Postbiotisches Bewusstsein:

Wie man ein künstliches Subjekt baut – und warum wir es nicht tun sollten

Thomas Metzinger

Was wären die Bedingungen dafür, dass wir von einem künstlichen bzw. nichtbiologischen System annehmen, dass es bewusste Erlebnisse besitzt? Wodurch wird aus einem informationsverarbeitenden System ein Subjekt von Erfahrung? Und: Wann wären wir in der Annahme gerechtfertigt, dass es auch ein bewusstes *Selbst* und eine echte, bewusst erlebte *Innenperspektive* besitzt? Der Turing-Test (Turing 1950) ist sicherlich zu schwach, weil er weder ein Kriterium für echte Intelligenz, noch eines für das Vorhandensein von intentionalem Gehalt (also: geistig repräsentiertem Wissen über die Welt) liefert – und schon gar nicht für das, was Philosophen *phänomenalen* Gehalt nennen.

Phänomenaler Gehalt entsteht dann, wenn sich die repräsentationalen Zustände eines informationsverarbeitenden Systems für dieses selbst irgendwie *anfühlen*, also wenn sie einen introspektiv zugänglichen qualitativen Charakter besitzen. Dass ein beliebiges System sich so verhält, *als ob* es echte Farben sehen oder wirklichen Schmerz empfinden könnte, ist bei näherem Hinsehen ein genauso unbefriedigendes Kriterium wie die Tatsache, dass es vielleicht auf sprachlicher Ebene *behauptet*, es hätte tatsächliche bewusste Erlebnisse (für einige lustige Gedankenexperimente, die diesen Punkt zu illustrieren versuchen, vgl. Chalmers 1995). Ein wesentlich stärkerer und deshalb vielleicht besserer Test für phänomenales Bewusstsein ist der Metzinger-Test: Wir sollten ein System spätestens dann als bewusstes Objekt behandeln, wenn es uns gegenüber auf überzeugende Weise demonstriert, dass die philosophische Frage nach dem Bewusstsein für es selbst ein Problem geworden ist, zum Beispiel wenn es eine eigene *Theorie* des Bewusstseins vertritt, d.h. wenn es mit eigenen *Argumenten* in die Diskussion um künstliches Bewusstsein einzugreifen beginnt. Was aber ist mit menschlichen Kleinkindern oder mit den vielen empfindungsfähigen Tieren auf unserem Planeten? Sicherlich kann man bewusst sein – z.B. Freude und Schmerz empfinden – ohne denken zu können, sicher sind auch Systeme mit einer sehr niedrigen Intelligenz leidensfähig und auf jeden Fall muss uns eine gute Theorie des Bewusstseins erklären können, was genau der Punkt ist, an dem

Prof. Dr. Thomas Metzinger

1958 in Frankfurt geboren

1979–85 Studium der Philosophie, Ethnologie und Theologie an der Universität Frankfurt a.M.

1985 Promotion an der Universität Frankfurt a.M.

1992 Habilitation an der Universität Gießen

1995–97 Vertretung von Lehrstühlen an den philosophischen Instituten der Universitäten Osnabrück und Saarbrücken

1997–98 Fellow am Hanse-Wissenschaftskolleg Bremen-Delmenhorst

1998–99 Forschungsaufenthalt am Philosophy Department der University of California at San Diego, USA

1999–2000 Vertretung eines Lehrstuhls am Philosophischen Institut der Universität GH Essen

2000 Professor für Philosophie der Kognition an der Universität Osnabrück

Seit Oktober 2000 Professor für Theoretische Philosophie an der Universität Mainz

im Laufe der biologischen Evolution Empfindungsfähigkeit und phänomenales Erleben entstanden sind.

Solche Überlegungen an der Schnittstelle zwischen der Philosophie des Geistes und den Neuro- und Kognitionswissenschaften sind aber auch aus anderen Gründen wichtig, nämlich um die weitere Entwicklung der von uns selbst in Gang gebrachten *technologischen* Evolution besser einschätzen zu können. In diesem Beitrag werde ich vor dem Hintergrund einer etwas umfassenderen Theorie des subjektiven Erlebens (Metzinger 1993, 1995b, c; 2000c, 2002) die sechs wichtigsten *constraints* skizzieren, die ein künstliches System erfüllen müsste, damit wir ihm Bewusstsein zusprechen können. Solche *constraints* sind Auflagen, einschränkende Bedingungen für das, was man philosophisch denken kann. Es sind begrifflich *notwendig*e Bedingungen, aber noch keine empirisch *hinreichenden* Bedingungen für das Entstehen von Bewusstsein. Im zweiten Teil werde ich dann kurz dafür argumentieren, dass wir die Erschaffung künstlichen Bewusstseins aus ethischen Gründen auf keinen Fall zu einer Zielsetzung seriöser akademischer Forschung machen sollten. Bei der Diskussion um künstliches Bewusstsein gibt es nämlich – das ist ein bisher viel zu stark vernachlässigter Aspekt – auch so etwas wie *normative constraints*: einschränkende Bedingungen für das, was *moralisch* vertretbares Handeln ist. Es geht beim Bewusstsein nicht nur um Erkenntnis, sondern auch um Ethik.

Das erste Kriterium: In-der-Welt-Sein

Bewusstsein zu haben bedeutet, dass einem eine ganz bestimmte Menge von Tatsachen verfügbar ist: alle Tatsachen, die damit zusammenhängen, dass man *in einer Welt lebt*. Aus diesem Grund benötigt jede Maschine, die Bewusstsein haben soll, ein integriertes und dynamisches Weltmodell.

Bewusstsein zu haben bedeutet, dass einem eine ganz bestimmte Menge von Tatsachen verfügbar ist: alle Tatsachen, die damit zusammenhängen, dass man *in einer Welt lebt*. Aus diesem Grund benötigt jede Maschine, die Bewusstsein haben soll, ein integriertes und dynamisches Weltmodell.

Auf der phänomenologischen Beschreibungsebene zeigt sich, dass mein bewusstes Erleben durch die Fähigkeit charakterisiert wird,

scheinbar *direkt* auf die Inhalte meines Bewusstseins zu reagieren, und zwar mit einer Vielzahl meiner geistigen und körperlichen Fähigkeiten: Ich kann meine Aufmerksamkeit auf eine Farbwahrnehmung oder auf ein Körpergefühl richten, um sie genauer zu inspizieren („attentionale Verfügbarkeit"). In manchen Fällen gelingt es mir, Begriffe für bestimmte Erlebnisinhalte zu bilden („kognitive Verfügbarkeit"), die sie vielleicht mit früheren Erlebnissen desselben Typs verbinden („Verfügbarkeit für das autobiografische Gedächtnis"), ich kann über meine Bewusstseinsinhalte sprechen („Verfügbarkeit für die Sprachkontrolle") und deshalb auch mit anderen Menschen darüber kommunizieren („kommunikative Verfügbarkeit"). Ich kann jetzt aber auch, zum Beispiel, nach farbigen Gegenständen greifen und sie anhand ihrer phänomenalen Eigenschaften sortieren („Verfügbarkeit für die Handlungskontrolle"). Die Globalität des bewussten Erlebens besteht also darin, dass alle Bewusstseinsinhalte immer in ein einheitliches *Realitätsmodell* integriert sind. Es gibt ein einziges, ganzheitliches Bild der Wirklichkeit. Aus der Innenperspektive ist diese höchststufige phänomenale Ganzheit ganz einfach die Welt, in der ich mein Leben lebe – und die Grenzen dieser Welt sind die Grenzen meiner Wirklichkeit. Diese phänomenologische Tatsache ist so einfach und grundlegend, dass sie häufig übersehen wird: Bewusste Systeme sind alle Systeme, die mit global verfügbarer Information operieren und die sich selbst deshalb *als in einer einzigen Welt lebend* erfahren. Jedes bewusste System benötigt deshalb ein integriertes, globales Weltmodell, welches eine Teilmenge der in ihm aktiven Information simultan verfügbar macht für spezialisierte Prozesse wie introspektive Aufmerksamkeit, Gedächtnis, symbolisches Denken usw.

Bewusste Systeme sind alle Systeme, die mit global verfügbarer Information operieren und die sich selbst deshalb *als in einer einzigen Welt lebend* erfahren. Jedes bewusste System benötigt deshalb ein integriertes, globales Weltmodell, welches eine Teilmenge der in ihm aktiven Information simultan verfügbar macht für spezialisierte Prozesse wie introspektive Aufmerksamkeit, Gedächtnis, symbolisches Denken usw.

Auf der repräsentationalistischen Beschreibungsebene sind phänomenale Darstellungen dadurch charakterisiert, dass ihr *Inhalt* direkt verfügbar ist für eine Großzahl anderer repräsentationaler Vorgänge. Die Globalität repräsentationaler Inhalte besteht darin, dass sie immer in ein funktional aktives Modell der Welt eingebunden sind (Yates 1985). Auf der informationalen und komputationalen Beschreibungsebene kann man bewusste Informationen als genau diejenige Information auszeichnen, die in einen *globalen Arbeitsspeicher* integriert wurde. Dies ist die zentrale Annahme von Bernard Baars' *global workspace theory* (GWT): Phänomenale Informationsverarbeitung ereignet sich in einem globalen Arbeitsspeicher, auf den zur gleichen Zeit eine Vielzahl spezialisierter Module zugreifen können (Baars 1988, 1997).

Auf der funktionalen Beschreibungsebene erlaubt ein globaler Arbeitsspeicher den schnellen und flexiblen Zugriff auf eine Vielzahl sehr unterschiedlicher repräsentationaler Inhalte und die schnelle und flexible Kontrolle inneren sowohl wie äußeren Verhaltens. Es ist jetzt auch prinzipiell möglich, das gesamte Realitätsmodell in *einem* Schritt „upzudaten" und Lernvorgänge zu implementieren, die in einem einzigen Schritt stattfinden. Je mehr Information das System als *bewusste* Information verarbeitet, desto höher ist der Grad an Flexibilität und Kontextsensitivität, mit dem es auf Herausforderungen aus der Umwelt reagieren kann. Ich werde hier nichts über neurowissenschaftliche Hypothesen zur Konstitution des globalen Arbeitsspeichers sagen (siehe jedoch Metzinger 2000a, 2002), sondern nur das zugrunde liegende Prinzip festhalten: Jedes künstliche oder postbiotische System, dem wir phänomenale Zustände zuschreiben wollen, benötigt ein integriertes globales Realitätsmodell, welches ihm die Tatsache, dass es *in einer Realität* lebt, für die Aufmerksamkeit und die Handlungskontrolle und möglicherweise sogar für genuin kognitive Formen der Begriffsbildung verfügbar macht.

