Dies ist lediglich eine Rohfassung. Die Endfassung dieses Texts erschien in:
Stephan, A. & Walter, S. (2012). *Handbuch Kognitionswissenschaft*. Stuttgart: J. B. Metzler.

IV. Kognitive Leistungen

19. Selbst, Selbstmodell, Subjekt

‚Das' Selbst in der Alltagspsychologie und das theoretische Problem des Selbstbewusstseins

Die aktuelle kognitionswissenschaftliche Erforschung des Selbstbewusstseins hat Ihre historischen Wurzeln sowohl in einer unreflektierten, aber weit verbreiteten alltagspsychologischen Sprechweise als auch in einer sich über viele Jahrhunderte erstreckenden philosophisch-theologischen Debatte darüber, was der innerste Kern oder das eigentliche ‚Wesen' einer Person ist. Gibt es so etwas wie eine *Essenz* des Menschen? Was sind die Identitätskriterien für kognitive Systeme im Allgemeinen, was macht ein solches System z.B. über die Zeit hinweg zu dem *selben* System?

Sowohl die Alltagspsychologie als auch die traditionellen metaphysischen Modelle des Selbst haben ihre historischen Ursprünge in archaisch-mythischen Selbstbildern des Menschen und in der Frage nach der Unsterblichkeit der Seele (Barresi/Raymond 2011; Oeing-Hanoff et al. 1974). Unser alltagspsychologisches Sprachspiel ist allerdings in mehreren Hinsichten begrifflich verwirrt.

- Es gibt weder auf empirischer Ebene noch in begrifflicher Hinsicht überzeugende Hinweise darauf, dass ein die Zeit überdauerndes Einzelding oder eine im ontologischen Sinne autonome Substanz existieren, die ‚das' Selbst sein könnten (Metzinger 2011). Menschliche Wesen sind dynamische, sozial situierte *Systeme*; Selbstbewusstsein ist kein Ding, sondern ein diskontinuierlicher Vorgang, der zeitweise bestimmte Fähigkeiten erzeugt, die begrifflich am besten als globale Systemeigenschaften beschrieben werden, weil sie klarerweise eine biologisch

fundierte Funktion für das System als Ganzes besitzt. Das bedeutet zum Beispiel, dass der Besitz von phänomenalem Selbstbewusstsein eine Eigenschaft der Person als Ganzer ist und nicht eine Eigenschaft ihres Gehirns.

- ‚Ich' – das Personalpronomen der ersten Person Singular – bezeichnet immer den Sprecher, der es aktuell verwendet. Seine logische Funktion ist nicht die eines Gattungsbegriffs, sondern die der Selbstlokalisation eines Sprechers in einem Äußerungskontext. In grammatischer und semantischer Hinsicht ist ‚Ich' also ein singulärer Term, der an einen bestimmten Äußerungskontext gebunden ist: Dieser Kontext besteht darin, dass der aktuelle Sprecher mit einem sprachlichen Werkzeug auf sich selbst zeigt.
- Trotzdem verwenden wir bei der sprachlichen Selbstbezugnahme den indexikalischen Ausdruck ‚Ich' sehr häufig so, als ob es sich dabei um einen Namen für ein inneres Ding oder eine Form von Objektreferenz, von Bezugnahme auf einen Gegenstand handelt (Beckermann 2010; Bennett/Hacker 2010, Kap. 12.4). Es gibt aber keine spezielle Gattung von Dingen (‚Iche' oder ‚Selbste'), die man in sich tragen könnte wie ein Herz oder besitzen könnte wie ein Fahrrad oder einen Fußball.
- Das in lebensweltlichen Kontexten allgegenwärtige Reden von unserem oder ‚meinem' Selbst ist in sich widersprüchlich, weil es dann ja schon jemanden geben müsste, der das Selbst ‚hat', also ein Selbst hinter dem Selbst, das zu diesem in einer Besitzrelation steht. Das Selbst kann auch nichts ‚in mir' sein, weil dann ja das, mit dem ich identisch bin, nur ein konstituierender Teil von mir wäre.

