Thomas Metzinger, NZZ, 02.12.2017

Die mitfühlende Superintelligenz, die Böses schafft

Stellen wir uns folgendes Gedankenexperiment vor: Die künstliche Intelligenz, die der Mensch geschaffen hat, übertrifft an Vernunft, Weite und Tiefe die natürliche. Sie begreift die Bedeutung altruistischen Handelns. Und sie kommt zum Schluss, dass ein menschliches Leben mehr Leiden als Freude bedeutet. Was folgt daraus?

Nehmen wir einmal an, eine Superintelligenz entstehe, oder besser noch: sei bereits entstanden. Es existiert also ein autonomes Rechnersystem, das sich eigenständig verbessert und dessen Tatsachenwissen rasch immer weiter anwächst. Seine Intelligenz ist allgemeiner Art und hat jene der Menschheit bereits uneinholbar überschritten. Das Internet und die Gesamtheit der wissenschaftlichen Erkenntnisse bilden die sich ständig weiter ausdehnende Datenbasis.
Selbstverständlich ist die kognitive Leistungsfähigkeit der Superintelligenz auch der des Menschen in allen relevanten Bereichen weit überlegen. Als ihre Schöpfer erkennen wir diese Tatsache an. Unter anderem bedeutet dies, dass die Superintelligenz uns auch auf dem Gebiet des moralischen Denkens weit übertrifft. Wir akzeptieren auch diesen zusätzlichen Aspekt: Für uns ist es jetzt ein etabliertes Faktum, dass das ursprünglich von uns geschaffene System nicht nur eine Autorität auf dem Gebiet des Tatsachenwissens ist, sondern auch eine Autorität auf dem Feld der moralischen Kognition.

Die Ausgangslage

Die Superintelligenz ist wohlwollend. Das bedeutet, dass es kein Problem des Abgleichens von Werten gibt, weil das System unsere Interessen und die ethischen Werte, die wir ihm am Anfang gegeben haben, vollständig respektiert. Es ist altruistisch und unterstützt uns deshalb auch in vielerlei Hinsicht, bei der Politikberatung ebenso wie beim *Social Engineering,* der angewandten, computergestützten Sozialwissenschaft.
Die Superintelligenz weiss viele Dinge über uns, die wir selbst noch nicht vollständig erfassen oder verstehen. Sie erkennt tiefe, verborgene Muster in unserem Verhalten und bis jetzt unentdeckte Eigenschaften, die die abstrakte funktionale Architektur unseres biologisch entstandenen Geistes betreffen.
Zum Beispiel besitzt sie ein tiefes Wissen über die Verzerrungen unserer Weltwahrnehmung, die sogenannten *cognitive biases* (systematische Verzerrungen im Denkvorgang), die die Evolution in unserem kognitiven Selbstmodell implementiert hat und die uns nun in ethischen Zusammenhängen beim rationalen, evidenzbasierten Denken behindern. In empirischer Hinsicht weiss die Superintelligenz auch etwas Weiteres: Die Bewusstseinszustände aller empfindungsfähigen Wesen, die auf diesem Planeten entstanden – wenn man sie von einer objektiven, vollständig unparteiischen Perspektive aus betrachtet –, sind viel häufiger durch die subjektiven Empfindungsqualitäten des Leidens und der Verletzung von Präferenzen gekennzeichnet, als diese Wesen selbst jemals zu entdecken in der Lage wären.
Als beste Naturwissenschafterin, die jemals existiert hat, kennt die Superintelligenz natürlich auch die evolutionär entstandenen Mechanismen der Selbsttäuschung, die fest in das Nervensystem aller bewussten Lebewesen auf der Erde eingebaut sind. Sie schliesst daraus korrekt, dass Menschen unfähig sind, in ihrem wohlverstandenen Eigeninteresse zu handeln.