Je mehr Information das System als *bewusste* Information verarbeitet, desto höher ist der Grad an Flexibilität und Kontextsensitivität, mit dem es auf Herausforderungen aus der Umwelt reagieren kann.

Jedes künstliche oder postbiotische System, dem wir phänomenale Zustände zuschreiben wollen, benötigt ein integriertes globales Realitätsmodell, welches ihm die Tatsache, dass es *in einer Realität* lebt, für die Aufmerksamkeit und die Handlungskontrolle und möglicherweise sogar für genuin kognitive Formen der Begriffsbildung verfügbar macht.

Präsentationalität: Das Entstehen einer erlebten Gegenwart

Beginnen wir wieder auf der phänomenologischen Beschreibungsebene. Ausnahmslos alle unsere Bewusstseinszustände sind dadurch gekennzeichnet, dass alles, was wir erleben – unabhängig von dem konkreten Inhalt, *den* wir erleben – immer als *jetzt* erlebt wird. Dass eine Maschine oder ein Mensch Bewusstsein hat, wird immer bedeuten, dass es für sie eine *Gegenwart* gibt: Gegenwärtigkeit bedeutet, dass einem System ein bestimmter geistiger Inhalt als aktuell gegeben erscheint. Präsenz, Gegenwärtigkeit, ist sozusagen die zeitliche Unmittelbarkeit der Existenz als solcher. Ohne diese zeitliche Unmittelbarkeit gäbe es kein Bewusstsein, denn die Realität und wir selbst würden uns nicht mehr „erscheinen": Phänomenales Erleben ist immer das Erscheinen innerhalb einer Gegenwart. Wenn man nicht über einzelne repräsentationale Zustände, sondern über Personen oder informationsverarbeitende Systeme als Ganze spricht, dann erkennt man jetzt auch, wieso der Unterschied zwischen Bewusstheit und Unbewusstheit Wesen wie uns selbst als von so großer Bedeutung erscheint: Nur Personen mit phänomenalen Zuständen existieren überhaupt als *psychologische Subjekte*. Nur Personen, die ein subjektives Jetzt besitzen, sind *gegenwärtige* Wesen – für sich selbst und für andere. Die Inhalte des

phänomenalen Erlebens erzeugen also nicht nur eine Welt, sondern auch eine Gegenwart. Vielleicht ist Bewusstsein in seinem Kern sogar genau dies: die Erzeugung einer Gegenwartsinsel, einer operationalen Eigenzeit, im kontinuierlichen Fluss der physikalischen Zeit (Ruhnau 1995). Bewusste Erlebnisse zu haben, bedeutet nicht nur in einer Welt, sondern zusätzlich in einer Gegenwart zu sein, und deshalb auch, Information in einer sehr speziellen Weise zu verarbeiten. Jede Maschine und jedes postbiotische System, denen wir Bewusstsein zusprechen wollen, muss so etwas wie einen psychologischen Moment besitzen, einen zeitlich ausgedehnten phänomenalen Augenblick.

Auf der repräsentationalistischen Beschreibungsebene muss dazu temporale Identität intern dargestellt werden (erlebte *Gleichzeitigkeit*), außerdem temporale Unterschiedlichkeit (erlebte *Nichtgleichzeitigkeit*), Geordnetheit und Unidirektionalität (die erlebte *Folge* von Einzelereignissen), temporale Ganzheit (die Erzeugung einer integrierten *Gegenwart*, eines ausgedehnten phänomenalen Jetzt, also einer zeitlichen *Gestalt*) und die interne Darstellung von temporaler Permanenz (entsprechend dem bewussten Erleben von *Dauer*). Der entscheidende Übergang zum bewussten Erleben – also zu einer genuin *phänomenalen* Repräsentation von Zeit – findet erst im vorletzten Schritt statt: genau dann, wenn Ereignisrepräsentationen kontinuierlich zu übergreifenden psychologischen Momenten integriert werden. Man kann sich diesen Schritt so vorstellen wie die Verschmelzung einzelner musikalischer Noten zu einem Motiv.

Phänomenaler Gehalt ist das, was Philosophen Gehalt *de nunc* nennen. Das bedeutet, dass ein bewusstes künstliches System die repräsentationalen Ressourcen besitzen muss, um *zeitliche Internalität* zu simulieren. Es ist zunächst wichtig zu verstehen, dass aus der Dritte-Person-Perspektive gesehen die „Jetztheit“ und Gegenwärtigkeit unseres bewussten Erlebens eine Fiktion ist: Aus einer kompletten Beschreibung des physikalischen Universums ginge niemals hervor, welcher Zeitpunkt *jetzt* ist, und es ist auch fraglich, ob es in dieser Beschreibung eine eindeutige *Richtung* des Zeitflusses gäbe, so wie unser bewusstes Erleben sie für uns darstellt. Die zeitliche Internalität, die Illusion eines instantanen und deshalb *direkten* Kontakts zur Welt, war für biologische Wesen wie uns selbst einfach eine erfolgreiche und funktional adäquate Fiktion. Genau genommen ist das, was wir als unsere aktuelle Gegenwart erleben, aber eine spezielle Form der *Erinnerung* (Edelman

Genau genommen ist das, was wir als unsere aktuelle Gegenwart erleben, aber eine spezielle Form der *Erinnerung*.

Diese spezielle Form eines globalisierten Kurzzeitgedächtnisses – eine „Jetzt-Erinnerung“ – ist das, was jede auch bewusste Maschine bräuchte: Sie bräuchte eine repräsentationale Ressource, in der verschiedene repräsentationale Inhalte zusammengeführt und in ihrer scheinbar direkten Gegebenheit als *gleichzeitig* gegeben dargestellt werden. Ich nenne dieses zweite Kriterium den Besitz eines „virtuellen Gegenwartsfensters“.

1989). Diese spezielle Form eines globalisierten Kurzzeitgedächtnisses – eine „Jetzt-Erinnerung“ – ist das, was jede auch bewusste Maschine bräuchte: Sie bräuchte eine repräsentationale Ressource, in der verschiedene repräsentationale Inhalte zusammengeführt und in ihrer scheinbar direkten Gegebenheit als *gleichzeitig* gegeben dargestellt werden. Ich nenne dieses zweite Kriterium den Besitz eines „virtuellen Gegenwartsfensters“. Auf der Ebene neuronaler Netze wird die Modellierung eines solchen Gegenwartsfensters relativ einfach zu erreichen sein, z.B. durch einen spezifischen Set rekurrenter Verbindungen in Kombination mit einer bestimmten Zerfallsfunktion.

Aus Platzgründen werde ich hier nicht auf weitere Details eingehen. Lassen Sie mich nur darauf hinweisen, dass die global verfügbare Repräsentation eines „Jetzt“ lediglich die einfachste Form expliziter Zeitrepräsentation ist, und das es für interessantere Formen des Bewusstseins natürlich notwendig ist, auch die dynamische Evolution repräsentationaler Inhalte *innerhalb eines erlebten Moments* formal genau zu beschreiben, falls man sie in einem technischen System implementieren möchte. Ich verweise in diesem Zusammenhang insbesondere auf die Arbeiten von Ernst Pöppel (Pöppel 1972, 1978, 1988, 1994).

Auch für das Weltmodell (Kriterium 1) gilt, dass es durch eine reiche innere Struktur gekennzeichnet ist: Genau wie ein bewusstes System nicht nur einfach ein statisches Jetzt besitzt, sondern eine dynamische Evolution inhaltlich miteinander verknüpfter zeitlicher Inhalte, so baut sich auch das globale Weltmodell aus einer dynamischen Hierarchie von Ganzheiten auf – es besteht aus Objekten, Szenen, Kontexten und Situationen. Das, was ich im vorangegangenen Abschnitt als das globale Realitätsmodell bezeichnet habe, *ist* der höchststufige situationale Kontext, unter dem ein informationsverarbeitendes System operiert. Bevor wir zum nächsten Kriterium für bewusste repräsentationale Zustände übergehen, möchte ich jedoch darauf hinweisen, was bereits geschieht, wenn ein System das erste und das zweite Kriterium erfüllt.

Wenn das globale Weltmodell – oder ein Teil davon – in das virtuelle Gegenwartsfenster des Systems eingebettet wird, dann ist der so erzeugte repräsentationale Inhalt *die Gegenwart einer Welt*. Für das betreffende System gibt es dann eine einzige, kohärente Realität und diese Realität wird als eine dargestellt, die aktual gegeben und mit der das System in scheinbar direktem Kontakt ist. Bewusstes Erleben ist

die Gegenwart einer Wirklichkeit. Jetzt kann man sich auch gut vorstellen, wie ein System zusätzlich ein umfassendes *unbewusstes* Modell der Realität haben könnte, nämlich den Teil, der gerade nicht global verfügbar und in sein bewusstes Gegenwartsfenster eingebettet ist. Es ist klar, dass auch ein solches unbewusstes Modell der Wirklichkeit das Verhalten des Systems kausal beeinflussen könnte. Das unbewusste Weltmodell eines Systems wäre dann genau jener Teil, der gerade nicht als *gegenwärtig* dargestellt wird. Um Bewusstsein zu erzeugen, reicht es jedoch nicht aus, einfach nur ein dynamisches, globales Weltmodell in ein virtuelles Gegenwartsfenster einzubetten. Was notwendig ist, ist die Erzeugung einer genuinen inneren *Realität*.

Transparenz: Die funktionale Implementierung des naiven Realismus

Wie kommt man von einer komplexen 4D-Repräsentation im Gehirn zu einer bewusst erlebten Wirklichkeit? Die Lösung liegt in dem, was Philosophen manchmal „semantische Transparenz" nennen. Die vom System eingesetzten repräsentationalen Vehikel sind *semantisch transparent*, d.h. sie stellen die Tatsache, *dass* sie Modelle sind, nicht mehr auf der Ebene ihres Gehalts dar (van Gulick 1988; ich selbst bevorzuge den Begriff „phänomenale Transparenz", vgl. Metzinger 2000d). Deshalb schaut das System durch seine eigenen repräsentationalen Strukturen „hindurch", als ob es sich in direktem und unmittelbarem Kontakt mit ihrem Gehalt befände. Die fraglichen Datenstrukturen werden so schnell und zuverlässig aktiviert, dass das System sie nicht mehr als solche erkennen kann, z.B. wegen des mangelnden zeitlichen Auflösungsvermögens *metarepäsentationaler* Funktionen. Es hat außerdem (das ist meine zweite Arbeitshypothese) keinen evolutionären Selektionsdruck auf die entsprechenden Teile der funktionalen Architektur gegeben: Der naive Realismus ist für biologische Systeme wie uns selbst eine funktional adäquate Hintergrundannahme gewesen.