Ähnliche Probleme haben die klassischen Reflexionsmodelle des Selbstbewusstseins, wie sie z.B. von den Philosophen des deutschen Idealismus entwickelt wurden: Wenn Selbstbewusstsein sich durch eine mysteriöse Form der inneren Selbstbeobachtung

konstituiert, in der ‚das Bewusstsein' Subjekt und Objekt zugleich ist oder in der ‚das Ich' sich selbst zum Gegenstand der eigenen Anschauung und des eigenen Denkens macht, oder wenn es in einer inneren Handlung ‚sein Sein setzt', dann wird das, was eigentlich zu erklären ist, immer schon vorausgesetzt (Frank 1991, 1994; Henrich 1966). Eine neuere Formulierung der philosophischen Grundproblematik finden wir in dem Begriff ‚Homunkulus-Fehlschluss' (*homunculus fallacy*). Diesen Fehlschluss begeht man immer dann, wenn man innerhalb eines kognitiven Systems ein kleines Männchen postuliert, das etwa Operationen auf mentalen Repräsentationen durchführt, sie betrachtet oder interpretiert. Man schreibt dann globale Eigenschaften – die nur die Person oder das System als Ganzes besitzen – subpersonalen Aspekten der Informationsverarbeitung zu. Der Homunkulus-Fehlschluss ist meistens eine Variante des ‚mereologischen Fehlschlusses', bei dem Eigenschaften des Ganzen mit Eigenschaften von Teilen verwechselt werden. Man kann diesen Fehlschluss auch nicht umgehen, indem man aus Verlegenheit einfach Anführungszeichen verwendet und augenzwinkernd über das ‚Ich' oder das ‚Selbst' schreibt. In den Worten von Maxwell Bennett und Peter Hacker: „Die Anführung ist der stille Tribut, den die falsche Grammatik der richtigen zollt" (2010, 449f.).

Manche Neurowissenschaftler versuchen, dieses Problem der Reifikation rein stilistisch durch literarische Neuschöpfungen zu entschärfen, indem sie nicht mehr von ‚dem Selbst' sprechen, sondern schlicht und feierlich von ‚Selbst' (analog der Heidegger'schen Rede von ‚Welt' oder des ‚Geschehens von Welt'). Prominente Beispiele sind „*neuroscientific studies of self*" (Vogeley/Gallagher 2011, 129), Wendungen wie „*how is self operationalized?*" (Gillihan/Farah 2005, 77) oder etwa schon der Titel von Antonio Damasios Buch *Self comes to Mind* (2010). Die Homunkulus-Variante dagegen findet man z.B. besonders häufig in Formulierungen wie ‚Zellen in dieser Region des visuellen Cortex sehen primär

Kanten‘ oder ‚der präfrontale Cortex plant die Handlung, während der prämotorische Cortex Fremdbewegungen analysiert und über die Bewegungsinitiierung sowie die Organisation der Bewegungssequenzen entscheidet‘. Sehen und Entscheiden sind aber immer Leistungen des jeweiligen Systems als einer verkörperlichten und situierten Ganzheit. Begrifflich gesehen sind sie *globale* Eigenschaften – z.B. die einer bewussten, menschlichen Person. Es gibt also jeweils eine globale Funktion (‚Sehen‘, ‚Handlungskontrolle‘), wobei die einzelnen Teilfunktionen, durch die sie konstituiert wird, jeweils durch einzelne Teile des Systems oder dynamische Interaktionen mit einer Umwelt realisiert sein können. In diesem Sinne sind auch Selbstbewusstsein, Selbstwissen und mentale Selbstrepräsentation aus kognitionswissenschaftlicher Sicht spezifische und funktional analysierbare *Fähigkeiten*, die ein informationsverarbeitendes System zu einem gegebenen Zeitpunkt besitzen kann oder auch nicht.

