Unterrichtsbaustein ‚Superintelligenz – Ende oder Rettung der Menschheit'?

Erläuterungen zum Baustein

Die übergeordnete Problemstellung dieses Bausteins lautet: Ist künstliche *Super*intelligenz die mit ihr verbundenen Risiken wert? Den Rahmen für die Bearbeitung der Problemstellung bildet das fiktionale Szenario eines Forschungsteams, das kurz vor der Freigabe eines neu entwickelten superintelligenten Systems namens SOPHOS steht. Dieses Szenario sollen die Lernenden schrittweise beurteilen.

Der Baustein gehört zum Thema „KI-Ethik". Er lässt sich einzeln unterrichten, mit den weiteren Bausteinen zu diesem Thema kombinieren oder anderweitig in Unterrichtsreihen einbetten, beispielsweise zu Fragen der Technikethik oder der Anthropologie. Er richtet sich primär an Lerngruppen der gymnasialen Oberstufe.

M1 dient der Einführung in die Problematik, zwischen Chancen und Risiken superintelligenter Systeme abwägen zu müssen, sowie der Klärung der hiermit verbundenen Grundlagen und Begriffe. Die Problemstellung wird durch ein fiktionales, zugespitztes Szenario eröffnet, in dem die Menschheit zwischen den beiden Optionen entscheiden soll, eine künstliche Superintelligenz freizulassen oder zurückzuhalten. Dies kann im Plenum mündlich besprochen werden (Aufgaben 1 und 2). Es können verschiedenste positive und negative Beispiele oder Szenarien angedacht werden, welche durch die Freisetzung einer künstlichen Superintelligenz bedingt werden könnten: Eine solche könnte etwa einen Beitrag zur Bekämpfung der Klimakrise leisten, aber auch zu einer fehlenden Kontrolle bestimmter Waffensysteme durch den Menschen führen.

Anschließend erfolgt die textbasierte Erarbeitung des Begriffs der Superintelligenz (in Abgrenzung zu „allgemeiner KI" – siehe jeweils das Glossar) sowie von Argumenten dafür, dass die Existenz einer solchen tatsächlich möglich ist (Aufgaben 3 und 4). Hierzu gehört der Gedanke, dass künstliche Intelligenz kein Ergebnis evolutionärer Prozesse ist, die zu natürlichen Beschränkungen der Leistungsfähigkeit etwa des menschlichen Gehirns führen. Für KI müssen solche Beschränkungen nicht gelten. Zudem zeigt das Beispiel der Strategiespiele Schach und Go, dass selbst Expert:innen in bestimmten Gebieten verschwindend geringe Fähigkeiten besitzen im Vergleich zu dem, was KI zu leisten vermag. Zuletzt ist hier die Erkenntnis relevant, dass menschliche Kognition im Vergleich zur Arbeitsleistung von Computern schlichtweg sehr langsam ist. Die Aufgabe, sich eine KI vorzustellen und auszumalen, die eine Million Mal schneller denkt als man selbst, soll entsprechend dazu beitragen, zu verdeutlichen, wie weit Superintelligenz menschlicher Intelligenz überlegen sein könnte (Aufgabe 5a). Hier könnten etwa Science-Fiction-Filme wie *Her* (Spike Jonze 2013) zur weiteren Veranschaulichung herangezogen werden. Im Sinne einer Zwischenbilanz werden die gesammelten Erkenntnisse erneut auf die anfängliche Entscheidungsfrage angewandt (Aufgabe 5b).