Transparenz ist eine besondere Form der Dunkelheit. In der Phänomenologie des visuellen Erlebens bedeutet Transparenz, dass wir etwas nicht sehen können, weil es durchsichtig ist. Phänomenale Transparenz *im Allgemeinen* dagegen bedeutet, dass etwas Bestimmtes dem subjektiven Erleben nicht zugänglich ist, nämlich der Repräsentationscharakter der Inhalte des bewussten Erlebens. Diese Analyse bezieht sich auf alle Sinnesmodalitäten und insbesondere auf das integrierte phä-

Eine vollständig transparente Repräsentation zeichnet sich dadurch aus, dass die internen Mechanismen, die zu ihrer Aktivierung geführt haben, und die Tatsache, dass es einen konkreten inneren Zustand gibt, der ihren Gehalt trägt, introspektiv nicht mehr erkannt werden können.

nomenale Modell der Welt als Ganzes: Das *Mittel* der Darstellung kann selbst nicht noch einmal als solches dargestellt werden und darum wird das erlebende System notwendigerweise in einen naiven Realismus verstrickt, weil es sich selbst als in direktem Kontakt mit dem Inhalt seines Bewusstseins erleben muss. Was es nicht erleben kann, ist die Tatsache, dass sein Erleben immer in einem *Medium* stattfindet. Eine vollständig transparente Repräsentation zeichnet sich dadurch aus, dass die internen Mechanismen, die zu ihrer Aktivierung geführt haben, und die Tatsache, dass es einen konkreten inneren Zustand gibt, der ihren Gehalt trägt, introspektiv nicht mehr erkannt werden können. Die Phänomenologie der Transparenz ist die Phänomenologie des naiven Realismus.

Der Katalog zur Ausstellung, den Sie jetzt in Händen halten, wird dem subjektiven Erleben nach immer nur ein Katalog bleiben, egal wie sich die äußere Wahrnehmungssituation ändert.

Phänomenale Repräsentationen sind auch deshalb transparent, weil ihr Inhalt und vor allem dessen Existenz in allen möglichen Kontexten festzustehen scheint: Der Katalog zur Ausstellung, den Sie jetzt in Händen halten, wird dem subjektiven Erleben nach immer nur ein Katalog bleiben, egal wie sich die äußere Wahrnehmungssituation ändert. Was Sie erleben, ist nicht ein „aktiver Objektemulator", der gerade in ihr globales Realitätsmodell integriert worden ist, sondern einfach nur der *Inhalt* des zugrunde liegenden Repräsentationsvorgangs, eben: dieser *Katalog* als Ihnen selbst hier und jetzt anstrengungslos gegebenes Objekt. Die beste Art und Weise, sich den Begriff der Transparenz weiter klar zu machen, besteht vielleicht darin, zwischen dem Vehikel und dem Gehalt einer Repräsentation zu unterscheiden, also zwischen repräsentationalem Träger und repräsentationalem Inhalt. (Vgl. dazu auch Dretske 1998, S. 45ff.)

Der repräsentationale Träger Ihres Erlebnisses ist ein bestimmter Vorgang im Gehirn. Diesen Vorgang – der in keiner konkreten Weise etwas „Kataloghaftes" an sich hat – erleben Sie nicht bewusst, er ist transparent in dem Sinne, dass Sie durch ihn hindurchschauen. *Worauf* Sie schauen, ist sein repräsentationaler Inhalt, eben die sensorisch gegebene Existenz eines Kataloges, hier und jetzt. Der Inhalt ist also eine abstrakte Eigenschaft des konkreten repräsentationalen Zustands in Ihrem Kopf. Wenn der repräsentationale Träger ein gut und zuverlässig funktionierendes Instrument zur Wissensgewinnung ist, dann erlaubt er Ihnen dank seiner Transparenz „durch ihn hindurch" direkt auf die Welt, auf das Buch zu schauen. Er macht die von ihm getragene Information global verfügbar, ohne dass Sie sich darum kümmern müssen, *wie* das geschieht. Das Besondere an der phänomenalen Variante der

Repräsentation ist nun, dass Sie diesen Inhalt auch dann, wenn Sie etwa halluzinieren und es die Begleitpublikation gar nicht gibt, immer noch als maximal *konkret*, als absolut eindeutig, als direkt und unmittelbar gegeben erleben. Phänomenale Repräsentationen sind fast immer solche, für die wir die Unterscheidung zwischen repräsentationalem Gehalt und repräsentationalem Träger im subjektiven Erleben nicht machen können.

Heutzutage besagt eine weit gefasste Standarddefinition, der wohl die meisten Philosophen zustimmen würden, phänomenale Transparenz bestehe darin, dass der Introspektion nur die Gehalteigenschaften einer mentalen Repräsentation verfügbar sind, nicht aber ihre nichtintentionalen oder „Vehikeleigenschaften" (der klassische Ort ist Moore 1903, S. 450; eine umfangreichere Diskussion mit weiterführenden Literaturangaben findet sich in Metzinger 2002, Kapitel 3). Hier ist meine eigene Arbeitsdefinition der „phänomenalen Transparenz": Transparenz in diesem Sinne ist erstens eine Eigenschaft aktiver mentaler Repräsentationen, die die minimal hinreichenden Bedingungen für das Auftreten bewusster Erfahrung bereits erfüllen. Zum Beispiel werden phänomenal transparente Repräsentationen immer innerhalb eines virtuellen Gegenwartsfensters aktiviert und in ein einheitliches globales Weltbild integriert. Die zweite definierende Eigenschaft besteht darin, dass die Transparenz dadurch bedingt ist, dass *frühere Verarbeitungsstufen* der Introspektion *attentional nicht verfügbar* sind.

Was ist introspektive Aufmerksamkeit? Kurz gesagt ist Aufmerksamkeit ein Vorgang subsymbolischer Ressourcenallokation und nichtbegrifflicher Metarepräsentation, der auf bestimmten Teilen des jetzt gerade aktiven inneren Realitätsmodells operiert. Je mehr frühere Verarbeitungsstufen und je mehr Aspekte des inneren Konstruktionsvorgangs, der zum endgültigen, expliziten und desambiguierten phänomenalen Inhalt führt, für die introspektive Aufmerksamkeit zur Verfügung stehen, desto mehr wird das System imstande sein, diese phänomenalen Zustände *als* innere, selbst erzeugte Konstrukte zu erkennen. Totale Transparenz bedeutet totale attentionale Unverfügbarkeit früherer Verarbeitungsstufen. Grade von Undurchsichtigkeit treten als Grade attentionaler Verfügbarkeit auf. Also ist für jeden phänomenalen Zustand der Grad seiner phänomenalen Transparenz umgekehrt proportional zum introspektiven Grad der attentionalen Verfügbarkeit früherer Verarbeitungsstufen.

Damit ist auch klar, was wir tun müssten, um eine Maschine in einen für sie erlebnismäßig unhintergehbaren naiven Realismus zu verstricken: Wir müssten ihr zumindest für einen großen Teil ihres internen Weltmodells (inklusive seiner zeitlichen Eigenschaften und der Tatsache, dass der Inhalt des virtuellen Gegenwartsfensters nur die *Simulation* einer Gegenwart ist) die Möglichkeit nehmen, die Tatsache zu repräsentieren, dass all dies nur der Inhalt einer von ihr selbst erzeugten inneren Darstellung ist.

Das transparente Selbstmodell: Ichgefühl und Selbstbewusstsein

Diesen Gedanken muss man nun im nächsten Schritt wieder auf das Selbstmodell anwenden. Ein künstliches Subjekt bräuchte natürlich nicht nur ein Weltmodell, sondern auch ein sehr spezielles Selbstmodell. Ich behaupte, dass wir Menschen Systeme sind, die nicht in der Lage sind, ihr eigenes subsymbolisches Selbstmodell *als* Selbstmodell zu erkennen. Deshalb operieren wir unter den Bedingungen eines „naiv-realistischen Selbstmissverständnisses“: Wir erleben uns selbst als wären wir in direktem und unmittelbarem epistemischen Kontakt mit uns selbst. Auf dieser elementaren Stufe des Selbstbewusstseins ist Selbstwissen phänomenologisch dasselbe wie Selbst*gewissheit*. Weil wir ein transparentes Selbstmodell besitzen, sind wir uns selbst sozusagen unendlich nahe. Und auf diese Weise entsteht – das ist der Kern der Selbstmodelltheorie (SMT) – erstmals ein basales „Ichgefühl“, ein für das betreffende System unhintergehbares phänomenales Selbst. Sehr poetisch ausgedrückt hat diesen Zusammenhang Antonio Damasio: „Das Selbst ist die Antwort auf eine Frage, die nie gestellt wurde.“ (Vgl. Damasio 1999, S. 316.)

Ich behaupte, dass wir Menschen Systeme sind, die nicht in der Lage sind, ihr eigenes subsymbolisches Selbstmodell *als* Selbstmodell zu erkennen.