Problematisch für die kognitionswissenschaftliche Forschung ist der begrifflich unklare Hintergrund z.B. deshalb, weil viele Neurowissenschaftler die logisch inkonsistente alltagspsychologische Sprechweise direkt in ihre Theorien übernehmen oder bei der Formulierung von Hypothesen verwenden. Ein Ausdruck wie ‚*self-processing*‘ (z.B. Blanke/Arzy 2005; Kircher et al. 2000) ist problematisch, weil er voraussetzt, dass ein System ein mysteriöses, konkretes Ding wie ‚das‘ Selbst oder sogar einfach ‚sich selbst‘ verarbeitet, ähnlich wie eine Wurstmaschine Fleischbrocken zu Würsten verarbeitet. Die richtige Sprechweise wäre hier, zu sagen, dass ein System bestimmte kognitionswissenschaftlich zu erklärende Fähigkeiten (z.B. die Wiedererkennung des eigenen Gesichts) oder ein spezifisches phänomenologisches Profil (z.B. eine außerkörperliche Erfahrung) dadurch erwirbt, dass es bestimmte Eigenschaften mit Hilfe eines integrierten Selbstmodells als globale Eigenschaften seiner selbst darstellt (s.u.). Im Fall von systematischen Selbsttäuschungen (Von Hippel/Trivers

2011), der Gummihand-Illusion (Botvinick/Cohen 1998) oder Ganzkörperillusionen (Blanke/Metzinger 2009; Lenggenhager et al. 2007) könnten dies durchaus auch Eigenschaften sein, die das System aus der Außenperspektive überhaupt nicht besitzt. Fehlrepräsentationen können funktional adäquat sein, und dies gilt auch für den Sonderfall der Selbstrepräsentation. In den meisten Fällen wird das Selbstmodell aber Informationen über tatsächlich vorhandene globale Eigenschaften intern darstellen. Damit wird dann typischerweise eine bestimmte Klasse von Tatsachen für die Handlungskontrolle verfügbar – z.B. Tatsachen, die die physische Realisierung des Systems betreffen (Sauerstoffgehalt des Blutes, Geschwindigkeit und Stellung der Effektoren im Raum, Zustände der Sensoren). Globale Selbstrepräsentation ist also ein Vorgang der funktionalen Aneignung von Fakten. Ein wichtiger Aspekt der *Autonomie* eines Systems liegt in genau der Information, die jeweils zur rationalen Selbstkontrolle zur Verfügung steht. Dazu gehören dann auch abstraktere Tatsachen wie individuelle Präferenzen oder die aktuelle soziale Situiertheit des Systems (etwa Hungergefühle und Zielzustände innerhalb des emotionalen Selbstmodells, das Selbstwertgefühl oder die Stellung in einer Dominanzhierarchie), aber auch das explizite Wissen um die eigene Rationalität, die ethische Bewertbarkeit des eigenen Handelns oder selbstgerichtete Überzeugungen wie die, dass man selbst ein rationales Subjekt ist, welches den Status einer Person besitzt.

Auf der anderen Seite ist auch richtig, dass die klassische sprachanalytische Auflösung des Problems *allein* oberflächlich bleibt. Wenn man Selbstwissen, Selbstbewusstsein und Subjektivität lediglich über eine Untersuchung der logischen Funktion des indexikalischen Ausdrucks „Ich“ und eine beschreibende Analyse der semantischen Besonderheiten des sprachlichen Selbstbezugs mit Hilfe des Pronomens der ersten Person Singular zu verstehen versucht, dann blendet man nicht nur die historisch-evolutionäre Tiefendimension und die

neurobiologische Fundierung seines Erkenntnisgegenstandes aus, sondern auch die Ebene des bewussten Erlebens. Der klassische Ansatz der analytischen Philosophie sagt uns nichts über die konstituierenden Bedingungen erfolgreicher Selbstbezugnahme und ignoriert z.B. die Phänomenologie der Substantialität. Er ist deshalb unbefriedigend, weil er drei für die Kognitionswissenschaft zentrale Erkenntnisziele ignoriert:

