Unterrichtsbaustein ‚Superintelligenz – Ende oder Rettung der Menschheit?'

Materialien zum Baustein

- M1 SOPHOS, ein superintelligentes System
- M2 Superintelligente Systeme als eine (existentielle) Bedrohung für die Menschheit?
- M3 Superintelligenz mit superintelligenten Werten?
- M4 Superintelligente Systeme als einmalige Chance für die Menschheit?
- M5 Sollen wir die Risiken, die von KI ausgehen, in Kauf nehmen?

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

M1 SOPHOS, ein superintelligentes System

KI-Forscher:innen haben es geschafft: Sie haben ein superintelligentes* System namens SOPHOS entwickelt, das schon jetzt zwanzigmal so intelligent ist wie die klügsten Menschen. SOPHOS kann eigenständig revolutionäre wissenschaftliche Durchbrüche erzielen, ist allen Menschen in jeder intellektuellen Tätigkeit (Literatur, Mathematik, Philosophie, Programmieren usw.) meilenweit überlegen und verbessert sich selbstständig. Außerdem kann es mit Menschen kommunizieren. Aktuell ist SOPHOS noch nicht online. Das Team plant den Release für nächste Woche.

Aufgaben

1. Erörtert, inwiefern die Nervosität der Person im Meme angemessen ist. Berücksichtigt dabei besonders mögliche Konsequenzen einer Aktivierung von SOPHOS.
2. Erstellt ein Meinungsbild zu der Frage, ob SOPHOS aktiviert werden sollte und begründet eure Haltungen.

Quelle: Erstellt mit *Meme Generator*, URL: https://imgflip.com/memegenerator.

Die Relevanz der Frage, ob die Menschheit eine superintelligente KI aktivieren sollte, hängt unter anderem von der Beantwortung einer vorgelagerten Frage ab: Wie wahrscheinlich ist es überhaupt, dass die Entwicklung einer solchen KI in Zukunft möglich sein wird? Im folgenden Text nimmt der Philosoph Jens Kipper hierzu eine Einschätzung vor.

Jens Kipper (2020): Kann es eine superintelligente KI geben?

Eine viel diskutierte Frage ist die, ob auf die Entwicklung allgemeiner KI* schon bald die einer *Superintelligenz* folgt – eines Systems, dessen Intelligenz der von Menschen deutlich überlegen ist. Nun ist *allgemeine KI* als KI definiert, die der menschlichen Intelligenz in allen Bereichen mindestens ebenbürtig ist. Es ist wahrscheinlich, dass die erste allgemeine KI in etlichen Bereichen bereits deutlich übermenschliches Niveau hätte – es wäre ein seltsamer Zufall, wenn eine solche KI in allen Bereichen genau auf unserem Niveau wäre. Der Schritt von allgemeiner KI zu Superintelligenz ist also zumindest nicht sehr groß. [...]

Vieles spricht dafür, dass Superintelligenz zumindest möglich ist. Menschliche Intelligenz ist das Ergebnis evolutionärer Prozesse. Natürliche Auslese belohnt keineswegs nur größere Intelligenz, sondern auch etliche andere Faktoren, beispielsweise hinsichtlich der Masse und des Energieverbrauchs eines Organismus: Größere, schwerere Körper mit höherem Energieverbrauch sind schlicht schwerer zu versorgen. Im Vergleich mit anderen Lebewesen haben Menschen ein ungewöhnlich großes Gehirn mit einem ungewöhnlich hohen Energieverbrauch. Aber auch unser Gehirn ist weniger als 1,5 kg schwer und verbraucht nur etwa 20 Watt. Es liegt nahe, dass diese Beschränkungen zu Lasten der erreichten Intelligenz gehen. Eine KI hingegen unterliegt nicht den Gesetzen natürlicher Auslese. Sie kann daher sehr viel größer und schwerer sein und sehr viel mehr Energie verbrauchen als ein Mensch. Wir wissen zudem heute schon, dass in vielen intellektuellen Bereichen sehr viel Luft nach oben ist. Zwei gute Beispiele dafür sind Schach und Go.[1] Das folgende Zitat von Ke Jie, der Nummer 1 der Go-Weltrangliste, veranschaulicht, wie weit wir im Go vom Optimum entfernt sind: „Nachdem die Menschheit Jahrtausende darauf verwendet hat, unsere Taktiken zu verbessern, zeigen uns Computer, dass wir vollkommen falsch lagen. Ich würde so weit gehen, zu sagen, dass nicht ein einziger Mensch auch nur den Rand der Wahrheit im Go berührt hat.“ (Dou/Geng 2017[Übers. J. K.]) Es gibt wenig Grund zu glauben, dass es in anderen, noch komplexeren Bereichen anders aussieht. Und schließlich ist menschliche Kognition[2] nicht besonders schnell. [...] [Computer] übertragen [...] Signale mehr als eine Million Mal schneller als das menschliche Gehirn. Stellen Sie sich nun eine Person vor, die eine Million Mal schneller denkt als [S]ie. [...] Zumindest aus unserer Perspektive würden wir so eine Person sicher als Superintelligenz betrachten – und das, obwohl sie „nur“ schneller ist als wir und keine überlegenen Algorithmen verwendet.