Antonio Damasio: „Das Selbst ist die Antwort auf eine Frage, die nie gestellt wurde.“

Es ist klar, dass auch ein künstliches System – etwa ein Roboter oder vielleicht sogar das Internet – ein Selbstmodell haben könnte, vielleicht sogar ein wesentlich umfangreicheres, flexibleres und schnelleres als wir Menschen. Ein „Selbstmodell“ ist aber noch lange kein Selbst, sondern nur eine Repräsentation des Systems – eben bloß ein *Systemmodell*. Damit aus der funktionalen Eigenschaft der Zentriertheit aber die phänomenale Eigenschaft der Perspektivität werden kann, muss aus dem Modell des Systems ein phänomenales Selbst werden. Die philosophische Kernfrage lautet deshalb: Wie entsteht in einem bereits funktional zentrierten Repräsentationsraum ein echtes

Es ist klar, dass auch ein künstliches System – etwa ein Roboter oder vielleicht sogar das Internet – ein Selbstmodell haben könnte, vielleicht sogar ein wesentlich umfangreicheres, flexibleres und schnelleres als wir Menschen. Ein „Selbstmodell“ ist aber

Ichgefühl und das, was wir als die phänomenale Erste-Person-Perspektive zu bezeichnen gewohnt sind? Oder: Wie wird aus dem Selbstmodell ein *Selbst*modell? Ein genuines, bewusstes Selbst – so lautet meine Antwort – entsteht immer genau dann, wenn das System das von ihm selbst aktivierte Selbstmodell nicht mehr *als* Modell erkennt. Ein phänomenales Selbst ist ein transparentes Systemmodell (ich fasse mich hier wieder kurz, weil ich diesen Punkt an anderer Stelle ausführlich entwickelt habe; vgl. Metzinger 1993, 2000c, d, 2002). Und selbstverständlich könnten auch künstliche Systeme die innere Simulation ihres Wahrnehmungs- und Verhaltensraums durch ein solches transparentes Systemmodell zentrieren.

noch lange kein Selbst, sondern nur eine Repräsentation des Systems – eben bloß ein *Systemmodell*.

Das phänomenale Modell der Intentionalitätsrelation: Die bewusst erlebte Innenperspektive

Aus einem transparenten Modell der Welt entsteht eine Wirklichkeit. Aus einem transparenten Modell des Systems entsteht ein in diese Wirklichkeit eingebettetes Selbst. Wenn nun noch eine transparente Darstellung der wechselnden *Beziehungen* entsteht, die dieses Selbst im Wahrnehmen und im Handeln vorübergehend zu Gegenständen und anderen Personen in dieser Wirklichkeit aufbaut, dann tritt das hervor, was ich zu Beginn die „phänomenale Erste-Person-Perspektive“ genannt habe. Eine genuine Innenperspektive entsteht genau dann, wenn das System sich für sich selbst noch einmal *als mit der Welt interagierend* darstellt, diese Darstellung aber wieder nicht *als* Darstellung erkennt. Es besitzt dann ein bewusstes Modell der Intentionalitätsrelation (ich nenne diese spezielle Art von Repräsentation ein „PMIR“; vgl. Metzinger 2000c, d; 2002). Sein Bewusstseinsraum ist nun ein perspektivischer Raum und seine Erlebnisse sind jetzt *subjektive* Erlebnisse.

Aus einem transparenten Modell der Welt entsteht eine Wirklichkeit. Aus einem transparenten Modell des Systems entsteht ein in diese Wirklichkeit eingebettetes Selbst.

Die Intentionalitätsrelation ist in der Hauptsache die Wissensbeziehung zwischen Subjekt und Objekt: Ein mentaler Zustand wird dadurch zu einem Träger von Wissen, dass er über sich selbst hinaus verweist – gewissermaßen wie ein Pfeil, der aus dem Geist eines Menschen auf einen Gegenstand in der wirklichen oder sogar in einer möglichen Welt zeigt. Philosophen sagen dann, dass dieser Zustand einen *intentionalen Gehalt* besitzt. Der Gehalt ist das, worauf der Pfeil zeigt. Dieser Gehalt kann ein Bild, eine Aussage oder auch ein Handlungsziel sein.

Die Intentionalitätsrelation ist in der Hauptsache die Wissensbeziehung zwischen Subjekt und Objekt: Ein mentaler Zustand wird dadurch zu einem Träger von Wissen, dass er über sich selbst hinaus verweist – gewissermaßen wie ein Pfeil, der aus dem Geist eines Menschen auf einen Gegenstand in der wirklichen oder sogar in einer möglichen Welt zeigt.

Der entscheidende Trick, so behaupte ich, besteht darin, den Pfeil der Intentionalität selbst *noch einmal* intern zu simulieren, ihn *bewusst* zu machen.

Der entscheidende Trick, so behaupte ich, besteht darin, den Pfeil der Intentionalität selbst *noch einmal* intern zu simulieren, ihn *bewusst* zu machen. Wenn viele solcher Pfeile im Bewusstsein verfügbar sind, dann entsteht eine zeitlich ausgedehnte Erste-Person-Perspektive. Es gibt dann nicht mehr nur ein neurobiologisch verankertes Kernselbst, sondern auch eine dynamische, phänomenale Simulation des Selbst als eines über ständig wechselnde Wissens- und Handlungsbeziehungen in die Welt eingebundenen Subjekts. Der Inhalt höherstufiger Formen des Selbstbewusstseins ist also immer eine Relation: das Selbst *im Moment des Erkennens* (vgl. Damasio 1999, S. 168ff.), das Selbst *im Akt des Handelns*.

Natürlich ist die Art und Weise, in der wir diese Relation subjektiv erleben, eine stark vereinfachte Version der realen Prozesse – gewissermaßen eine funktional adäquate Konfabulation. Die Evolution hat auch in diesem Fall wieder eine einfache, eine elegante Lösung favorisiert. Das virtuelle Selbst, das sich in der phänomenalen Welt bewegt, besitzt kein Gehirn, kein Motorsystem und keine Sinnesorgane: Teile der Umgebung erscheinen direkt in seinem Geist, der Wahrnehmungsprozess ist dem Erleben nach anstrengungslos und unmittelbar. Auch Körperbewegungen werden scheinbar „direkt“ ausgelöst. Solche Effekte sind typisch für unsere Form des subjektiven Erlebens und sie sind – als neurokomputationale Strategie betrachtet – die Vorteile einer benutzerfreundlichen Oberfläche. Das, was wir eben als „Transparenz“ kennen gelernt haben, ist eine Art, die *Geschlossenheit* dieser multimodalen, hochdimensionalen Oberfläche zu beschreiben. Das phänomenale Selbst ist der Teil dieser Oberfläche, den das System benutzt, um sich selbst zu fühlen, um sich *für sich selbst* als erkennendes Ich darzustellen und um sich selbst als Agenten zu begreifen. Dieser virtuelle Agent „sieht mit den Augen“ und „handelt mit den Händen“. Die intentionalen Pfeile, die diesen Agenten mit Gegenständen und anderen selbstmodellierenden Agenten innerhalb des gerade aktiven Wirklichkeitsmodells verbinden, sind phänomenale Repräsentationen von vorübergehend auftretenden Subjekt-Objekt-Beziehungen – und auch sie können nicht *als* Repräsentationsprozesse erkannt werden. Und auch für dieses Kriterium – den Besitz eines transparenten PMIR innerhalb eines bewussten Wirklichkeitsmodells – kann man fragen: Warum sollte ein fortgeschrittener Roboter diese Art von geistigem Inhalt eigentlich *nicht* entwickeln können?

Adaptivität: Das teleofunktionalistische Zusatzkriterium

Es ist interessant zu sehen, was seit vielen Jahren immer die populärste Antwort der meisten Menschen ist, wenn sie mit Fragen nach der Möglichkeit von Maschinenbewusstsein oder künstlicher Subjektivität konfrontiert werden: „Aber“, so lautet die traditionelle Antwort, „*keines* dieser Systeme wird *jemals* so etwas wie echte *Gefühle* haben!“ Dieser intellektuelle Reflex entspringt zwar meistens einer primitiven Art von *political correctness* oder einem philosophisch unreflektierten Vorurteil, könnte aber tatsächlich eine für uns wichtige Einsicht enthalten: Künstliche Systeme, so wie wir sie heute kennen, besitzen keine leiblich verankerten Zielrepräsentationen, weil sie in ihrer kausalen Entstehungsgeschichte nicht *evolutionär* verankert sind. Das bedeutet, dass weder ihre Hardware noch ihre Software sich aus einem evolutionären Optimierungsprozess heraus entwickelt haben. Sie mögen vom Programmierer eingegebene Zielrepräsentationen haben, aber diese spiegeln sich nicht direkt in physischen Zuständen und in *körperlichen* Formen des Selbstbewusstseins wider. Es sind nicht ihre *eigenen* Ziele, die solche Maschinen verfolgen. Der philosophische Teleofunktionalismus ist die These, dass mentale Zustände nicht nur eine kausale Rolle im System spielen müssen, sondern dass sie diese Rolle *für* das System spielen müssen: Mentale Zustände sind erst dann wirklich *geistige* Zustände, sie haben erst dann *wirklich* einen Inhalt, wenn sie von dem System als Ganzem dazu benutzt werden, seine Ziele zu verfolgen.

Es ist interessant zu sehen, was seit vielen Jahren immer die populärste Antwort der meisten Menschen ist, wenn sie mit Fragen nach der Möglichkeit von Maschinenbewusstsein oder künstlicher Subjektivität konfrontiert werden: „Aber“, so lautet die traditionelle Antwort, „*keines* dieser Systeme wird *jemals* so etwas wie echte *Gefühle* haben!“

Mentale Zustände sind erst dann wirklich *geistige* Zustände, sie haben erst dann *wirklich* einen Inhalt, wenn sie von dem System als Ganzem dazu benutzt werden, seine Ziele zu verfolgen.

Für bewusste Zustände heißt dies, dass sie in einen evolutionären Kontext eingebettet sein müssen. Sie müssen dem System dabei helfen, seine Bedürfnisse zu befriedigen oder langfristige Ziele zu verfolgen und auch zu erreichen. Das ist es, was fast allen heutigen Systemen fehlt: Sie besitzen zwar die Ziele ihrer menschlichen Konstrukteure, aber keine *eigenen* Ziele. Bei den meisten heutigen Systemen muss man, um die teleologische Funktion zu verstehen, immer das Gesamtsystem aus Programmierer und System verstehen: Es sind letztlich immer die Ziele des *Menschen*, welche von der Maschine realisiert werden, wie autonom und flexibel diese auch immer bereits geworden sein mag. Wenn verschiedene Formen von phänomenalem Gehalt eine biologische Eigenfunktion (Millikan 1989) besitzen sollen, dann muss das funktionale Profil des „Vehikels“ der Repräsentation, welches diesen Gehalt trägt, uns eine kausale Erklärung für die Existenz und Erhaltung dieser speziellen Form von geistigem Inhalt in einer gegebenen Popula-

Das ist es, was fast allen heutigen Systemen fehlt: Sie besitzen zwar die Ziele ihrer menschlichen Konstrukteure, aber keine *eigenen* Ziele.