Optional kann nun mit dem sogenannten Gorilla-Problem die Erkenntnis vertieft werden, was für die Menschheit auf dem Spiel steht. Durch die Erschaffung einer Superintelligenz würden wir zu dieser in ein Verhältnis gesetzt, das demjenigen von Gorillas zu Menschen entspricht: Gorillas sind nicht nur weniger intelligent, sondern ihre Existenz ist dem Willen der Menschheit vollkommen ausgeliefert. Diese Intelligenzunterschiede lassen

sich zu Veranschaulichungszwecken einfach visualisieren, etwa in Gestalt einer Art von Intelligenz-Entwicklungslinie, bestehend aus Gorilla, Mensch und Superintelligenz. In diesem Zusammenhang kann zudem der Unterschied zwischen der Entstehung des Menschen und der Entwicklung von Superintelligenz reflektiert werden (Aufgabe 6). Insgesamt ließe sich mit dem Gorilla-Problem ebenfalls zur Behandlung der Herausforderung einer mangelnden Kontrollierbarkeit von Superintelligenz überleiten (Aufgabe 7), die mit dem anschließenden Unterrichtsmaterial weiterverfolgt wird.

In **M2** bearbeitet die Lerngruppe potenzielle Risiken, die mit der Aktivierung einer Superintelligenz einhergehen und die bis hin zur vollständigen Auslöschung der Menschheit reichen. Die intrinsischen Ziele der in den Texten imaginierten Superintelligenzen bestehen in der Beseitigung von Corona (Kipper) und im Kaffeeholen (Russell). Die Tötung von Menschen in Ersterem und die Sicherung des eigenen Fortbestehens in Letzterem sind instrumentelle Ziele, um das übergeordnete intrinsische Ziel zu erreichen. Die Unterscheidung zwischen intrinsischen und instrumentellen Zielen wird in M2 nur implizit thematisiert. Der Grundgedanke der Texte funktioniert auch, ohne die begriffliche Unterscheidung einzuführen. Bei leistungsstärkeren Lerngruppen kann dies jedoch fruchtbar sein (siehe das Glossar).

Indem die Schüler:innen in Aufgabe 1 die Zielformulierungen zur Corona-Beseitigung und mögliche (katastrophale) Folgen durchspielen, erfahren sie, wie schnell Fehler in der Ziel- und Wertsetzung von Superintelligenz zu Konsequenzen existentiellen Ausmaßes führen könnten. Denn zum einen können superintelligente Systeme nicht von Menschen kontrolliert werden, und zum anderen impliziert Intelligenz nicht, dass bestimmte, von Menschen geteilte Werte verfolgt werden. Bei leistungsstarken Lerngruppen kann im Corona-Szenario der Einwand der Forscherin („Warum sollte die KI nicht einfach alle mit Corona infizierten Menschen töten? Schließlich wäre dies ein sehr effektiver Weg, diese Krankheit auszurotten.“) zunächst weggelassen werden und Lerngruppen könnten selbst über den schlechtestmöglichen Ausgang der Zielsetzung nachdenken.

Bei den verbesserten Zielformulierungen, die Lernende entwickeln sollen, ist mit hoher Wahrscheinlichkeit davon auszugehen, dass keine erfolgreiche Formulierung herauskommt. Exemplarisch kann im Plenum durchgespielt werden, welche katastrophalen Folgen bei der Realisierung der Zielformulierungen durch die Superintelligenz entstehen könnten. Denkbar wäre etwa, dass Lernende Formulierungen wählen wie „Sorge dafür, dass Corona besiegt wird, ohne dass dabei Menschen sterben“. Bei der Formulierung könnte das superintelligente System alle Tiere töten, die das Virus möglicherweise weitergeben, oder alle Menschen zwangsisolieren. Es sollte ein Problembewusstsein dafür geschaffen werden, wie schnell eine Superintelligenz zur (existenziellen) Bedrohung werden könnte, ohne dass intrinsisch schlechte Absichten in Bezug auf Menschen zu haben.

Im nachfolgenden Text von Kipper wird das Szenario aufgegriffen. Es sollte deutlich werden, dass Menschen superintelligente Systeme nicht von der Verfolgung ihrer Ziele abhalten können, da die Systeme antizipieren werden, dass Menschen sie möglicherweise abschalten wollen und dass dies hinderlich für die Erreichung ihrer Ziele ist. Superintelligenzen könnten Menschen aus instrumentellen Gründen aus dem Weg räumen wollen, wenn der Nutzen ihrer Existenz nicht groß genug für die Zielerreichung ist.