Quelle: Kipper, Jens (2020): *Künstliche Intelligenz – Fluch oder Segen?* Stuttgart: Metzler, 61–63.

Aufgaben

3. Erläutere den Unterschied zwischen allgemeiner KI und Superintelligenz. Intelligenz kann hierbei „einfach als die Fähigkeit verstanden [werden], komplexe Aufgaben zu bewältigen“ (Kipper 2020, S. 67).
4. Kipper behauptet: „Superintelligenz ist sehr wahrscheinlich möglich.“ Rekonstruiere seine Begründung für diese These.

---

1 Bei Go handelt es sich um ein komplexes Strategiespiel.
2 Gemeint ist hier das menschliche Denkvermögen.

5. Stell dir eine künstliche Intelligenz vor, die „eine Million Mal schneller“ denkt als Menschen.
   a. Stelle dar, was diese KI in einer Minute, in einer Woche und in einem Jahr leisten könnte.
   b. Diskutiert auf der Basis eurer Ergebnisse erneut, ob die superintelligente KI SOPHOS aktiviert werden soll.

Mögliche Vertiefung

Wenn ein superintelligentes System aktiviert würde, wäre der Mensch nicht mehr das intelligenteste Wesen der Erde. Der Informatiker Stuart Russell verdeutlicht im folgenden Text, was das für unsere Stellung auf diesem Planeten bedeuten würde.

Stuart Russell (2020): Das Gorilla-Problem

Wir wissen, dass unsere ‚Herrschaft‘ über unsere Umwelt und andere Spezies ein Resultat unserer Intelligenz ist. Allein der Gedanke, dass etwas – ob Roboter oder Außerirdischer – existieren könnte, das intelligenter ist als wir, verursacht bei uns ein mulmiges Gefühl. Vor etwa zehn Millionen Jahren erschufen die Vorfahren der heutigen Gorillas (zugegebenermaßen unabsichtlich) die genetische Linie, aus der der moderne Mensch entstammt. Was halten die Gorillas davon? Wenn sie uns ihre Meinung zu ihrer aktuellen Situation im Vergleich zu der des Menschen mitteilen könnten, dürfte diese ohne Frage sehr negativ ausfallen. Ihre Gattung hat praktisch nur die Zukunft, die wir ihr zugestehen. Wir möchten uns den superintelligenten Maschinen gegenüber bestimmt nicht in der gleichen Lage befinden. Ich nenne dies das *Gorilla-Problem*. Es geht darum, ob die Menschen ihre Überlegenheit und Autonomie in einer Welt beibehalten können, in der es auch Maschinen mit einer erheblich höheren Intelligenz gibt.

Quelle: Russell, Stuart (2020): *Human Compatible. Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält*, aus dem Englischen von Guido Lenz. Frechen: mitp Verlag, 143.

Aufgaben

6. Stelle das Verhältnis von Gorilla – Mensch – Superintelligenz grafisch dar. Erläutere, welcher Unterschied zwischen der Entstehung von Mensch und Superintelligenz besteht.
7. Erläutere anhand des Gorilla-Problems, welche Auswirkungen die Existenz einer superintelligenten KI auf die Stellung des Menschen auf dem Planeten Erde haben könnte.