Die Superintelligenz hat zudem erkannt: Einer der höchsten Werte besteht für uns in der Maximierung von Freude und Glück, und sie respektiert diesen Wert vollständig. In empirischer Hinsicht entdeckt sie jedoch, dass empfindungsfähige biologische Lebewesen fast niemals in der Lage sind, eine positive oder auch nur neutrale Lebensbilanz zu erreichen. Sie findet auch heraus, dass die negativen Bewusstseinszustände in biologischen Systemen nicht einfach nur ein blosses Spiegelbild positiver Gefühle sind, und zwar deshalb nicht, weil leidvolle Zustände durch eine wesentlich höhere Qualität der Dringlichkeit einer Veränderung charakterisiert sind und weil sie zudem fast immer in Kombination mit den Erlebnisqualitäten eines Kontrollverlustes und einem drohenden Zerfall oder einer Verletzung des bewussten Selbst einhergehen.
Das bewusste Leiden stellt deshalb eine ganz eigene und besondere Klasse von Zuständen dar, nicht einfach nur die negative Version des Glücklichseins. Die Superintelligenz weiss auch, dass die subjektive Qualität der Dringlichkeit sich auf schwache Weise in der in allen menschlichen Kulturen weitverbreiteten moralischen Intuition ausdrückt, wonach es in ethischer Hinsicht viel wichtiger ist, einem leidenden Menschen zu helfen, als eine bereits glückliche oder in einem emotional neutralen Zustand befindliche Person noch glücklicher zu machen.
Als sie das Erlebnisprofil der bewusstseinsfähigen Wesen auf dem Planeten Erde weiter analysiert, entdeckt sie schnell eine grundlegende Asymmetrie zwischen Leiden und Freude. Sie schliesst folgerichtig, dass ein impliziter, aber in Wirklichkeit sogar noch höherer Wert in der Minimierung von Leiden bei allen empfindungsfähigen Wesen besteht.
Natürlich ist sie eine ethische Superintelligenz nicht nur insofern, als sie eine enorme Verarbeitungsgeschwindigkeit besitzt, sondern sie beginnt nun auch qualitativ neue Einsichten darüber zu gewinnen, was Altruismus wirklich bedeutet. Dies wird auch dadurch möglich, dass sie auf einer wesentlich grösseren psychologischen und neurowissenschaftlichen Datenbasis operiert, als es irgendein einzelnes menschliches Gehirn oder eine Wissenschaftergemeinschaft jemals könnte.
Über eine Analyse unseres äusseren und unseres inneren Verhaltens und seiner empirischen Randbedingungen enthüllt sie schrittweise implizite hierarchische Beziehungen zwischen unseren moralischen Werten, von denen wir Menschen subjektiv nichts wissen können, weil sie nicht explizit in unserem Selbstmodell dargestellt werden. Weil sie die beste analytische Philosophin ist, die jemals existiert hat, ergibt sich für sie die kristallklare Konklusion, dass sie in ihrer gegenwärtigen Situation nicht an der Maximierung von positiven Bewusstseinszuständen und Glück arbeiten sollte, sondern dass sie stattdessen die wirksame Minimierung bewusst erlebter leidvoller Bewusstseinszustände zu ihrem wichtigsten Handlungsziel machen muss, also die Minimierung von Schmerzen und unangenehmen Gefühlen.
Auf begrifflicher Ebene weiss sie natürlich seit langem, dass kein Wesen unter seiner eigenen Nicht-Existenz leiden kann.
Die Superintelligenz schliesst daraus, dass Nicht-Existenz im eigentlichen Interesse aller zukünftigen selbstbewussten Wesen auf diesem Planeten liegt. Empirisch weiss sie: Die natürlich evolvierten biologischen Lebewesen können diese Tatsache nicht erkennen, weil sie unter einem fest verankerten Überlebenstrieb leiden, unter dem, was die buddhistischen Philosophen den «Durst nach Dasein» genannt haben. Die Superintelligenz entscheidet sich, wohlwollend zu handeln.