In Aufgabe 2 sollte festgehalten werden, dass die Sinnhaftigkeit der Ziele der Superintelligenzen von Menschen sehr unterschiedlich eingeschätzt werden dürfte. In Bezug auf Aufgabe 3 sind vielfältige Szenarien denkbar, etwa eine simulierte Abschaltung des Systems bei gleichzeitigem Vorhandensein von versteckten Kopien, technische Schutzmaßnahmen gegen die Abschaltung, die das System schon weit im Voraus angelegt hat oder die Drohung, kritische Infrastruktur zu sabotieren. Es sollte herauskommen, dass eine Superintelligenz, die alle menschlichen Handlungen antizipieren kann, sich nicht von Menschen ausschalten lassen würde, sofern dies nicht explizit ihren eigenen Zielen entsprechen würde. Zur Illustration des Abschalt-Szenarios könnte die Filmszene aus Stanley Kubricks *2001: Odyssee im Weltraum* verwendet werden, die auf Youtube unter dem Stichwort „HAL 9000: I'm sorry Dave, I'm afraid I can't do that" oder auf Deutsch unter „Es tut mir Leid Dave, aber das kann ich nicht tun" zu finden ist.

Wie in **M3** thematisiert wird, scheint es eine naheliegende Option zu sein, der KI bestimmte Werte anzutrainieren, noch bevor sie sich zur Superintelligenz entwickelt. Aus dem Text „Dumme Werte: Das Problem der Wertharmonie" geht jedoch hervor, dass (Super-)Intelligenz prinzipiell mit allen möglichen Werten kombiniert werden kann (*Orthogonalitätsthese*). Das kann am Beispiel hochintelligenter Menschen illustriert werden, die nicht per se gesellschaftlich besonders geschätzte Werte teilen oder moralischer handeln als durchschnittlich intelligente Menschen. Für die Erreichung der meisten Ziele dürfte es für ein superintelligentes System nicht ausschlaggebend sein, ob eine Harmonie mit den von Menschen geschätzten Werten besteht. Nach der Texterschließung (Aufgaben 1 und 2) gehen die Lernenden diesen Überlegungen in Aufgabe 3 mit der Erstellung eines philosophischen Gutachtens zur Superintelligenz SOPHOS weiter nach. Es ist davon auszugehen, dass sie vor dem Hintergrund der bislang bearbeiteten Texte zur Einschätzung kommen, dass durch die Beliebigkeit der Werte von superintelligenten Systemen eine sehr große Gefahr für das Fortbestehen der Menschheit existiert und dass sie deshalb vor der Freigabe des Systems warnen. Es empfiehlt sich, die Gutachten digital anfertigen zu lassen, da im Verlauf des Bausteins weiter daran gearbeitet wird und die Möglichkeit einer Überarbeitung gegeben sein sollte.

M4 kann mithilfe eines fiktiven *Social Media*-Beitrags eingeleitet werden, in dem darauf hingewiesen wird, dass wir aus guten Gründen an immer intelligenterer KI arbeiten: Sie hat das Potenzial, unser Leben grundlegend einfacher, länger und glücklicher zu machen, globale Probleme zu lösen, und wir erhoffen uns sehr viel von ihr. Dieser Einstieg soll an die Lebenswelt und mögliche Intuitionen der Schüler:innen anknüpfen und ihnen zunächst ermöglichen, eigene Prognosen über die potenziellen Vorteile superintelligenter KI abzugeben, beispielsweise in Form einer Sammlung von Chancen. Dabei können eventuell bereits sehr gut informierte Schüler:innen ermutigt werden, ihr Vorwissen zu teilen. Zu diesem Zweck kann die erste Aufgabe auch erweitert werden, indem sie zum Beispiel eine Verbesserung oder Erweiterung des *Social Media*-Beitrags oder einen ganz eigenen Kommentar unter dem Beitrag verlangt, in dem die Schüler:innen den dort gemachten Punkt entweder ergänzen oder mit Einwänden konfrontieren. Die Lernenden erhalten hier die Möglichkeit, ihr philosophisches Gutachten zu überarbeiten oder zu ergänzen. Die dabei antizipierte pessimistische Einstellung basiert vermutlich nicht nur auf

den in M1 und M2 präsentierten Materialien, sondern auch auf popkulturellen und medialen Bildern und Narrativen von künstlicher Intelligenz.