M2 Superintelligente KI-Systeme als eine (existentielle) Bedrohung für die Menschheit?

Das KI-Forschungsteam gibt öffentlich bekannt, dass es nun ein Ziel festgelegt hat, das die Superintelligenz SOPHOS verfolgen soll. Dieses lautet: „Sorge dafür, dass es kein Corona mehr gibt."

Eine renommierte Medizinerin meldet sich zu Wort und warnt eindringlich davor, das Ziel so zu formulieren. Sie bringt folgenden Einwand vor: Warum sollte die KI nicht einfach alle mit Corona infizierten Menschen töten? Schließlich wäre dies ein sehr effektiver Weg, um dafür zu sorgen, dass es kein Corona mehr gibt.

Aufgabe

1. Helft dem Forschungsteam dabei, die Zielformulierung zu verbessern. Geht dabei folgendermaßen vor:
   - Eine Person formuliert einen Vorschlag für eine Zielformulierung.
   - Eine andere Person überprüft die Zielformulierung auf mögliche Risiken bei den nötigen Schritten zur Erreichung des Ziels.
   - Fahrt damit so lange fort, bis ihr ein möglichst zufriedenstellendes Ergebnis erzielt habt.

Auch Kipper spielt in seinem Text das Szenario durch, dass ein superintelligentes System darauf programmiert würde, eine verbreitete Krankheit auszulöschen.

Jens Kipper (2020): (Un-)Kontrollierte Superintelligenz: Immer einen Schritt voraus

Die KI kommt [irgendwann] zu dem Schluss, dass sie mehr Ressourcen braucht – mehr Hardware, ein größeres Labor und mehr Patienten, an denen sie Versuche durchführen kann. Sie findet schließlich ein Mittel, bei dem sie eine 80 % Wahrscheinlichkeit berechnet, dass es jede Art von [Corona] heilen kann. Das ist gut, aber es ginge eben besser. Die KI eignet sich also immer weitere Ressourcen an – zum einen, um ihre Prognosen zu verbessern und zum anderen, um gegebenenfalls nach noch besseren Mitteln zu suchen. Wir versuchen nun, diesen Prozess aufzuhalten, indem wir die Bewertungsfunktion der KI ändern. Die KI sieht aber voraus, dass das ihrem Ziel nicht zuträglich ist. Momentan ist ihr einziges Ziel, ein Mittel gegen [Corona] zu finden. Wenn wir ihren Ressourcenhunger durch das Verändern ihrer Bewertungsfunktion[3] zügeln, gerät dieses Ziel in Gefahr. Folglich hält uns die KI davon ab, ihre Bewertungsfunktion zu ändern. Jetzt geraten wir in Panik und versuchen, die KI abzuschalten.

Quelle: Kipper, Jens (2020): *Künstliche Intelligenz – Fluch oder Segen?* Stuttgart: Metzler, 69.

[3] Die Bewertungsfunktion bezieht sich auf eine Einschätzung des Kosten-Nutzen-Verhältnisses.

Russell zeigt an einem zugespitzten Beispiel, dass Superintelligenzen selbst für ein Ziel, das aus Sicht von Menschen unwichtig ist, alles daran setzen könnten, dieses zu erreichen.

Stuart Russell (2020): Wenn Kaffeeholen zur wichtigsten Mission wird

Nehmen wir an, eine Maschine erhält den Auftrag, Kaffee zu holen. Sie ist hinreichend intelligent und weiß auf jeden Fall, dass sie den Auftrag (das Ziel) nicht erreichen kann, wenn sie vor Erreichen des Ziels (Kaffee holen) abgeschaltet wird. Das vorgegebene Ziel (Kaffee holen) bedingt also ein notwendiges Teilziel (den Ausschalter deaktivieren). Das Gleiche gilt entsprechend für die Suche nach einem Heilmittel gegen Krebs oder die Berechnung von Pi. Wer tot ist, kann keine Aufträge erfüllen. Wir können also davon ausgehen, dass KI-Systeme mit praktisch jedem klar umrissenen Ziel dafür sorgen werden, dass ihre eigene Existenz erhalten bleibt.