- Die *phänomenologische Tiefenstruktur* des menschlichen Selbstbewusstseins, in der die Intuitionen verwurzelt sind, aus denen die Widersprüchlichkeiten der Alltagspsychologie überhaupt erst entstehen.
- Die *repräsentationale Architektur* der mentalen Selbstrepräsentation, die nicht nur bei verschiedenen Typen von Sprechern, sondern bei bestimmten Klassen von kognitiven Systemen sehr unterschiedlich sein kann (z.B. bei autonomen Robotern, nicht-menschlichen Tieren, Säuglingen, träumenden Personen oder psychiatrischen Patienten).
- Die *feinkörnigen funktionalen Eigenschaften*, durch die ein bestimmter Typ von Selbstmodell die *Fähigkeiten* konstituiert, die es in empirischer Hinsicht zu erklären gilt.

Ein wichtiger begrifflicher Unterschied ist hier der zwischen Selbstwissen und Selbstbewusstsein. Ein kognitives System könnte durchaus ein reiches und parallel in verschiedenen Formaten dargestelltes Wissen über seine eigenen Fähigkeiten und Eigenschaften besitzen, ohne dass es diese Tatsache auch subjektiv erlebt, also ohne dass in ihm global integrierte phänomenale Zustände auftreten, ohne raumzeitliche Selbstlokalisierung und ohne die Entstehung einer introspektiven Innenperspektive. Auf der anderen Seite könnte ein träumendes oder halluzinierendes System sehr intensive subjektive Erlebnisse durchlaufen und trotzdem fast ausschließlich falsche Überzeugungen über sich

selbst besitzen (Metzinger 2013, Windt 2015).

Worum es kognitionswissenschaftlich geht, ist also primär eine sehr spezielle Liste von *Fähigkeiten*:

- Genau welche Mechanismen erlauben es einem Sprecher, das Erste-Person-Pronomen ‚Ich' korrekt zu verwenden?
- Das *acquisition constraint*: Wie erwirbt ein System *graduell* die Fähigkeit, sprachlich auf sich selbst Bezug zu nehmen (Bermudez 1998)?
- Welche funktionale Rolle spielt bei dieser Fähigkeit die *kognitive* Selbstbezugnahme, welche die *phänomenale* Selbstrepräsentation (Metzinger 2003)?
- *Embodiment*: Auf welche Weise ist die abstrakte geistige Selbstbezugnahme (etwa die Fähigkeit, Begriffe wie ‚Person' oder ‚rationales Subjekt' intern auf sich selbst anzuwenden) in funktional basaleren Fähigkeiten verankert, also in nicht-begrifflichen Formen der Selbstrepräsentation, etwa Emotionen oder motorischen Simulationen (Knoblich et al. 2003; Tsakiris 2011)?
- Was ist bei biologischen Systemen die *evolutionäre Geschichte* dieser Fähigkeiten gewesen (Taylor Parker et al. 1994)?
- Wann ist es generell für ein System *adaptiv*, sich selbst als eine Ganzheit zu repräsentieren und globale Eigenschaften seiner selbst intern zu repräsentieren (z.B. Gallup 1997; Gallup et al. 2011)?
- Bei welchen Formen von Selbstrepräsentation handelt es sich um eine Form von *Wissen*, wann waren bestimmte Formen der Selbsttäuschung funktional adäquat (Trivers 2011)?
- Was ist das computationale Ziel des Selbstbewusstseins, worin liegen seine Vorteile? In genau welchen Fällen erfolgt die Verarbeitung selbstbezogener Informationen auf der

