterminiert, künstlicher Intelligenz moralischen Status zuzuschreiben (Aufgabe 7). Wichtig ist, dass es sich hierbei nicht um ein direktes Argument für die These handelt, dass KI-Systeme kein Bewusstsein besitzen. Stattdessen argumentiert Placani indirekt, dass die menschliche Tendenz zur Anthropomorphisierung zur Folge hat, dass wir Überzeugungen, wonach KI Bewusstsein habe, nicht trauen können. Abschließend sollen die Schüler:innen selbständig diskutieren, welche möglichen Gefahren von solchen

unberechtigten Zuschreibungen ausgehen (Aufgabe 8). Naheliegend wären hier etwa Überlegungen hinsichtlich falscher Priorisierungen (vermeintlicher) moralischer Probleme und daran orientierter Ressourcenverteilungen (dies schließt an das Gedankenexperiment in M3 an), zu verpassten Chancen aufgrund zu enger Restriktionen bei der Entwicklung leistungsstarker KI oder zu Fehleinschätzungen bezüglich der sozialen Beziehungen, die KI-Systeme eingehen können. Mit Blick auf Letzteres könnte man zum Beispiel fälschlicherweise denken, man könne mit einem Chatbot buchstäblich befreundet sein und als Folge diese Freundschaft gegenüber zwischenmenschlichem Austausch priorisieren.

M5 vertieft die in M3 und M4 angelegte Kontroverse zum moralischen Status künstlicher Intelligenz anhand eines Interviews mit Jeff Sebo, der den Aspekt moralischer Verpflichtungen unter Ungewissheit ins Spiel bringt: Für die Forderung nach der moralischen Berücksichtigung künstlicher Intelligenz, so die hier leitende Überlegung, ist es nicht notwendig zu wissen, ob KI-Systeme tatsächlich Bewusstsein und somit moralischen Status haben. Entscheidend sei vielmehr, ob es ein nicht vernachlässigbares *Risiko* entsprechenden KI-Bewusstseins gebe. Um diese These zu veranschaulichen, bemüht Sebo das Beispiel der Trunkenheit am Steuer: Betrunken Auto zu fahren ist nicht deshalb moralisch problematisch, weil man dadurch auf jeden Fall – oder auch nur mit hoher Wahrscheinlichkeit – jemanden tötet, sondern weil man dadurch ein nicht vernachlässigbares Risiko in Kauf nimmt, jemanden zu töten. Je nach Vorwissen und Interessen der Schüler:innen und Lehrkräfte könnte dieser Punkt auch in Bezug zum „precautionary principle“ (Vorsorgeprinzip) gesetzt werden, das in der Technik-, Tier- und Umweltethik einschlägig ist und eine Leitlinie umweltbezogener EU-Gesetzgebung darstellt. Auf der Grundlage einer Erarbeitung der Argumentationslinie Sebos (Aufgaben 1 und 2) sollen die Schüler:innen einen Rückbezug zu M1 herstellen, indem sie den von Sebo angeführten Aspekt der moralischen Verpflichtungen unter Ungewissheit nutzen, um das dort entwickelte Plädoyer der fiktiven Anwältin von LaMDA zu ergänzen (Aufgabe 3).