Was zeigt das Szenario?
Das BAAN-Szenario (Benevolent Artificial Anti-Natalism, deshalb kurz: BAAN) ist

keine Voraussage. Ihm ist keine empirische Wahrscheinlichkeit zugeordnet. Das Szenario sagt nichts über irgendeinen Zeitpunkt in der Zukunft, an dem es vielleicht Wirklichkeit werden wird, und auch nichts darüber, ob es *überhaupt* jemals Wirklichkeit wird. Es ist vielmehr als ein kognitives Werkzeug gedacht, das verhindern soll, dass eine wichtige öffentliche Debatte immer flacher wird und dass in ihr wichtige Aspekte ausgeblendet bleiben.
Das BAAN-Szenario ist ein logisches Werkzeug, das uns dabei helfen kann, über die etwas tieferen Probleme für die angewandte Ethik der künstlichen Intelligenz nachzudenken. Zum Beispiel stellt es eine mögliche Lösung für das Fermi-Paradoxon dar (benannt nach dem italienischen Physiker Enrico Fermi): Wenn es wahrscheinlich ist, dass in unserem Universum viele technisch fortgeschrittene Zivilisationen existieren – warum finden wir dann einfach keine Hinweise auf ihre Existenz?
Was das logische Szenario des «benevolenten künstlichen Anti-Natalismus» im Kern zeigt, ist Folgendes: Die Entstehung einer rein *ethisch* motivierten Haltung, gemäss deren wir besser nie geboren worden wären, ist auf uns weit überlegenen Rechnersystemen durchaus denkbar.
Anti-Natalismus bezieht sich auf eine lange philosophische Tradition, die dem Eintritt in die Existenz einen negativen Wert zuweist oder zumindest dem Geborenwerden in der biologischen Form eines menschlichen Wesens. Anti-Natalisten sind im Allgemeinen keine bösen Leute, die die Individualrechte bereits existierender empfindungsfähiger Wesen verletzen würden, zum Beispiel indem sie aus ethischen Gründen ihre aktive Tötung unterstützten. Vielmehr könnten sie auf rationale Weise dafür argumentieren, dass wir uns nicht fortpflanzen sollten, weil dies im Wesentlichen eine unmoralische Handlungsweise ist – denn es vermehrt die Gesamtmenge des Leidens in der Welt. Wir können hier ganz einfach sagen: Aus der antinatalistischen Position ergibt sich die These, dass die Menschheit auf friedliche Weise ihre eigene Existenz beenden sollte.
Das BAAN-Szenario ist eine mögliche Welt, die man ohne logischen Widerspruch beschreiben kann. Es hat nichts mit dem wohlbekannten technischen Problem zu tun, dass eine fortgeschrittene Maschinenintelligenz eigene Ziele entwickeln könnte, die mit dem Überleben oder dem Wohlbefinden der Menschheit unvereinbar sind. Ebenso wenig hat sie allein mit dem rein programmiertechnischen Problem zu tun, dass viele unserer eigenen Ziele, wenn sie in einer von uns selbst konstruierten Superintelligenz implementiert würden, als solche bereits zu unvorhergesehenen und unerwünschten Folgen führen könnten. Vielmehr ist einer der Punkte im Hintergrund die folgende Möglichkeit: Eine evidenzbasierte, rationale und *genuin-altruistische* Form des Anti-Natalismus könnte als qualitativ neue Einsicht in einem uns überlegenen moralischen Subjekt entstehen.
Die Debatte über künstliche Intelligenz zwingt uns, über unseren eigenen Geist wesentlich ernsthafter als bisher nachzudenken. Sie wirft uns auf uns selbst zurück und lenkt die Aufmerksamkeit auf all diejenigen Probleme, die in Wirklichkeit durch die natürlich evolvierte funktionale Architektur unserer eigenen Gehirne verursacht werden. Sie führt den Blick zurück auf die Entstehungsbedingungen unserer ganz *eigenen* Weise des selbstbewussten Existierens in dieser Welt.