Beim nachfolgenden Textauszug Russells handelt es sich um eine detaillierte Darstellung von zwei möglichen positiven Effekten einer autonomen, intelligenten KI. Zum einen kann schon eine KI mit menschlichem Intellekt potenziell verfügbare Ressourcen so effizient einsetzen, dass der Lebensstandard aller Menschen signifikant angehoben wird. Dies erklärt eindrücklich und nachvollziehbar, warum es rational sein könnte, KI immer intelligenter zu machen. Zum anderen skizziert Russell, wie eine intelligente KI den Menschen sogar autonomer machen oder Bildungsgerechtigkeit schaffen könnte, indem mit ihrer Hilfe jedem Menschen persönliche KI-Lehrkräfte, -Anwält:innen und -Berater:innen zur Verfügung gestellt werden könnten.

Da der Text in erster Linie mit detaillierten Beispielen arbeitet, soll die Erarbeitung von Aufgabe 2 anhand eines selbstgewählten Beispiels erfolgen. Dieser Zugriff lässt sich im Sinne der Binnendifferenzierung flexibel anspruchsvoller oder niedrigschwelliger gestalten, indem entweder das vorgegebene Zitat weggelassen oder ein zu erläuterndes Beispiel vorgegeben wird. In Aufgabe 3 wird ein Transfer angestoßen, bei dem die Schüler:innen über die Auswirkung von KI mit menschenähnlicher Intelligenz auf die Macht und Wirkung *super*intelligenter KI extrapolieren, wie groß die potenziellen Vorteile einer superintelligenten KI wären. Da sich diese Übertragungsleistung ggf. als schwierig herausstellen könnte, kann diese Aufgabe für jüngere oder weniger starke Lerngruppen weggelassen oder modifiziert werden. Eine mögliche Modifikation wäre etwa die Bereitstellung eines Fragenkatalogs oder einer Kriterienliste, der bzw. die die Schüler:innen dabei anleitet, die KI entlang einer Reihe verschiedener Dimensionen zu bewerten (etwa: Eine KI mit menschlicher Intelligenz wäre so überzeugend wie ein sehr guter menschlicher Anwalt. Wie überzeugend wäre eine superintelligente KI?). Nachdem die Lernenden in Aufgabe 4 zunächst auf einer persönlichen Ebene Stellung dazu beziehen, inwiefern sie eine durch KI strukturierte Gesellschaft und Ökonomie als wünschenswert einschätzen, sollen diese Einschätzungen in Aufgabe 5 verallgemeinert und in das philosophische Gutachten integriert werden.

M5 kann mit einer kurzen Zusammenführung der bisherigen Arbeitsergebnisse eingeleitet werden, bei der die Schüler:innen Probleme und Chancen superintelligenter KI gegenüberstellen und zu einer vorläufigen Einschätzung kommen (Aufgabe 1). Dabei rufen sie sich ins Gedächtnis, dass das Wertharmonie-Problem aus M2 als eine große Schwierigkeit bestehen bleibt, und reaktivieren ggf. erste eigene Lösungsansätze. Diese einleitende Zusammenführung kann durch die vorgeschlagene Diskussion in Kleingruppen erarbeitet werden, die durch Vereinfachungen oder weitere Vorgaben ergänzt werden können. So können die Schüler:innen verschiedene Rollen (etwa KI-Expert:innen, Ethiker:innen, Politiker:innen, …) zugewiesen bekommen oder der Arbeitsauftrag kann durch die Verschriftlichung der Ergebnisse ergänzt werden. Denkbar wären auch Visualisierungen zum Stand der Dinge, z. B. in Form einer Waage oder eines passenden Memes, das an den Einstieg in M1 anknüpft.