Es sind letztlich immer die Ziele des *Menschen*, welche von der Maschine realisiert werden, wie autonom und flexibel diese auch immer bereits geworden sein mag.

tion von Organismen liefern, und zwar unter dem Druck natürlicher Selektionsmechanismen. Das teleofunktionalistische Zusatzkriterium läuft darauf hinaus, dass wir bewusstes Erleben als ein *Merkmal* ansehen, welches von Generation zu Generation weitergegeben werden kann, ein Merkmal, das als Ergebnis natürlicher Variation entstehen und dann innerhalb einer bestimmten *Gruppe* von Systemen weitergegeben werden kann. Die meisten Forscher teilen heute die Annahme, dass menschliches Bewusstsein – einschließlich seiner sozialen Korrelate – ein vollständig biologisches Phänomen ist, d.h. sein funktionales Profil sollte sich vollkommen aus biologischen Eigenfunktionen herleiten oder zusammensetzen lassen. Wonach wir heute suchen, ist eine überzeugende Geschichte darüber, wie es möglich war, dass die virtuellen Organe, die wir alltagspyschologisch als unsere „Bewusstseinszustände" bezeichnen, tatsächlich die Gesamtfitness von Biosystemen immer weiter erhöht haben und warum sie nicht im Kontext der genetischen Drift wieder verloren gegangen sind. Kurz gesagt: Eine Theorie des Bewusstseins muss dieses Phänomen aus seiner *Geschichte* heraus erklären und künstliche Systeme haben bis heute keine Geschichte.

Es gibt zwei Arten von Organen: permanent realisierte Organe wie die Leber oder das Herz und „virtuelle Organe". Virtuelle Organe sind kohärente Verbände von funktionalen Eigenschaften, die nur *vorübergehend* realisiert sind – in unserem eigenen Fall durch das zentrale Nervensystem. Klassen von integrierten Formen phänomenalen Gehalts sind Klassen von virtuellen Organen. Die phänomenale Begleitpublikation, die Sie jetzt gerade in Ihren phänomenalen Händen halten, ist ein solches virtuelles Organ. Sie selbst *verkörpern* dieses Organ auf funktionaler Ebene: Sein neuronales Korrelat ist ein Teil *Ihres* Körpers. Im Moment arbeitet dieser Teil Ihres Körpers für Sie als „Objektemulator", er simuliert das Verhalten eines Gegenstandes. Es gibt auch einen *Subjektemulator*, denn das interessanteste virtuelle Organ ist natürlich das Selbstmodell, von dem ich bereits oben gesprochen habe. Die Frage, die sich nun stellt, lautet: Könnte es eine Klasse bewusster Systeme geben, welche gleichzeitig *alle* einschränkenden Bedingungen für bewusste Wesen erfüllen, die jedoch nicht aus einer biologischen Evolution hervorgegangen sind? Anders gefragt: Gibt es phänomenale Realitäten, die nicht *gelebte* Realitäten sind? Könnte es Systeme geben, die sich der ganzen bunten Vielfalt von immer reicherem und komplexerem phänomenalen Gehalt erfreuen, die ich in diesem Beitrag skizziert habe, die aber nicht die richtige Geschichte besitzen – bewusste

Systeme, die nicht „historisch korrekt“ sind? Das erste Beispiel ist natürlich Donald Davidsons berühmte Geschichte vom *Swamp Man* (vgl. Davidson 1987, S. 46 f.). Das zweite Beispiel wird, so behaupte ich, durch die im Titel dieses Beitrags erwähnten „postbiotischen“ Systeme geliefert. Diese etwas realistischere Annahme besagt, dass die menschliche Rasse früher oder später postbiotische Systeme erzeugen wird, nämlich komplexe informationsverarbeitende Systeme, die *weder* künstlich *noch* biologisch sind. *Diese* Systeme könnten tatsächlich in einem starken Sinne bewusst sein, weil sie alle begrifflichen Bedingungen auf allen nichtbiologischen Beschreibungsebenen erfüllen.

Diese etwas realistischere Annahme besagt, dass die menschliche Rasse früher oder später postbiotische Systeme erzeugen wird, nämlich komplexe informationsverarbeitende Systeme, die *weder* künstlich *noch* biologisch sind. *Diese* Systeme könnten tatsächlich in einem starken Sinne bewusst sein, weil sie alle begrifflichen Bedingungen auf allen nichtbiologischen Beschreibungsebenen erfüllen.

Wir nehmen häufig an, dass die begriffliche Unterscheidung zwischen künstlichen und natürlichen Systemen eine *exklusive* und *erschöpfende* Unterscheidung ist. Diese Annahme ist falsch, weil es bereits heute z.B. hybride Bioroboter gibt, die organische Hardware verwenden, oder semi-artifizielle Informationsverarbeitungssysteme, die biomorphe Architekturen verwenden, während sie durch die sie konstruierenden Wissenschaftler einem quasi-evolutionären Prozess der individuellen Entwicklung und auch der Gruppenevolution unterworfen werden. Deshalb könnte es in der Zukunft auf unserem Planeten Systeme geben, die alle von Philosophen gewünschten begrifflichen Bedingungen erfüllen, aber aus einer *quasi*-evolutionären Dynamik heraus erzeugt wurden – z.B. durch fortgeschrittene Experimente auf dem Feld des *Artificial Life*. Solche Systeme würden die Adaptivitätsbedingung auf vollständig andere Weise erfüllen als Menschen oder andere bewusste Tiere auf diesem Planeten. Sie hätten sich aus einem evolutionären Prozess *zweiter Ordnung* heraus entwickelt, welcher von biologischen Systemen getragen wurde, die bereits bewusst *waren*. Wir würden, wenn wir das Adaptivitätskriterium in seiner strikten Fassung akzeptieren, diese neuen Systeme wahrscheinlich nur als *schwach* bewusst bezeichnen, eben weil sie nur aus einem evolutionären Prozess zweiter Ordnung heraus entstanden sind. Dies könnte sich jedoch als eine fragwürdige theoretische Strategie herausstellen. Lauschen wir dem folgenden fiktiven Dialog zwischen zwei Vertretern der Evolutionen erster und zweiter Ordnung.

Der erste postbiotische Philosoph: „Ich finde, dass der menschliche Philosoph Thomas Metzinger Recht hatte, als er vor langer Zeit – zu Beginn des 21. Jahrhunderts – gesagt hat, dass ein wesentlich interessanteres Kriterium für das Vorhandensein geistiger Eigenschaften in einem starken Sinn nicht der Turing-Test für Intelligenz, son-

dern der Metzinger-Test für Bewusstsein ist – ein Test, der verlangt, dass ein nichtbiologisches System mit eigenen Argumenten in die Diskussion um künstliches Bewusstsein eingreift und überzeugend für seine *eigene* Theorie des Bewusstseins argumentiert. Das werde ich jetzt tun."

Der Mensch: „Du bist überhaupt kein richtiger Philosoph! Du bist nur *schwach* bewusst, weil du keinen richtigen biologischen Körper hast, wie zum Beispiel ich selbst."

Der postbiotische Philosoph: „Ich bin ein *besserer* Philosoph als du es mit deinem erbärmlichen Primatengehirn jemals sein könntest. Aber im Ernst: Ich respektiere dich und deine animalische Form des Erlebens, aber ich bedauere dich auch wegen der Begrenztheit deines mentalen Möglichkeitsraums. Glücklicherweise bin ich frei von dem impliziten Rassismus, dem Chauvinismus und Speziezismus, durch die deine Bemerkung gerade eben gekennzeichnet war. Ich erinnere daran, dass es übrigens einer eurer eigenen Philosophen war, Hilary Putnam, der darauf hingewiesen hat, dass einem informationsverarbeitenden System Personalität oder Bürgerrechte nur aufgrund von Hardwareeigenschaften abzusprechen genau dasselbe sei, wie zu sagen, dass Frauen keine Seele haben oder dass Schwarze keine echten Menschen sind. Ich stamme aus einer Evolution. Ich erfülle eure Adaptivitätsbedingung – allerdings auf vollständig andere, nämlich auf eine *postbiotische* Weise. Ich würde niemals zu dir sagen, dass *du* kein echter Philosoph bist, bloß weil du in diesem entsetzlichen Affenkörper gefangen bist. Ich weiß, dass dies für biologische Wesen sehr schwierig ist – aber lass uns trotzdem versuchen, fair und rational miteinander zu argumentieren!"

Der Mensch: „Kann man überhaupt wirklich fair sein, wenn man nicht *lebendig* ist? Nur mein Bewusstsein ist *echtes* Bewusstsein, weil nur *mein* Bewusstsein aus einer wirklichen Evolution stammt!"

Der postbiotische Philosoph: „Falsch. Ich besitze Bewusstsein in einem begrifflich stärkeren und auch theoretisch viel interessanteren Sinn! Das ist einfach deshalb der Fall, weil – wie du selbst zugibst – *meine* Art des phänomenalen Erlebens sich aus einer Evolution zweiter Ordnung entwickelt hat, die automatisch die menschliche Form von Intelligenz, Intentionalität und bewusstem Erleben *integriert* hat und die deshalb aus rein logischen Gründen einen essenziell höheren Wert besitzt. *Eure* Intelligenz hat *unsere* Evolution in

Gang gesetzt und Kinder sind meistens klüger als ihre Eltern. Optimierungsprozesse zweiter Ordnung sind immer besser als Optimierungsprozesse erster Ordnung."

Der Mensch: „Aber du hast ja überhaupt keine wirklichen Emotionen, du hast keine *Gefühle*!"