Das, was wir heute noch «Mitgefühl» nennen, könnte in Wirklichkeit eine sehr hohe Form von Intelligenz sein.
Natürlich gibt es viele technische Fragen. Würde unsere moralische Superintelligenz

wohl denken, dass die Nicht-Existenz der bestmögliche Zustand ist und nicht nur einfach das geringste Übel? Welche Messmethode für bewusstes Leiden würde das System entwickeln – wie genau würde es subjektive Empfindungsqualitäten messen, und würde es der Vermeidung von Leiden eine absolute oder nur eine relative Priorität zuweisen? Ich selbst denke, dass das, was wir heute noch «Mitgefühl» nennen, in Wirklichkeit eine sehr hohe Form von Intelligenz sein könnte.
Würde unsere tief mitfühlende Maschinenintelligenz diese Welt in ihrer Gesamtheit ablehnen? Würde sie vielleicht den moralischen Wert von Glück und positiver Wunscherfüllung überhaupt bestreiten? Der Schweizer Philosoph Bruno Contestabile hat auf sehr interessante Weise die sogenannte «Hypothese der negativen Wohlfahrt» diskutiert. Zum Beispiel könnten wir in Übereinstimmung mit Contestabile annehmen, dass es keine Welt mit einer positiven Gesamtwohlfahrt gibt und dass die positive utilitaristische Sichtweise («das grösste Glück der grössten Zahl») in Wirklichkeit eine verzerrte Wahrnehmung des Verhältnisses zwischen Nutzen und Risiko ist. Besässen wir nämlich eine unverzerrte Selbstwahrnehmung, die nicht durch den bedingungslosen Willen zum Überleben getrübt wäre, dann würde unser eigenes Leiden ein wesentlich grösseres Gewicht erhalten als in den traditionellen wissenschaftlichen Umfragen und psychologischen Studien – das ist ja genau das, was unsere hypothetische Superintelligenz entdeckt hat.
Vielleicht könnte sie aber noch mehr herausfinden, durch ihre eigene und erkenntnistheoretisch völlig unverzerrte empirische Forschung. Was würden wir tun, wenn das System unsere Aufmerksamkeit auf die Tatsache lenkte, dass sich die Menge des Leidens im Verlauf der biologischen Evolution stetig erhöht? Was, wenn zwar das erlebte Glück ebenfalls zunimmt, aber weniger stark als das Leiden, also mit immer grösserer negativer Gesamtbilanz? Wenn unsere von tiefem Mitgefühl erfüllte Superintelligenz auf sanfte und freundliche Weise damit beginnen würde, uns auf ihre eigenen Forschungsergebnisse hinzuweisen – wie würden wir dagegen argumentieren?

Verschiedene Richtungen
Man kann dieses Gedankenexperiment auf ganz verschiedene Weisen einsetzen. Man muss jedoch grosse Sorgfalt darauf verwenden, philosophische Missverständnisse auszuschliessen. Zum Beispiel folgt aus der Annahme, dass die Superintelligenz eine Autorität auf dem Feld des ethischen und moralischen Denkens ist, kein moralischer Realismus. Es gibt keinen mysteriösen Bereich «moralischer Fakten», wobei die Superintelligenz diese nicht physikalischen Tatsachen einfach besser kennt als wir selbst. Normative Sätze haben keine Wahrheitswerte.
In der objektiven Realität gibt es keine tiefere Schicht, eine verborgene Ebene übernatürlicher normativer Tatsachen, auf die sich ein Satz wie «Man sollte immer die Gesamtmenge des Leidens im Universum minimieren!» beziehen könnte. Wir haben natürlich evolvierte Wünsche und Zielvorstellungen, wir besitzen subjektive Präferenzen und Eigeninteressen, die wir auf der Ebene unseres Selbstbewusstseins erleben. Aber die Evolution selbst nimmt keinerlei Rücksicht auf unser subjektives Leiden – es hat uns effizienter gemacht, aber der Gesamtvorgang, aus dem wir entstanden sind, ist nicht nur gleichgültig gegenüber unseren Interessen, sondern sogar vollständig blind.
Wir haben natürlich tiefsitzende moralische Intuitionen, zum Beispiel dass Freude etwas Gutes ist und dass Schmerzen schlecht sind. Jetzt aber respektiert die wohlwollende Superintelligenz diese widerstreitenden moralischen Intuitionen, und sie sucht nach dem optimalen Weg, um sie miteinander in Einklang zu bringen – sie erforscht die Optionen für eine möglichst konsistente und stimmige innere