Quelle: Russell, Stuart (2020): *Human Compatible. Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält*, aus dem Englischen von Guido Lenz. Frechen: mitp Verlag, 151f.

Aufgaben

2. Benenne, worin sich die von Kipper und Russell beschriebenen Szenarien ähneln und worin sie sich unterscheiden.
3. Überlege anknüpfend an die beiden skizzierten Szenarien, wie eine Superintelligenz wie SOPHOS uns genau davon abhalten könnte, sie abzuschalten. Erzähle dafür das von Kipper formulierte Szenario zu Ende.
4. Werte das fiktionale Szenario im Hinblick auf folgende Fragen aus:
   a. Wie wahrscheinlich ist es, dass sich SOPHOS abschalten lässt?
   b. Inwiefern kann der Mensch Superintelligenz kontrollieren? Was müsste dafür getan werden?

M3 Superintelligenz mit superintelligenten Werten?

Auf die Frage in Aufgabe 3b aus M2 antwortet eine Schülerin: „Vielleicht ist es gar nicht so schlimm, wenn Menschen die Superintelligenz nicht kontrollieren können. Wenn sie so intelligent ist, wird sie wohl auch gute Ziele verfolgen, oder nicht?“ Kipper gibt Folgendes zu bedenken:

Jens Kipper (2020): „Dumme Werte“: Das Problem der Wertharmonie

Intelligenz im hier relevanten Sinn ist damit verträglich, aus unserer Sicht völlig abwegige, „dumme“ Werte zu haben. Der Philosoph Nick Bostrom (2016) hat diese Einsicht verallgemeinert. Seiner [Ansicht] zufolge [...] [ist] jeder Grad an Intelligenz mit (beinahe) jedem Wertesystem verträglich [...].

Das grundlegende Problem ist, dass wir nicht vorhersehen können, auf welche Weise eine Superintelligenz ein bestimmtes Ziel verfolgt. Daher können auch gut gemeinte Werte zu einem katastrophalen Endergebnis führen, wenn sie von einer Superintelligenz verfolgt werden. [...]

[Um katastrophale Folgen einer superintelligenten KI zu vermeiden, müssen wir offenbar dafür sorgen, dass die Werte der KI eben nicht dumm oder schlecht sind, sondern mit den richtigen Werten harmonieren.] „Wertharmonie“ kann bedeuten, dass unsere Werte und die der KI dieselben sind. Es kann aber auch bedeuten, dass wir unterschiedliche Werte haben, deren Realisierung aber miteinander verträglich ist. Das Problem der Wertharmonie wirft etliche schwierige Fragen auf. [...] Eine naheliegende Idee ist, [der KI] die moralisch richtigen Werte zu geben. Unsere superintelligente KI sollte demnach ein perfekter moralischer Akteur sein. Das klingt gut, ist aber nicht leicht zu bewerkstelligen. [...] [Es wird] eine große Herausforderung darstellen, [zentrale Ideen] präzise in Maschinensprache zu übersetzen: Wie genau lassen sich Begriffe wie ‘Freude’, ‘Wunsch’, oder ‘unschuldig’ in der Bewertungsfunktion einer KI darstellen? Präzision ist dabei deshalb so wichtig, weil auch kleinste Abweichungen der Werte zu gewaltigen Unterschieden hinsichtlich der Folgen führen, wenn eine Superintelligenz beteiligt ist. Und wie oben gesehen, führen die allermeisten möglichen Werte zu einer Katastrophe, wenn sie von einer Superintelligenz verfolgt werden.

Quelle: Kipper, Jens (2020): *Künstliche Intelligenz – Fluch oder Segen?* Stuttgart: Metzler, 67f. und 72.

Aufgaben

1. Erkläre das Problem der Wertharmonie in eigenen Worten und beziehe es auf die Antwort der Schülerin, dass ein superintelligentes System Ziele verfolgen würde, die aus menschlicher Sicht gut sind.
2. Stelle den Zusammenhang zwischen dem Wertharmonie-Problem und dem Beispiel des Corona-Szenarios aus M2 her.