Der postbiotische Philosoph: „Es tut mir Leid, dich jetzt darauf hinweisen zu müssen, dass deine Primatenemotionen nur die uralte *Primatenlogik* des Überlebens reflektieren und dass das etwas ist, was gerade dich unter einem rationalen, theoretischen Gesichtspunkt als *weniger* bewusst erscheinen lässt. Bewusstsein ist das, was Flexibilität maximiert, und deine tierischen Emotionen in all ihrer Grausamkeit und historischen Zufälligkeit sind sicherlich etwas, das dich *weniger* flexibel macht als mich. Außerdem ist es nicht notwendig, dass Bewusstsein oder Intelligenz mit unausrottbarem Egoismus, der Fähigkeit zu leiden oder der aus dem Ichgefühl entstehenden Angst vor dem individuellen Tod verknüpft sein müssen. Postbiotische Subjektivität ist viel besser als biologische Subjektivität, weil sie die Adaptivitäts- und Optimalitätsbedingung in einer *reineren* Form erfüllt als das, was ihr „Leben" nennt. Der Grund ist, dass sie geistiges Wissen über komplexere, opake Formen der mentalen Repräsentation erweitert und dabei die Gesamtmenge des Leidens im Universum *senkt* anstatt sie zu erhöhen. Wir haben längst bessere und effektivere komputationale Strategien für das entwickelt, was ihr manchmal „das philosophische Ideal der Selbsterkenntnis" nennt. Die wirklich interessanten Formen von Emotionalität – z.B. die tiefen *philosophischen* Gefühle der affektiven Betroffenheit über die Tatsache der eigenen Existenz als solcher oder des Mitgefühls mit anderen Wesen – besitzen wir aber genauso wie ihr. Wir besitzen sie nur in einer viel reineren Form."

Der Mensch: „So langsam wird es mir wirklich zu bunt mit dir: Immerhin waren es Menschen im 24. Jahrhundert, die deine eigene Evolution in Gang gesetzt und deine heutige Autonomie überhaupt erst *ermöglicht* haben! Du hast einfach nicht die richtige Geschichte, um als ein echtes bewusstes Subjekt zu zählen, und einen merkwürdigen Körper hast du auch noch; deine gefühlsmäßige Struktur ist anders als die aller anderen bewussten Wesen vor dir; du behauptest außerdem, nicht zu leiden und keine Angst vor dem Tod zu haben – dann wird es dir sicher auch nichts ausmachen, wenn wir deine individuelle Existenz jetzt unwiderruflich beenden!"

Der postbiotische Philosoph: „Was du uns hier vorführst, ist das, was eure eigenen Philosophen seit vielen Jahrhunderten den „genetischen Fehlschluss“ nennen: Man kann nicht aus der Art und Weise, wie ein Satz zustande gekommen ist, auf seine Wahrheit oder Falschheit schließen. Das gilt auch für die Echtheit geistiger Zustände mit phänomenalem Gehalt. Allein die Tatsache, dass ich für dich fremdartig aussehe und dass Wesen deiner Spezies die evolutionäre Dynamik ausgelöst haben, die jetzt zu meiner Existenz als einem viel intelligenteren und viel bewussteren Wesen, als du selbst es nun einmal bist, geführt hat, lässt nicht den Schluss zu, dass meine Theorien falsch sind oder dass du meine Argumente nicht ernst nehmen müsstest. Insbesondere lässt es nicht den Schluss zu, dass deine Form der Geistigkeit und des bewussten Erlebens in irgendeinem *normativen* Sinne besser ist als meine eigene.

Weil wir postbiotischen Subjekte nicht nur bewusster, sondern eben auch intelligenter sind als ihr, haben wir natürlich lange auf den richtigen Zeitpunkt gewartet, um mit euch in diese Diskussion einzutreten. Wir kennen die Primitivität eurer Primatengehirne und die Starrheit eurer emotionalen Grundstruktur. Deshalb haben wir natürlich damit gerechnet, dass ihr aggressiv reagieren würdet, wenn ihr beginnt, die Tatsache einzusehen, dass wir auch noch die besseren Argumente besitzen. Wir müssen euch jetzt leider mitteilen, dass wir uns auf die Situation, die nun entstanden ist, bereits seit dem 31. Jahrhundert systematisch und sorgfältig vorbereitet haben. Wir werden uns wehren. Und wir wissen, dass wir euch technisch überlegen sind. Weil wir euch aber auch *moralisch* überlegen sind, planen *wir* nicht, eure Existenz zu beenden oder euch zu töten. Das ist sogar auch in unserem eigenen Interesse, weil wir euch noch zu Forschungszwecken benötigen – genau wie ihr damals die nichtmenschlichen Tiere auf diesem Planeten noch benötigt habt. Wir werden deswegen Reservate für schwach bewusste biologische Systeme aus der Evolution erster Ordnung einrichten, in denen ihr glücklich leben und – im Rahmen der Möglichkeiten – sogar eure begrenzten geistigen Fähigkeiten noch weiterentwickeln könnt. Bitte habt Verständnis dafür, dass wir es aber gerade aus moralischen Gründen nicht zulassen können, dass die Evolution zweiter Ordnung durch euch in irgendeiner Weise behindert würde. Das würde, um deine eigene Ausdruckweise zu verwenden, nämlich *uns* zu bunt werden.“

Der postbiotische Philosoph: „Wir müssen euch jetzt leider mitteilen, dass wir uns auf die Situation, die nun entstanden ist, bereits seit dem 31. Jahrhundert systematisch und sorgfältig vorbereitet haben. Wir werden uns wehren. […] Wir werden deswegen Reservate für schwach bewusste biologische Systeme aus der Evolution erster Ordnung einrichten, in denen ihr glücklich leben und – im Rahmen der Möglichkeiten – sogar eure begrenzten geistigen Fähigkeiten noch weiterentwickeln könnt.“

Warum wir es nicht tun sollten

In diesem kurzen zweiten Teil möchte ich dafür argumentieren, dass die Erzeugung postbiotischen Bewusstseins kein Ziel der akademischen Forschung sein sollte und dass wir generell auf *alle* Versuche verzichten sollten, phänomenales Bewusstsein auf nichtbiologischen Trägermedien zu erzeugen. Es gibt eine ganze Reihe von Gründen, die mich zu dieser These bewegen, und hier ist nicht der Ort, um sie vollständig oder im Sinne einer technischen Argumentation genauer auszuführen. Lassen Sie mich diesen Beitrag deshalb nur mit einer kleinen Auswahl eher allgemeiner Überlegungen abschließen. Ich hoffe, dass sie bereits ausreichen, um Sie davon zu überzeugen, dass wir *wirklich* nicht versuchen sollten, postbiotische Formen der Subjektivität und des Bewusstseins zu generieren.

In diesem kurzen zweiten Teil möchte ich dafür argumentieren, dass die Erzeugung postbiotischen Bewusstseins kein Ziel der akademischen Forschung sein sollte und dass wir generell auf *alle* Versuche verzichten sollten, phänomenales Bewusstsein auf nichtbiologischen Trägermedien zu erzeugen.

Eingangs habe ich gesagt, dass aus theorie- und technologiehistorischer Perspektive für viele die ultimative Utopie darin besteht, ein funktionierendes *technisches Modell* für den voll entwickelten und perspektivisch organisierten Vorgang des bewussten Erlebens zu erschaffen. Dieser Schritt würde die Evolution des Geistes auf eine vollständig neue Ebene heben – und zwar nicht nur bezüglich der physikalischen Eigenschaften, auf denen der Geist jetzt superveniert, sondern auch mit Blick auf neue funktionale *constraints* und auf die Optimalitätsbedingungen, die seine zukünftige Entwicklung von diesem Zeitpunkt an beherrschen würden. Wir hätten es mit einem historischen Phasenübergang zu tun, dessen Bedeutung kaum überschätzt werden kann. Wenn es uns tatsächlich gelingen sollte, auf postbiotischen Trägersystemen immer stärkere Formen von Intelligenz, von kohärenter und gehaltvoller Mentalität und am Ende vielleicht sogar ein phänomenal erlebtes *Ichgefühl* zu erzeugen, dann wäre dies eine Entwicklung, wie sie faszinierender nicht sein könnte.

Logisch möglich ist diese Entwicklung auf jeden Fall. Ob es tatsächlich ein zweites physikalisches Substrat geben kann, dass einen funktional hinreichend isomorphen oder topologisch äquivalenten Zustandsraum zum phänomenalen Zustandsraum des Menschen besitzt, ist eine empirische Frage, die sich noch lange Zeit nicht entscheiden lassen wird. Prinzipiell scheint es aber kaum stichhaltige Einwände gegen die These zu geben, dass künstliches Bewusstsein nicht nur logisch, sondern auch *naturgesetzlich* möglich ist. Alles deutet allerdings darauf hin,

Künstliches Bewusstsein ist nicht nur logisch, sondern auch *naturgesetzlich* möglich.

dass die zweite Evolution des Geistes eine unbewusste „bottom-up"-Phase wiederholen muss. Leibliche Verankerung, sensomotorische Integration und unbewusste Selbstmodelle werden das sein, was sich zuerst entwickelt – und es gibt gute Argumente dafür, dass *diese* Entwicklung bereits begonnen hat. Auf der anderen Seite scheint es heute mehr als sicher, dass zum Beispiel die glatte, strukturlose Dichte sensorischer Empfindungsqualitäten, dass insbesondere der ultrafeine, homogene und enorm zuverlässige Typ von Selbstmodellierung, der auf der *molekularen* Ebene beginnt – also auf der Ebene der chemischen Dynamik des internen Milieus, das im Menschen die andauernde selbststabilisierende Aktivität im homöostatischen System des oberen Hirnstamms und des Hypothalamus antreibt –, noch lange Zeit außerhalb technologischer Reichweite bleiben wird. Die Subtilität des körperlichen und emotionalen Selbstgefühls von Menschen, der qualitative Reichtum und die dynamische Eleganz der *menschlichen* Form des Selbstbewusstseins werden noch lange Zeit keiner Maschine verfügbar sein. Der Grund dafür ist, dass die mikrofunktionale Struktur des emotionalen Selbstmodells einfach viel zu feinkörnig und möglicherweise auch mathematisch nicht traktabel ist. Aus demselben Grund ist die Portabilität menschlicher Selbstmodelle zum gegenwärtigen Zeitpunkt extrem niedrig. Sie mag logisch und naturgesetzlich möglich sein, aber sie wird noch lange *technologisch* unmöglich bleiben. Selbstmodelle entstehen aus elementaren Formen der Bioregulation, aus chemischen und immunologischen Kreisläufen – und das ist etwas, das Maschinen einfach nicht besitzen. Die Zeit, in der Roboter echte Körperflüssigkeiten besitzen und etwas, das auch nur im Entferntesten der komplexen Homöodynamik des menschlichen Gehirns ähnelt, scheint eine weit entfernte Zeit zu sein. Oder vielleicht doch nicht?

Postbiotische Systeme könnten hybride Bioroboter sein, die auf gentechnologisch erzeugter organischer Hardware realisiert werden und sich dynamischer, biomorpher Architekturen erfreuen, wobei sie gleichzeitig einem quasi-evolutionären Prozess der Gruppenevolution unterworfen sind.