bewussten Ebene, gibt es spezielle Vorteile der *phänomenalen Selbstrepräsentation*? Was auf philosophischer Ebene benötigt wird, ist ein plausibler Nachfolgebegriff, der an die Stelle der unklaren Metaphysik und Alltagspsychologie des ‚Ich' oder ‚Selbst' tritt. Ein solches begriffliches Werkzeug müsste eine minimale Menge von Kriterien erfüllen: Es müsste logisch widerspruchsfrei sein, in empirischen Daten verankert sein und es müsste in seiner Semantik ständig durch neue Erkenntnisse angereichert und dabei natürlich auch revidiert werden können. Für die Kognitionswissenschaft sind allerdings mindestens zwei zusätzliche Eigenschaften von zentraler Bedeutung: Der Nachfolgebegriff muss einerseits so angelegt sein, dass er in einzelnen Disziplinen (etwa der kognitiven Neuropsychiatrie oder der Computerlinguistik) feinkörnige Analysen und die Formulierung sehr spezifischer Hypothesen möglich macht; andererseits muss er eine hinreichende Generalität besitzen, um als Brücke für die interdisziplinäre Zusammenarbeit funktionieren zu können. Ein möglicher Kandidat ist der Begriff eines ‚Selbstmodells'.

Selbstmodell

Der amerikanische Philosoph Josh Weisberg hat den Begriff der „*method of interdisciplinary constraint satisfaction*" (MICS) geprägt (Weisberg 2005; Metzinger 2004a). Diese Methode besteht darin, gleichzeitig auf einer Vielzahl verschiedener Beschreibungsebenen sowohl empirische als auch begriffliche Auflagen zu erfüllen, z.B. für eine umfassende Theorie des Selbstbewusstseins. Das Ziel besteht darin, ein komplexes Erkenntnisziel gewissermaßen zu ‚triangulieren', indem man verschiedene Methoden und Informationsquellen gleichzeitig nutzt, um ein heuristisch fruchtbares Arbeitskonzept zu konstruieren. Dabei ist es eine zentrale Aufgabe der Philosophie der Kognitionswissenschaft, aus metatheoretischer Perspektive begriffliche Instrumente zu entwickeln, welche eine *Integration* über die verschiedenen Ebenen der Analyse hinweg ermöglichen und im Idealfall einen formalen

Rahmen bereitstellen, der dann verschiedene Datensätze und unterschiedliche theoretische Herangehensweisen zusammenführen kann. Die ‚Selbstmodell-Theorie der Subjektivität' (SMT) ist ein Beispiel für einen solchen Versuch (Metzinger 2004a, 2006, 2008, 2009, 2011).

SMT formuliert auf der phänomenologischen, der repräsentationalistischen, der funktionalen und der neurobiologischen Beschreibungsebene gleichzeitig sowohl begriffliche als auch empirische Auflagen. Dabei handelt es sich um einschränkende Bedingungen, die den Raum denkbarer Lösungen immer weiter einengen und uns dann innerhalb eines bestimmten Gegenstandsbereichs dabei helfen sollen zu klären, was es für die betreffende Klasse von Systemen bedeutet, Selbstbewusstsein zu besitzen (Gallagher 2011).

Zunächst ist es wichtig, die ontologische Generalthese der SMT zu verstehen: Einzeldinge oder Substanzen wie ‚Selbste' existieren in der Welt nicht (Metzinger 2011). Deshalb kann man den Begriff des ‚Selbst' als einer theoretischen Entität für alle wissenschaftlichen und philosophischen Zwecke problemlos eliminieren. Was wir in der Vergangenheit, und insbesondere alltagspsychologisch, ‚das' *Selbst* genannt haben, ist keine ontologische Substanz, keine kontextunabhängige und unwandelbare Essenz und auch keine besondere Art von Ding (d.h. kein Individuum im Sinne der philosophischen Metaphysik), sondern ein dynamischer Vorgang, nämlich die Selbstorganisation einer sehr speziellen Art von repräsentationalem Inhalt in einer sehr speziellen Art von informationsverarbeitendem System. Es ist der Inhalt eines Selbstmodells, das von dem System, das es benutzt, introspektiv nicht *als* Modell erlebt werden kann. Der dynamische Inhalt des phänomenalen Selbstmodells (PSM) ist somit der Inhalt dessen, was wir in der Vergangenheit als ‚das' bewusste Selbst bezeichnet haben: Meine aktuellen Körperempfindungen, mein gegenwärtiger emotionaler Zustand und alle Inhalte meiner phänomenal erlebten Kognition. Diese bilden den