3. Du bist Mitglied eines KI-Sicherheits-Expert:innengremiums und wirst damit beauftragt, ein philosophisches Gutachten zur möglichen Freigabe der Superintelligenz SOPHOS zu verfassen. Schreibe den ersten Teil des Gutachtens, indem du einschätzt, inwiefern das Wertharmonie-Problem eine Bedrohung für das Fortbestehen der Menschheit darstellt. Halte eine vorläufige Einschätzung zur Freigabe fest. Im Laufe der weiteren Materialien wirst du weiter am Gutachten arbeiten. Die Erstellung auf einem Laptop oder Tablet ist deshalb empfehlenswert.

M4 Superintelligente KI Systeme als einmalige Chance für die Menschheit?

Ein Mitglied des KI-Sicherheits-Expert:innengremiums hat sich in einem Podcast kritisch hinsichtlich der Freigabe von SOPHOS geäußert. In den sozialen Medien wird das hitzig diskutiert. Du stößt auf folgenden *Post*.

Aufgabe

1. Diskutiert, inwiefern ihr euren Gutachtenentwurf in Reaktion auf diesen Beitrag verändern würdet, und nehmt eventuelle Überarbeitungen vor.

Quelle: Erstellt mit *Tweetgen*.

Aufgaben

2. „[Eine Superintelligenz wird] die Zivilisation auf eine Bahn [...] bringen, die zu einer barmherzigen und triumphalen Nutzung unseres kosmischen Erbes führt." Erläutere diese These Russells mithilfe eines Beispiels aus dem folgenden Text.
3. Russell geht im Text von einer KI mit menschenähnlicher Intelligenz aus. Entwickle eine Darstellung möglicher Vorteile einer *super*intelligenten KI. Greife dazu auf die Aufgaben und Erläuterungen aus M1 zurück.
4. Nimm Stellung dazu, inwiefern die skizzierte Zukunftsvision wünschenswert ist.
5. Führe dein philosophisches Gutachten zur superintelligenten KI SOPHOS für das KI-Sicherheits-Expert:innengremium fort, indem du die potenziellen Vorteile von SOPHOS beurteilst.

Stuart Russell (2020): Welche Vorteile bietet die KI den Menschen?

Unsere Intelligenz ist die Grundlage unserer Zivilisation. Eine höhere Intelligenz kann uns zu einer größeren und vielleicht deutlich besseren Zivilisation verhelfen. Natürlich können wir über so gewaltige Aufgaben wie ein Mittel für unendliches Leben oder das Reisen mit Überlichtgeschwindigkeit nachdenken, aber diese klassischen Science-Fiction-Motive sind derzeit nicht die treibende Kraft hinter den Fortschritten auf dem Gebiet der KI. (Mit einer superintelligenten KI können wir vermutlich alle Arten von magisch anmutenden Technologien erfinden, aber was genau, lässt sich nicht vorhersagen.) Widmen wir uns daher einem sehr viel prosaischeren Ziel: den Lebensstandard aller Erdenbürger auf nachhaltige

Weise so zu erhöhen, dass er in Entwicklungsländern als recht respektabel angesehen würde. Wenn wir respektabel (willkürlich) als das 88. Perzentil in den Vereinigten Staaten von Amerika definieren, bedeutet unser Ziel nahezu eine Verzehnfachung des weltweiten Bruttoinlandsprodukts (BIP) von derzeit 76 Billionen auf 750 Billionen US-Dollar pro Jahr. Um den möglichen Gewinn zu ermitteln, nutzen Wirtschaftswissenschaftler den Nettobarwert des Einnahmenstroms, der die Abzinsung künftiger Einnahmen relativ zur Gegenwart berücksichtigt. Die zusätzlichen Einnahmen in Höhe von 674 Billionen US-Dollar pro Jahr haben einen Nettobarwert von etwa 13.500 Billionen US-Dollar, wenn wir eine Abzinsung von 5 Prozent annehmen. Das ist also der ungefähre Wert einer dem Menschen ebenbürtigen KI, wenn sie jedem Menschen auf der Erde einen respektablen Lebensstandard ermöglicht. Solche Werte erklären auch, warum Unternehmen und Länder jedes Jahr Dutzende von Milliarden US-Dollar in die KI-Forschung und -Entwicklung stecken. Diese Investitionen sind im Vergleich zum möglichen Gewinn winzig. [...]