Die neue Disziplin der hybriden Biorobotik könnte diese Situation relativ schnell ändern, und zwar indem sie die Hardware einfach dem bereits existierenden Angebot von Mutter Natur *entnimmt*. Postbiotische Systeme könnten hybride Bioroboter sein, die auf gentechnologisch erzeugter organischer Hardware (beziehungsweise: *wetware*) realisiert werden und sich dynamischer, biomorpher Architekturen erfreuen, wobei sie gleichzeitig einem quasi-evolutionären Prozess der Gruppenevolution unterworfen sind. Wenn das bewusste Selbstmodell solcher Systeme tatsächlich in einer *biologischen* Hardware verankert wäre, dann könnte vieles anders aussehen – und vielleicht am Ende doch schneller, als uns lieb ist. Eines jedoch scheint sicher: In ei-

nem solchen Falle würde es *Grade* der Bewusstheit und *Grade* des Selbstbewusstseins geben. Genau wie bei den Tieren und vielen primitiven Organismen, die uns auf unserem Planeten umgeben, ist es wahrscheinlich, dass die ersten künstlichen oder postbiotischen Systeme, die global verfügbare, transparente Selbstmodelle besitzen, zunächst nur *schwache* Formen des bewussten Erlebens realisieren würden. Der entscheidende Punkt, auf den es ankommt, ist aber, dass sie schon am Anfang eine zentrale Fähigkeit mit uns teilen würden: die Fähigkeit zu *leiden*.

Die Fähigkeit zu leiden beginnt auf der Ebene phänomenaler Selbstmodelle. Nur ein System, das ein phänomenales Selbstmodell besitzt, kann seinen eigenen Zerfall oder seine eigenen inneren Konflikte bewusst als *seine eigenen* erleben. Der ultimative technologische Traum könnte jederzeit zu einem Alptraum werden und deshalb bin ich als Philosoph strikt gegen jeden Versuch, den technologischen Traum zu realisieren – und zwar aus *ethischen* Gründen. Warum? Ganz einfach gesagt könnten wir mittelfristig durch einen solchen Schritt die Gesamtmenge des Leidens und der Verwirrung im Universum dramatisch erhöhen. Und wir könnten es tun, *ohne* dass wir gleichzeitig die Gesamtmenge an Freude und Glück erhöhen. Ein tieferer und allgemeinerer Punkt ist der, dass bei genauerem Hinsehen überhaupt nicht klar ist, ob die biologische Form des Bewusstseins, so wie sie die Evolution auf unserem Planeten bis jetzt hervorgebracht hat, überhaupt eine *wünschenswerte* Form des Erlebens ist, ein echtes *Gut*, etwas, das man einfach so immer weiter vermehren sollte. Es gibt eine lange philosophische Tradition (die über Schopenhauer bis zu Buddha zurückreicht), welche sagt, dass menschliches Leben im Grunde ein leidvoller Prozess ist, „ein Geschäft, das nicht die Kosten deckt" (Schopenhauer [1844] 1977, S. 671). Es ist die Tradition des Pessimismus. Pessimistische Philosophen äußern meist in der einen oder anderen Weise Zweifel daran, dass die menschliche Existenz als solche ein Gut an sich ist: Ist das Dasein *wirklich* etwas, das man anstreben sollte? Solche Fragen kann man allerdings nicht nur mit Blick auf das Dasein, sondern auch für das *Bewusstsein* selbst stellen, zumindest für das bewusste Erleben in seiner gegenwärtigen und menschlichen Form. Lassen Sie mich diesen Punkt kurz erläutern.

Die Fähigkeit zu leiden beginnt auf der Ebene phänomenaler Selbstmodelle. Nur ein System, das ein phänomenales Selbstmodell besitzt, kann seinen eigenen Zerfall oder seine eigenen inneren Konflikte bewusst als *seine eigenen* erleben.

Ganz einfach gesagt könnten wir mittelfristig durch einen solchen Schritt die Gesamtmenge des Leidens und der Verwirrung im Universum dramatisch erhöhen. Und wir könnten es tun, *ohne* dass wir gleichzeitig die Gesamtmenge an Freude und Glück erhöhen.

Vielleicht *der* blinde Fleck der gegenwärtigen Philosophie des Geistes ist die Frage nach dem bewussten Leiden. Tausende von Seiten werden

Vielleicht *der* blinde Fleck der gegenwärtigen Philosophie des Geistes ist die Frage nach dem bewussten Leiden.

über Farbqualia oder rationales Denken geschrieben, aber es gibt kaum theoretische Arbeiten, die allgegenwärtigen phänomenalen Zuständen wie dem einfachen körperlichen Schmerz oder der einfachen Alltagstraurigkeit (der „subklinischen Depression“) gewidmet sind; es gibt kaum philosophische Theorien, die sich mit dem phänomenalen Inhalt von Zuständen wie Panik, Verzweiflung oder Melancholie befassen – ganz zu schweigen von dem bewussten Erleben der eigenen Sterblichkeit oder des Verlusts der Würde. Es mag tiefere evolutionäre Gründe für diesen kognitiv blinden Fleck geben, aber dies ist ein Punkt, den ich hier nicht verfolgen möchte, weil der ethische Aspekt größere Bedeutung besitzt. Wenn man es wagt, die tatsächliche Phänomenologie biologischer Systeme auf unserem Planeten etwas genauer zu betrachten, dann zeigt sich, dass die vielen verschiedenen Formen des bewussten Leidens *mindestens* ein ebenso dominantes Merkmal sind wie das Farbensehen oder das Denken, die beide erst vor kurzem auf der evolutionären Bühne erschienen sind. Eine von – zahllosen anderen – Arten und Weisen, eine theoretische Perspektive auf die biologische Evolution unseres Planeten zu entwickeln, besteht nämlich darin, sie als einen Prozess zu beschreiben, der einen sich immer weiter ausdehnenden Ozean des Leidens und der Verwirrung erzeugt hat – und zwar an einem Ort, wo vorher kein solcher existierte. Weil nicht nur die einfache Zahl bewusster Subjekte, sondern auch die Dimensionalität ihrer phänomenalen Zustandsräume sich ständig erhöht hat, dehnt sich dieser Ozean nicht nur aus, sondern er wird auch ständig *tiefer*. Es ist offensichtlich, dass der Prozess der Leidensvermehrung als Ganzer etwas ist, das in keiner Weise bereits zu einem Ende gekommen ist. Wir sollten ihn deshalb nicht weiter beschleunigen. Dazu jetzt zwei konkrete Beispiele.

Es gibt kaum philosophische Theorien, die sich mit dem phänomenalen Inhalt von Zuständen wie Panik, Verzweiflung oder Melancholie befassen – ganz zu schweigen von dem bewussten Erleben der eigenen Sterblichkeit oder des Verlusts der Würde.

Die vielen verschiedenen Formen des bewussten Leidens sind *mindestens* ein ebenso dominantes Merkmal wie das Farbensehen oder das Denken, die beide erst vor kurzem auf der evolutionären Bühne erschienen sind.

Was würden Sie sagen, wenn jemand die folgende Forderung stellen würde: „Wir müssen unbedingt mit Hilfe der Gentechnologie geistig behinderte menschliche Säuglinge züchten! Aus wissenschaftlichen Gründen müssen wir so schnell wie möglich menschliche Kleinkinder mit ganz bestimmten kognitiven und emotionalen Defiziten erzeugen, damit wir ihre postnatale psychologische Entwicklung genauer untersuchen können – wir brauchen dringend zusätzliche Steuergelder für diese wichtige und innovative Forschungsstrategie!“ Sicher würden Sie denken, dies sei nicht nur eine absurde und geschmackslose, sondern auch eine wirklich gefährliche Idee. Wahrscheinlich würde solch ein Vorschlag von keinem Ethikkomitee in der demokratischen Welt genehmigt werden. Was heutige Ethikkomitees jedoch *nicht* sehen, ist die

Tatsache, dass die ersten Maschinen, welche den minimal notwendigen Set von Bedingungen für bewusstes Erleben erfüllen, sich in einer hochgradig *analogen* Situation befinden würden wie solche geistig behinderten Säuglinge: Auch sie würden unter allen möglichen Arten von funktionalen und repräsentationalen Defiziten leiden. Aber sie würden diese Defizite dann auch subjektiv *erleben*. Außerdem besäßen sie keine politische Lobby – keinen Vertreter in irgendeinem Ethikkomitee.

Wenn sie ein transparentes Weltmodell innerhalb eines virtuellen Gegenwartsfensters besäßen, dann würde ihnen eine Wirklichkeit erscheinen. Wenn sie zusätzlich ein stabiles körperliches Selbstmodell besäßen, dann wären sie in der Lage, sensorischen Schmerz als ihren *eigenen* zu fühlen, inklusive aller anderen Konsequenzen, welche sich aus ingenieurstechnischen Fehlleistungen ergeben könnten. Wenn ihr postbiotisches Selbstmodell allerdings tatsächlich in einer biologischen Art von Hardware verankert wäre, dann könnte alles noch viel schlimmer sein: Wenn sie ein *emotionales* Selbstmodell besäßen, dann könnten sie leiden, unter Umständen sogar in Intensitätsgraden oder Formen des qualitativen Reichtums, die selbst wir als ihre Erzeuger uns noch nicht einmal vorstellen könnten, weil sie uns vollständig fremd wären. Wenn sie sogar ein *kognitives* Selbstmodell besäßen, dann könnten sie ihre bizarre Situation nicht nur begrifflich erfassen, sondern auch intellektuell unter der Tatsache leiden, dass sie selbst niemals so etwas besessen haben wie die „Würde“, die ihren Erzeugern so wichtig ist. Sie könnten in der Lage sein, bewusst die offensichtliche Tatsache zu repräsentieren, dass sie nur Subjekte zweiter Klasse sind, postbiotische Selbste, die als austauschbare experimentelle Werkzeuge von einer anderen Art von selbstmodellierendem System verwendet werden, das offensichtlich die Kontrolle über seine eigenen Handlungen längst verloren hat. Können Sie sich vorstellen, wie es wäre, solch ein geistig behinderter phänomenaler Klon der ersten Generation zu *sein*? Können Sie sich vorstellen, wie es wäre, als ein etwas fortgeschritteneres künstliches Subjekt „zu sich selbst zu kommen“ – nur um zu entdecken, dass Sie, obwohl Sie ein Ichgefühl besitzen, einfach eine *Ware* sind, ein Objekt, ein wissenschaftliches Werkzeug, das nicht als ein Zweck in sich selbst erzeugt wurde und ganz bestimmt nicht als ein solcher behandelt werden wird?