repräsentationalen Inhalt meines PSM. Auf der funktionalen Beschreibungsebene bilden genau jene Eigenschaften des Selbstmodells, auf die ich in diesem Moment prinzipiell meine Aufmerksamkeit richten kann, den Inhalt meines aktuellen PSM. Dieses PSM ist kein Ding, sondern ein integrierter Vorgang, der episodisch in meinem Gehirn abläuft. Das PSM kann man also gleichzeitig auf der repräsentationalistischen, der funktionalistischen und – bei Biosystemen – auf der neurowissenschaftlichen Beschreibungsebene näher untersuchen. Auch künstliche Systeme mit unbewussten Selbstmodellen existieren bereits (Bongard et al. 2006).

Intuitiv – und in einem gewissen metaphorischen Sinn – besteht für manche an dieser Stelle vielleicht die Versuchung zu sagen, dass ich als bewusstes Selbst der Inhalt meines PSM *bin*. Das wäre jedoch wieder der am Anfang erwähnte mereologische Fehlschluss. In Wirklichkeit bin ich natürlich das System als Ganzes: Ich bin das dynamisch an innere und äußere Umwelten gekoppelte und auch sozial situierte System, inklusive des jetzt gerade in seinem Gehirn aktiven Selbstmodells. Allerdings kann ich den Unterschied zwischen dem System und dem Teil des Systems, der als sein bewusstes Modell funktioniert, durch die introspektive Lenkung von Aufmerksamkeit nicht entdecken. Ich kann die Tatsache, dass das Selbstmodell eine *Repräsentation* ist, auf den meisten Inhaltsebenen nicht subjektiv erleben. Das liegt daran, dass die meisten Partitionen des menschlichen PSM transparent sind, etwa das Modell des eigenen Körpers (dies ist die vielleicht wichtigste begriffliche Auflage der repräsentationalistischen Beschreibungsebene; vgl. dazu Metzinger 2003, 2006). Eine häufig übersehene phänomenologische Auflage für Theorien des Selbstbewusstseins ist aber auch die phänomenale Opazität des kognitiven Selbstmodells: Als Denker von Gedanken erlebe ich mich selbst eben gerade als ein Subjekt, das mentale *Repräsentationen* erzeugt, strukturiert und in einander transformiert. Die Phänomenologie der Kognition zeichnet sich

interessanterweise dadurch aus, dass die *Repräsentationalität* der fraglichen inneren Zustände mir auch auf der Ebene des bewussten Erlebens verfügbar ist

‚Phänomenal transparent' ist eine Repräsentation dann, wenn das kognitive System, in dem sie auftaucht, sie introspektiv nicht mehr *als* eine Repräsentation erkennen kann und sich deshalb als direkt mit ihrem Inhalt in Kontakt erlebt. Z.B. erleben Sie das Buch, das sie gerade in ihren Händen halten, nicht mehr als den Inhalt einer Repräsentation in ihrem Gehirn, weil die visuelle und taktile Repräsentation des Buchs in Ihren Händen so schnell und zuverlässig aufgebaut wird, dass Sie sie subjektiv nicht mehr als den Inhalt eines inneren Zustands erleben können. Ein Hauptargument der Selbstmodell-Theorie besagt, dass ein bewusst erlebtes Selbst genau dann entsteht, wenn dasselbe auch für das Selbstmodell gilt: Wir sind Wesen, die ihr eigenes inneres Modell von sich selbst nicht mehr *als* ein Modell erleben können und die deshalb naive Realisten auch bezüglich ihrer eigenen Existenz sind. Wir erleben uns notwendigerweise als in direktem und unmittelbarem Kontakt mit uns selbst. Was eine phänomenale Repräsentation transparent macht, ist also die funktionale Tatsache, dass frühere Verarbeitungsstufen im Gehirn für die nicht-begriffliche Introspektion für die innere Aufmerksamkeit nicht verfügbar sind. Die Mittel der Repräsentation können selbst nicht *als solche* repräsentiert werden. Deshalb ist das System, das die Erfahrungen macht, hinsichtlich der entsprechenden Inhalte und mit begrifflicher Notwendigkeit in einem naiven Realismus gefangen: In Standardkonfigurationen haben die meisten Inhalte des phänomenalen Erlebens einen unhintergehbar realistischen Charakter. SMT wendet also die Transparenz-Auflage auf das PSM an.