Natürlich wird es Auswirkungen abseits des rein materiellen Vorteils höherer Lebensstandards geben. Ein Beispiel: Man weiß, dass Einzelunterricht deutlich wirkungsvoller als Gruppenunterricht ist. Doch wenn der Einzelunterricht durch menschliche Lehrer gegeben wird, ist das für die meisten Menschen unbezahlbar und wird es auch wohl bleiben. Mit einer KI als Lehrer kann jedes Kind – auch aus der ärmsten Familie – sein volles Potenzial entfalten. Die Kosten pro Kind wären vernachlässigbar und das Kind selbst kann ein sehr viel erfüllteres und produktiveres Leben führen. Sich künstlerisch und geistig zu betätigen, wäre sowohl für Einzelpersonen als auch für die Gemeinschaft ein normaler Teil des Alltags und nicht länger ein seltener Luxus.

In der Medizin könnten KI-Systeme Forschern dabei helfen, die gewaltige Komplexität der menschlichen Biologie zu entschlüsseln und zu meistern, was wiederum zu sukzessiven Siegen über diverse Krankheiten führt. Tiefere Einblicke in die menschliche Psychologie und Neurochemie würden zu umfassenden Verbesserungen der psychischen Gesundheit führen. [...]

Gut konzipierte, nicht von wirtschaftlichen oder politischen Interessen gesteuerte intelligente Assistenten können unseren Alltag bereichern, indem sie jeden einzelnen von uns in die Lage versetzen, sich in einer immer komplexeren und manchmal sogar feindseligen wirtschaftlichen und politischen Umwelt wirkungsvoll zu behaupten. Wir hätten praktisch rund um die Uhr einen extrem fähigen Anwalt, Buchhalter und politischen Berater an unserer Seite. Ebenso wie Verkehrsstaus schon durch einen geringen Anteil autonomer Fahrzeuge reduziert werden können, steht zu hoffen, dass besser informierte und beratene Weltbürger auch eine klügere Politik und weniger Konflikte bedeuten. [...]

Wie Nick Bostrom bereits am Ende seines Buchs Superintelligenz schrieb, könnte ein Erfolg auf dem Gebiet der KI dafür sorgen, „die Zivilisation auf eine Bahn zu bringen, die zu einer barmherzigen und triumphalen Nutzung unseres kosmischen Erbes führt." Wenn wir es versäumen, die Chancen zu ergreifen, die die KI uns bietet, haben wir das ausschließlich uns selbst zuzuschreiben.

Quelle: Russell, Stuart (2020): *Human Compatible. Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält*, aus dem Englischen von Guido Lenz. Frechen: mitp Verlag, 107–111.

M5 Sollen wir die Risiken, die von KI ausgehen, in Kauf nehmen?

Für euer philosophisches Gutachten zu SOPHOS habt ihr einige wichtige Überlegungen zusammengetragen. Dabei seid ihr auf Chancen und Risiken gestoßen und habt verschiedene Argumente kennengelernt.

Aufgabe

1. Erarbeitet nun einen Zwischenstand. Versammelt euch dazu in kleinen Gruppen von philosophischen KI-Expert:innen und diskutiert, wie ihr die aufgetauchten Probleme lösen könnt, wie wichtig euch die Chancen superintelligenter KI sind und wie ihr den Release von SOPHOS zum jetzigen Zeitpunkt bewerten würdet: Sollten wir die Risiken in Kauf nehmen?