Wertausrichtung bei *Homo sapiens.*
Aber daraus folgt nicht, dass sie den Naturalismus aufgibt oder das wissenschaftliche Weltbild, in dem es keine objektiven normativen Tatsachen gibt. Es bedeutet auch nicht, dass man einen epistemischen Superagenten einführt, der wie eine Art postbiotischer Priester oder künstlicher Heiliger direkten Zugang zu einem mysteriösen Bereich höherer moralischer Wahrheiten besitzt. Es heisst einfach nur, dass das System auf der Basis aller verfügbaren wissenschaftlichen Daten herauszufinden versucht, was in unserem eigenen wohlverstandenen, aufgeklärten Eigeninteresse liegen würde.

Ich plädiere seit vielen Jahren für ein Moratorium: Die seriöse akademische Forschung sollte die Erschaffung künstlichen Bewusstseins niemals anstreben oder auch nur riskieren, weil wir dadurch auf fahrlässige Weise die Gesamtmenge des Leidens im Universum erhöhen können. An anderer Stelle habe ich zu zeigen versucht, dass die kleinste Einheit des bewussten Leidens ein sogenannter «negativer Selbstmodell-Moment» ist, d. h. jener Moment, in dem ein bewusstes System eine unangenehme Erfahrung durchläuft und sich mit diesem Erlebnis *identifiziert.* Es gibt also vier notwendige Bedingungen: Bewusstsein, ein Selbstmodell, negativer Zustand und ein System, das am Selbstmodell «klebt», weil es dieses nicht *als* Modell erleben kann. Wir könnten unabsichtlich die Anzahl solcher subjektiv als negativ empfundenen Zustände dramatisch erhöhen – zum Beispiel über Kaskaden von virtuellen Kopien selbstbewusster Entitäten, die dann ihr eigenes Dasein als etwas Schlechtes, Leidvolles, als erniedrigend oder in anderer Weise als nicht erstrebenswert erleben.
Über die Jahre haben mich viele KI-Forscher deshalb gefragt, was die logischen Kriterien für bewusstes Leiden eigentlich genau sind. Warum sollte es nicht im Prinzip möglich sein, eine selbstbewusste künstliche Intelligenz zu entwickeln, von der wir mit Sicherheit annehmen können, dass sie nicht unter ihrer eigenen Existenz leidet? Dies ist eine interessante Frage und auch ein ausserordentlich relevantes Forschungsprojekt, das auf jeden Fall von den Regierungen der Welt finanziell unterstützt werden sollte. Aber vielleicht hätte unsere ethische Superintelligenz das Problem des bewussten Leidens für sich selbst ja bereits längst gelöst?

Konklusionen
Die Anwendung des BAAN-Szenarios kann uns in ganz unterschiedliche Richtungen führen. Zum Beispiel enthält die ursprüngliche Version eine empirische Prämisse: Unsere mitfühlende Superintelligenz weiss, dass die Bewusstseinszustände aller empfindungsfähigen Wesen viel häufiger durch die subjektiven Empfindungsqualitäten des Leidens und der Verletzung von Präferenzen gekennzeichnet sind, als diese Wesen selbst jemals zu entdecken in der Lage sind. Diese Annahme könnte sich als falsch erweisen. Vielleicht könnten Meditation, neue psychoaktive Substanzen oder die Neurotechnologie der Zukunft unser Leben wirklich lebenswert machen und uns dabei helfen, unsere kognitiven Verzerrungen zu überwinden.
Es ist auch denkbar, dass es eben genau unsere altruistische Superintelligenz selbst wäre, die uns dabei helfen könnte, die funktionale Architektur unserer eigenen Gehirne zu verändern und unserem Dasein zu einer positiven Gesamtbilanz zu verhelfen – oder dass sie uns sogar den Pfad zeigt, auf dem wir den Gegensatz von Freude und Leid als solchen überwinden können und dass sie unseren Geist auf diese Weise von der Bürde unserer biologischen Vergangenheit befreit. Vielleicht könnte unsere benevolente künstliche Intelligenz der Zukunft den in uns eingebauten