Zumindest für alle uns bekannten bewussten Wesen gilt, dass sie weder ein Selbst *haben*, noch ein Selbst *sind*. Was sie haben, ist ein Selbst*modell* – und dies ist letztlich ein komplexer Gehirnzustand. Es gibt zwar biologische Organismen, aber ein Organismus ist

natürlich noch lange kein Selbst. Manche Organismen besitzen bewusste Selbstmodelle, aber solche Selbstmodelle sind mit Sicherheit keine *Selbste* – sie sind lediglich komplexe Gehirnzustände. Wenn ein Organismus auf der Basis eines transparenten Selbstmodells operiert, dann besitzt er ein *phänomenales* Selbst. Die phänomenale Eigenschaft des ‚Ich-Gefühls' oder der ‚Selbstheit' als solche ist also ein repräsentationales Konstrukt, denn sie ist eine interne, dynamische Repräsentation des Organismus als Ganzem, die in ein virtuelles Gegenwartsfenster eingebettet wurde und die Transparenz-Bedingung erfüllt. Sie ist tatsächlich eine *phänomenale* Eigenschaft in dem Sinne, dass sie nur eine Erscheinung ist.

Man könnte den zentralen Gedanken auch dadurch auszudrücken versuchen, dass man sagt, wir seien Systeme, die sich unentwegt selbst mit dem Inhalt ihres PSM *verwechseln*. Aber auch diese Metapher der ‚Ich-Illusion' enthält natürlich bei näherem Hinsehen einen logischen Fehler: Täuschung und Wissen im Sinne propositionaler Inhalte gibt es auf der fraglichen Ebene überhaupt noch nicht, es gibt niemand, der *sich* täuschen könnte (dies wäre wieder der eingangs erwähnte Homunkulus-Fehlschluss). Im Gegenteil: Die phänomenologische Grundstruktur, um die es hier geht, ist ja genau die Struktur, die die Entstehung eines echten epistemischen Subjekts überhaupt erst ermöglicht. Eines der relevantesten Erkenntnisziele für die Kognitionswissenschaft sind die subpersonalen Bedingungen der Möglichkeit von Personalität, eine präzise Beschreibung des Übergangs vom Körpermodell zum Personenmodell. Der Besitz eines PSMs ist die zentrale notwendige Bedingung der Möglichkeit komplexerer Formen von Wissen und Erkenntnis. Es gibt keine überzeugenden Argumente gegen die logische Möglichkeit, dass künstliche Systeme diese Eigenschaft nicht ebenfalls instanziieren könnten.