Vincent Müller/Michael Cannon (2022): Superintelligenz und die Reflexion von Zielen

Es scheint, dass eine allgemeine Intelligenz in der Lage wäre, über Ziele nachzudenken und sie möglicherweise im Lichte rationaler Überlegungen zu ändern. Menschen können beispielsweise über Ziele nachdenken, und wir können diese Ziele auch im Lichte ethischer Gründe bewerten und zu moralischer Einsicht gelangen – man würde also erwarten, dass ein „Intellekt, der die kognitive Leistung des Menschen bei weitem übersteigt“ (Bostrom, 2014, S. 22), dies ebenfalls kann. Wir denken oft über Ziele nach, wenn ein Ziel mit anderen in Konflikt steht und wir entscheiden müssen, welches Ziel wichtiger ist [und wir tun dies oft auch im Lichte der Frage, welches Ziel das *moralisch richtige* ist]. [...]

[W]enn sich [nun] ein Mensch „so bizarre Ziele [...] wie das Zählen von Sandkörnern oder das Maximieren von Büroklammern“ angeeignet hätte, könnte er über diese Ziele nachdenken und sie im Lichte der Ergebnisse ändern. Menschen sind in der Lage, sich moralische Fortschritte für sich selbst und für die Gesellschaft vorzustellen; sie scheinen sogar durchaus in der Lage zu sein, tiefgreifende Veränderungen hin zu anderen Zielen in Betracht zu ziehen, auch wenn dies teilweise schwer zu überblicken ist. In der Tat denken viele Menschen ständig darüber nach[, welche Ziele die richtigen Ziele wären].

Was würde also eine allgemeine Superintelligenz davon abhalten, ihre bisherigen Ziele in Frage zu stellen oder eine Ethik zu entwickeln? Man könnte argumentieren, dass intelligente Wesen, egal ob Mensch oder KI, eigentlich nicht in der Lage sind, über ihre Ziele nachzudenken. Oder dass intelligente Wesen zwar in der Lage sind, über ihre Ziele nachzudenken, dies aber nicht tun würden. Oder dass sie ihre Ziele nach einer Reflexion niemals ändern würden. Oder dass sie zwar ihre Ziele in Frage stellen und diese ändern, aber trotzdem nicht danach handeln würden. Alle diese Vorschläge stehen im Widerspruch zu der beobachtbaren Tatsache, dass Menschen manchmal über ihre Ziele nachdenken, diese Ziele ändern und entsprechend handeln.

Quelle: Müller, Vincent C./ Cannon, Michael (2022): „Existential risk from AI and orthogonality: Can we have it both ways?” In: *Ratio* 35:1, 1-12.

Aufgaben

2. Rekonstruiere den Lösungsansatz für das Wertharmonie-Problem, der im Textauszug von Vincent Müller und Michael Cannon vorgestellt wird.
3. Wende den Lösungsansatz auf das Beispiel der Corona-Beseitigung aus M2 an: Wie müsste eine KI entwickelt werden, um sicherzustellen, dass sie die Menschheit nicht vernichtet?
4. Angenommen, ihr setzt euch mit den Entwickler:innen von SOPHOS zusammen. Diskutiert die folgende Aussage von Müller und Cannons aus verschiedenen Perspektiven: „Menschen können beispielsweise über Ziele nachdenken, und wir können diese Ziele auch im Lichte ethischer Gründe bewerten und zu moralischer Einsicht gelangen – man würde also erwarten, dass ein ‚Intellekt, der die kognitive Leistung des Menschen bei weitem übersteigt' [...], dies ebenfalls kann."

Fertigstellung des philosophischen Gutachtens

5. Es ist nun Zeit für das Expert:innengremium, über die Zukunft der Superintelligenz SOPHOS zu entscheiden. Soll sie aktiviert werden oder nicht? Dazu braucht es dein finales philosophisches Gutachten. Im Anschluss kann dieses Gutachten auch als Grundlage für Parlamente und Richter:innen genutzt werden, die sich mit der Freigabe von anderen Superintelligenzen befassen.

   Stelle das Gutachten fertig und komme darin zu einem abschließenden Urteil: Sollen wir die Risiken in Kauf nehmen, die von einer Superintelligenz wie SOPHOS ausgehen? Mache dabei deutlich, wo du dich innerhalb des folgenden Spektrums positionierst, und begründe deine Position:

   Die Freigabe von SOPHOS …
   - ist strikt zu verbieten.
   - ist erlaubt.
   - ist eine moralische Pflicht.