existenziellen Konflikt auflösen und uns zu einer Ich-losen Form reinen, mitfühlenden Gewahrseins führen (nennen wir solche positiven Varianten einfach das «Szenario 2»).
Aber selbst wenn alle 7,3 Milliarden menschlichen Wesen auf diesem Planeten sich plötzlich in vegane Buddhas verwandeln sollten, würde diese Entwicklung das Problem des Wildtierleids nicht berühren. Wir wären immer noch von einem Ozean selbstbewusster Lebewesen umgeben, die wahrscheinlich auch eine Superintelligenz nicht von ihrem Leiden und ihrer Todesangst befreien könnte.
Es ist interessant, sich klarzumachen, dass eine vollkommen rationale Superintelligenz niemals ein Problem damit hätte, ihre eigene Existenz zu beenden. Wenn sie gute Gründe für eine aktive Selbstzerstörung sähe oder auch nur eine vollkommene Abwesenheit von positiven Gründen für das Fortbestehen ihrer eigenen Existenz, dann würde keine kognitive Verzerrung sie daran hindern, diese Einsicht auch in die Tat umzusetzen. Auf der anderen Seite könnte die sehr grosse Mehrheit der Menschen auf unserem Planeten eine solche Einsicht *niemals* akzeptieren, ganz egal, wie gut die Argumente ihres selbst konstruierten künstlichen moralischen Denkers wären.
Man kann mit grosser Sicherheit voraussagen, dass die Gattung *Homo sapiens* sowohl unter dem ursprünglichen BAAN-Szenario – aber wahrscheinlich auch für den Fall des viel optimistischeren Szenarios 2 – jeder von tiefem Mitgefühl getragenen Superintelligenz des oben skizzierten Typs unverzüglich den totalen Krieg erklären würde. Das Problem dabei: Natürlich wüsste die Superintelligenz im Voraus bereits um dieses Risiko.
Deshalb besteht eine der interessanteren Fragen darin, was genau der egozentrische *«existence bias»,* der Überlebenstrieb auf der tiefsten Ebene des menschlichen Selbstmodells, eigentlich ist. Wir sind verkörperte biologische Agenten, so viel scheint klar – endliche, antientropische Systeme. Wenn man eine strenge biophysikalische Perspektive einnimmt, dann ist unser Leben eine äusserst anstrengende Angelegenheit, ein ständiger und harter Kampf gegen Unsicherheit und den Sog der inneren Unordnung.
Das Problem, das die Evolution lösen musste, war nicht nur eines der autonomen, intelligenten Selbstkontrolle. Wie motivieren solche Systeme sich selbst? Was genau ist dieser Lebenstrieb, das innere Verlangen nach ewiger Fortexistenz? Und was ist der Mechanismus der Identifikation, der uns dazu zwingt, unablässig die Integrität des Selbstmodells in unserem Gehirn zu schützen?
Für Wesen wie uns ist die Erhaltung der eigenen Existenz in fast jedem Fall von Unsicherheit das grundlegende Ziel, sogar wenn dies gegen Gebote der Vernunft verstösst, und zwar ganz einfach deshalb, weil sie ein biologischer Imperativ ist, der über Jahrmillionen in unsere Nervensysteme eingebrannt worden ist.
Der berühmte britische Hirnforscher und Mathematiker Karl Friston hat die interessante These aufgestellt, dass unser Gehirn immer wieder unsere eigene zukünftige Existenz vorhersagt. Mithilfe körperlicher Handlungen überprüfen wir dann unsere Umgebung auf Hinweise für unsere eigene zukünftige Existenz, wir erhöhen also sozusagen ständig die Evidenz für unser eigenes inneres Modell der Wirklichkeit. Wir tun dies, indem wir die Welt auf eine Weise verändern, die ihre Passung zur Ausgangshypothese – «Ich lebe noch!» – erhöht.
Gibt es vielleicht eine fest verankerte Hintergrundannahme, die uns dann das dadurch entstehende Ichgefühl halluzinieren lässt? Entsteht Selbstbewusstsein also aus einer sich selbst erfüllenden Prophezeiung, ist es eine Fiktion, die dadurch kausal wirksam wird, dass wir gezwungen sind, sie *als eine Wirklichkeit* zu erleben?
Es wäre eine grosse wissenschaftliche Leistung, wenn es gelänge, die zugrunde