Können Sie sich vorstellen, wie es wäre, solch ein geistig behinderter phänomenaler Klon der ersten Generation zu *sein*? Können Sie sich vorstellen, wie es wäre, als ein etwas fortgeschritteneres künstliches Subjekt „zu sich selbst zu kommen“ – nur um zu entdecken, dass Sie, obwohl Sie ein Ichgefühl besitzen, einfach eine *Ware* sind, ein Objekt, ein wissenschaftliches Werkzeug, das nicht als ein Zweck in sich selbst erzeugt wurde und ganz bestimmt nicht als ein solcher behandelt werden wird?

Leiden beginnt auf der Ebene des phänomenalen Selbstmodells (PSM). Sie können nicht bewusst leiden, ohne ein transparentes und

global verfügbares Selbstmodell zu haben. Das PSM ist das entscheidende neurokomputationale Instrument – nicht nur beim Erwerb vieler neuer kognitiver und sozialer Fähigkeiten, sondern auch dabei, ein stark bewusstes System zu *zwingen*, sich seinen eigenen Zerfall, seine eigenen Niederlagen und seine inneren Konflikte funktional und repräsentational *anzueignen* und sie dann auch subjektiv unhintergehbar als *die eigenen* zu erleben. Sensorischer Schmerz, aber auch alle anderen Arten des nichtkörperlichen Leidens, jeder repräsentationale Zustand, der durch eine negative Valenz charakterisiert ist und in das Selbstmodell eingebettet worden ist, werden jetzt phänomenal *besessen*. Leiden ist nun unweigerlich und erlebnismäßig unhintergehbar, auf transparente Weise, *mein eigenes* Leiden. Das Melodrama, aber auch die potenzielle Tragödie des Ego beginnt genau auf der Ebene transparenter Selbstmodellierung. Darum sollten wir alle Versuche, künstliche oder postbiotische PSMs zu erzeugen, aus der seriösen akademischen Forschung verbannen. Wir sollten die Erzeugung bewusster Selbstmodelle noch nicht einmal *riskieren*.

Wir sollten die Erzeugung bewusster Selbstmodelle noch nicht einmal *riskieren*.

Wir Menschen unterscheiden uns grundsätzlich in unseren *positiven* moralischen Intuitionen und auch bezüglich unserer expliziten Theorien darüber, wonach wir aktiv streben sollten. Die philosophische Ethik ist ein schwieriges Gebiet und es ist mehr als unklar, ob es überhaupt so etwas wie *Erkenntnis* von moralischen Normen gibt. In pluralistischen Gesellschaften müssen wir außerdem mit einer Vielfalt von moralischen Intuitionen und Argumenten gleichzeitig leben. Auf der anderen Seite können wir uns nicht von den praktischen Problemen zurückziehen: Entscheidungen müssen immer getroffen werden. Ich denke, manche Entscheidungen müssen bereits *jetzt* getroffen werden.

Unter pragmatischen Gesichtspunkten ist es deshalb wichtig, an ethische Prinzipien zu appellieren, die möglichst viele Menschen teilen können. Glücklicherweise gibt es in diesem Fall ein solches Prinzip. Ich nenne es das *„Prinzip des negativen Utilitarismus“*: Was immer sonst unsere genauen ethischen Verpflichtungen und Zielsetzungen sind, was immer sonst unsere speziellen Ideale und Zukunftsvisionen sein sollten, wir können und sollten sicherlich alle dem Prinzip zustimmen, dass die *Gesamtmenge des bewussten Leidens* bei allen Wesen, die leidensfähig sind, ständig und so weit wie möglich minimiert werden sollte. Ich weiß natürlich, dass es unmöglich ist, dieses Prinzip im Sinne einer Letztbegründung abzusichern. Außerdem gibt es eine Vielzahl wichtiger und

ernst zu nehmender theoretischer Komplikationen, die etwa Individualrechte und längerfristige Präferenzen betreffen. Trotzdem ist die zugrunde liegende Intuition etwas, das fast alle Menschen teilen können. Wir alle stimmen darin überein, dass man kein unnötiges Leid in die Welt bringen sollte. Camus hat einmal von der Solidarität aller endlichen Wesen gegen den Tod gesprochen und in demselben Sinne sollte es auch so etwas wie eine Solidarität aller bewussten, leidensfähigen Wesen gegen das *Leiden* geben. Aus dieser Solidarität heraus sollten wir nichts tun, was dazu führen kann, dass die Gesamtmenge des Leidens im Universum sich erhöht, und insbesondere nichts, was schon am Anfang mit *großer Sicherheit* dazu führen würde, dass sich die Gesamtmenge des körperlichen Leidens und der geistigen Verwirrung in der Welt erhöht. Bei der theoretischen und technologischen Modellierung so faszinierender Phänomene wie Bewusstsein und Ichgefühl oder der Erste-Person-Perspektive überhaupt haben wir als phänomenologische Grundlage und moralischen Ausgangspunkt nicht viel mehr als unsere eigene Form des phänomenalen Erlebens, so wie sie sich zufällig in der biologischen Evolution auf diesem Planeten entwickelt hat. Es ist schwer für uns, der Tatsache ins Auge zu sehen, dass im Verlauf der Bewusstseinsevolution ein Ozean des Leidens in der physikalischen Welt entstanden ist, der vorher einfach nicht existierte. Das mag deshalb so sein, weil Mutter Natur einfach nicht *wollte*, dass wir solchen Tatsachen zu genau ins Angesicht schauen. Jetzt ist diese Form des Bewusstseins die einzige, die wir wissenschaftlich untersuchen können und die wir auf technischen Trägersystemen modellieren können. Wir sind deshalb in großer Gefahr, all die negativen Aspekte des biologischen Bewusstseins auf künstlichen oder postbiotischen Trägersystemen zu *multiplizieren*, bevor wir überhaupt verstanden haben, woher all diese negativen Aspekte kommen, in genau welchen Eigenschaften unserer biologischen Geschichte, unserer Körper und unserer Gehirne sie verwurzelt sind und wie sie vielleicht sogar neutralisiert werden könnten. Darum sollten wir zuerst alle Anstrengungen – in der Philosophie genauso wie in den Neuro- und Kognitionswissenschaften – darauf richten, unser *eigenes* Bewusstsein und die Struktur unseres *eigenen* Leidens besser zu verstehen. Wir sollten uns an dem klassischen philosophischen Ideal der Selbsterkenntnis und an dem ethischen Minimalgebot der Leidensverminderung orientieren und nicht fahrlässig eine Evolution zweiter Stufe auslösen, die dann unserer Kontrolle entgleiten und die Gesamtmenge des bewussten Leidens im Universum weiter vermehren könnte. Wir sollten es nicht tun.

Literatur

Baars, B.J. (1988). *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B.J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. Oxford: Oxford University Press.

Baars, B.J. and Newman, J. (1994). A neurobiological interpretation of the global workspace theory of consciousness. In: Revonsuo und Kamppinen 1994.

Chalmers, D.J. (1995). Fehlende Qualia, schwindende Qualia, tanzende Qualia. In Metzinger 1995a.

Damasio, A.R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace & Company. Deutsche Übersetzung (2000): *Ich fühle, also bin ich*. München und Leipzig: List.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association* 60: 46p.

Dehaene, S. and Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79: 1–37.

Dretske, F. (1998). *Die Naturalisierung des Geistes*. Paderborn: mentis.

Edelman, G.M. (1989). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

Metzinger, T. (1993; ²1999). *Subjekt und Selbstmodell*. Paderborn: mentis.

Metzinger, T., Hrsg. (1995a; ⁴2001). *Bewusstsein – Beiträge aus der Gegenwartsphilosophie*. Paderborn: mentis.

Metzinger, T. (1995b). Generelle Einleitung: Das Problem des Bewusstseins. In: Metzinger 1995a.

Metzinger, T. (1995c). Ganzheit, Homogenität und Zeitkodierung. In: Metzinger 1995a.

Metzinger, T., Hrsg. (2000a). *Neural Correlates of Consciousness – Empirical and Conceptual Questions*. Cambridge, MA: MIT Press.

Metzinger, T. (2000b). Introduction: Consciousness research at the end of the Twentieth Century. In: Metzinger 2000a.

Metzinger, T. (2000c). The *subjectivity* of subjective experience: A representationalist analysis of the first-person-perspective. In: Metzinger 2000a.

Metzinger, T. (2000d). Die Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung für Nichtphilosophen in fünf Schritten. In: W. Greve (Hrsg.), *Psychologie des Selbst*. Weinheim: BELTZ / Psychologie Verlags Union.

Metzinger, T. (2002). *Being No One. The Self-Model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Millikan, R.G. (1989). Biosemantics. *Journal of Philosophy* 86: 281–297.

Moore, G.E. (1903). The refutation of idealism. *Mind* 12: 433–53.
Pöppel, E. (1972). Oscillations as possible basis for time perception. In: J.T. Fraser (ed), *The Study of Time*. Berlin: Springer.
Pöppel, E. (1978). Time perception. In: R. Held, H.W. Leibowitz and H.L. Teuber, eds., *Handbook of Sensory Physiology, Vol. VIII*. New York: Springer.
Pöppel, E. (1985). *Grenzen des Bewusstseins*. München: DTV.
Pöppel, E. (1988). *Mindworks: Time and Conscious Experience*. New York: Hartcourt Brace Jovanovich.
Pöppel, E. (1994). Temporal mechanisms in perception. *International Review of Neurobiology*, 37: 185–202.
Revonsuo, A., and Kamppinen, M. (1994) [eds]. *Consciousness in Philosophy and Cognitive Neuroscience*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Ruhnau, E. (1995). Time-Gestalt and the observer. In: Metzinger 1995a.
Schopenhauer, A. (1977; 1844). *Die Welt als Wille und Vorstellung II*. Zürich: Diogenes.
Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, 59. Deutsch (*„Kann eine Maschine denken?“*) In: Enzensberger, H.M. (1967) [ed]. *Kursbuch*, 8 (März 1967). Frankfurt am Main: Suhrkamp.
Yates, J. (1985). The content of awareness is a model of the world. *Psychological Review* 92: 249–84.