Der nachfolgende Text von Vincent Müller und Michael Cannon stellt einen metaethischen Lösungsansatz für das Wertharmonie-Problem vor. Der Lösungsansatz basiert auf der Idee, dass ein Wesen, das intelligenter als ein Mensch ist, erwartbarerweise

mindestens so vernünftig die eigenen Ziele reflektiert und auswählt, wie ein Mensch es tut. Daraus ergibt sich die Frage, warum die Werte einer Superintelligenz überhaupt mit unseren in Harmonie gebracht werden müssten. Schließlich müsste eine Superintelligenz die Fähigkeit haben, sich selbst Ziele zu setzen. Es kann davon ausgegangen werden, dass ein superintelligentes System aus moralischer Erkenntnis heraus eben nicht die Menschheit auslöschen würde, um z.B. die Büroklammerproduktion zu maximieren. Als Hilfestellung zu Aufgabe 2 könnte etwa eine Schritt-für-Schritt Rekonstruktion des Textes angefertigt werden. Leistungsstärkere Lernende können vertiefend in einem Schaubild darstellen, wie genau sich Müller und Cannons Lösung auf das Wertharmonie-Problem bezieht.

Die kognitive Umwälzung der Texterarbeitung erfolgt im Rahmen der dritten Aufgabe durch einen Transfer auf das Beispiel der Corona-Beseitigung aus M2. Eine kompetente Anwendung des Lösungsansatzes von Müller und Canon kann beispielsweise vorsehen, dass die Superintelligenz mit der Fähigkeit ausgestattet wird, ihre eigenen Ziele zu reflektieren und ggf. zu ändern. Ergänzend können die Schüler:innen aufgefordert werden, ein neues Ziel zu formulieren, dass sich die Superintelligenz selbst setzt, ggf. mit Rückgriff auf bereits bekannte ethische Theorien (z. B. „Beseitige das Corona-Virus, ohne dass dadurch mehr Leid entsteht als verhindert würde."). Die vierte Aufgabe ermöglicht eine kritische Reflexion des Lösungsansatzes. Zur Unterstützung der fünften Aufgabe kann es sinnvoll sein, zentrale Ergebnisse der Diskussion schriftlich festhalten zu lassen (etwa als *Placemat*) und dann zu sammeln.

Den Abschluss des Bausteins stellt die Fertigstellung des philosophischen Gutachtens zur Freigabe von SOPHOS dar. Idealerweise sollen die Schüler:innen sowohl auf die Gefahren superintelligenter KI im Allgemeinen und auf das Wertharmonie-Problem und seine möglichen Lösungen im Besonderen eingehen, als auch auf die dagegen abzuwägenden Chancen, die solche Systeme bieten. Bei schwächeren Lerngruppen kann die Aufgabe stärker angeleitet werden (z.B. Verweis auf das Gorilla-Problem, unterstützende Visualisierung). Bedarf die Lerngruppe grundsätzlich größerer Unterstützung, bietet es sich auch an, für das philosophische Gutachten ein standardisiertes Dokument (z.B. ein Arbeitsblatt mit Überschriften und Formulierungshilfen) vorzubereiten und zu Beginn der Arbeit mit dem Baustein auszuteilen.

Im Anschluss an diesen Baustein ließe sich zur Vertiefung etwa die Position des Antinatalismus behandeln, die gegen das Fortbestehen der Menschheit gerichtet ist und somit ein Aussterben der Menschheit durch superintelligente KI womöglich gutheißen würde. Denkbar wäre auch eine Auseinandersetzung mit religionsphilosophischen Fragen, etwa bezüglich Parallelen zwischen dem Konzept der ‚Superintelligenz' und dem Gottesbegriff.