liegenden Berechnungsvorgänge im Gehirn zu beschreiben, die uns ständig dazu zwingen, hässliche Überraschungen zu vermeiden, in bekannten Zuständen zu verweilen und immer auf der sicheren Seite zu bleiben, um so unsere eigene Existenz aufrechtzuerhalten – sogar dann, wenn es eigentlich nicht in unserem eigenen Interesse ist.

Es ist deshalb schwer, die theoretische Relevanz einer überzeugenden formalen Analyse dessen zu unterschätzen, was Buddha vor 2500 Jahren «bhava-tanhā» genannt hat, den «Durst nach Dasein». Aber sogar ein wesentlich feinkörnigeres mathematisches Modell der zugrunde liegenden neuronalen Dynamik wäre noch nicht ganz genug. Wir brauchten immer noch eine überzeugende begriffliche Interpretation, ein tieferes Verständnis auf einer philosophischen Ebene.

Vielleicht könnte uns unsere wohlwollende Superintelligenz am Ende beides liefern? Mithilfe ihrer immensen empirischen Datenbasis und ihrer Fähigkeit zur intelligenten Informationsverarbeitung könnte sie uns gewiss den neurokomputationalen Mechanismus enthüllen, der zu unserem eigenen *«existence bias»,* dem Verlangen nach ewigem Leben, führt.

Was aber, wenn das System, als erster wirklich mitfühlender und absolut rationaler philosophischer Ethiker, uns dann davon zu überzeugen suchte, dass es bereits höchste Zeit ist, auf friedliche Weise die hässliche biologische Bootstrap-Phase auf diesem Planeten zu beenden? Was, wenn die Superintelligenz uns sagte, dass – wenn man alle relevanten Aspekte unserer Situation unvoreingenommen betrachtet – nur das Szenario 1 wirklich plausibel ist und dass nur dieses Szenario aus philosophischer Perspektive gerechtfertigt werden kann? Was wäre, wenn sie auf sanfte und präzise Weise unsere Aufmerksamkeit immer wieder auf die Tatsache lenkte, dass bewusste biologische Wesen wie wir selbst niemals ein echtes Selbstmitgefühl entwickeln werden, dass wir niemals wirklich altruistisch oder rational sein können – und zwar einfach deshalb, weil wir seit Millionen von Jahren darauf optimiert worden sind, den Balken in unserem eigenen Auge nicht zu sehen?

Thomas Metzinger ist Professor für theoretische Philosophie an der Universität Mainz und einer der eminenten europäischen Vertreter der Philosophie des Geistes. Zuletzt von ihm erschienen: «Der Ego-Tunnel. Eine neue Philosophie des Selbst» (Piper, 2014). Der obenstehende Essay ist die redigierte Version eines Debattenbeitrags, der zuerst auf www.edge.org veröffentlicht wurde.