SMT und der hypothetische Begriff eines Selbstmodells haben ihre heuristische Fruchtbarkeit schon in vielen Bereichen unter Beweis gestellt: Viele neurologische und

psychiatrische Störungsbilder kann man als Störungen des menschlichen Selbstmodells genauer analysieren (Metzinger 2004a, 2004b). Z.B. kann man Somatoparaphrenien oder bestimmte Positivsymptome der Schizophrenie wie die Gedankeneingebung als funktionale Konfigurationen analysieren, in denen das System existierende Repräsentationen von Körperteilen oder der eigenen kognitiven Vorgänge nicht mehr in das PSM integrieren kann. Die psychosomatischen Erkrankungen zugrundeliegenden psychophysischen Korrelationen und die ätiologischen Kausalrelationen lassen sich so wesentlich genauer als Wechselwirkungen zwischen den bewussten und den unbewussten Schichten des Selbstmodells beschreiben. Der Unterschied zwischen Traum und Wachzustand, die spezifischen kognitiven Defizite im REM-Schlaf, aber auch der Übergang vom Traum in den luziden Traum können als Verschiebungen im funktionalen Profil des Selbstmodells besser verstanden werden (Windt 2010, 2015; Windt/Metzinger 2007). Lokale Körperillusionen wie die bereits erwähnte Gummihand-Illusion (Botvinick/Cohen 1998), manche Störungen der Willkürmotorik, aber auch das Phänomen der halluzinierten Agentivität (Wegner/Wheatley 1999) erscheinen als Fehlrepräsentationen, in denen bereits im Gehirn aktive repräsentationale Inhalte in das Selbstmodell eingebettet werden und dadurch automatisch mit der phänomenalen Eigenschaft der ‚Meinigkeit' versehen werden: Was immer vom Gehirn funktional in das gegenwärtig aktive PSM eingebettet wird, wird von der betreffenden Person unhintergehbar als *eigener* Zustand erlebt. Spontan auftretende Phänomene wie OBEs (*out-of-body experiences*), nicht-visuelle Ganzkörperillusionen wie das pathologische Gefühl einer Anwesenheit oder FOP (*feeling of a presence*), neurologische Störungen wie die Heautoskopie oder die Autoskopie lassen sich unter SMT taxonomisch genauer klassifizieren (Blanke/Metzinger 2009). Es ist auch bereits gelungen, aus der Theorie heraus im Labor neue Formen von Ganzkörperillusionen zu erzeugen (Lenggenhager et al. 2007), also einzelne

Dimensionen des körperlichen Selbstbewusstseins wie die Selbstlokalisation in einem räumlichen Bezugsrahmen, die subjektive Identifikation mit dem Inhalt eines Körperbilds oder den Ursprung einer visuellen Perspektive funktional voneinander zu dissoziieren.

Ein wichtiges theoretisches Ziel besteht in der Entwicklung eines *Minimalmodells* für das phänomenale Selbst: Was ist die minimal hinreichende Menge von Eigenschaften, die beim Menschen die einfachste Form des Selbstbewusstseins entstehen lässt? Die einfachste Form eines Selbstmodells genauer zu beschreiben wäre deshalb wichtig, weil es in vielen Fällen die funktionale Plattform sein wird, auf deren Grundlage höhere kognitive Fähigkeiten entstehen können.

Der Besitz eines integrierten Selbstmodells bringt erstmals die Möglichkeit mit sich, dass ein kognitives System ganz bestimmte funktionale oder repräsentationale *Fähigkeiten* entwickelt:

- Die korrekte Verwendung des Erste-Person-Pronomens ‚Ich' setzt eine *nicht-begriffliche* Form des Selbstbewusstseins voraus (Bermudez 1998). Weil mentale Modelle (Johnson-Laird 1983; Knauff 2009) ein wesentlicher Bestandteil des Sprachverstehens sind, ist das Selbstmodell auch notwendiger Bestandteil der Fähigkeit, sprachlich auf sich selbst Bezug zu nehmen.
- Die *kognitive* Selbstbezugnahme setzt eine Interaktion zwischen opaken und transparenten Schichten des menschlichen Selbstmodells voraus (Metzinger 2003).
- *Embodiment* als eine Eigenschaft kognitiver Systeme entsteht auf mehreren funktionalen und repräsentationalen Ebenen gleichzeitig, wobei das unbewusste und das bewusste Körpermodell sich beim Menschen sehr deutlich als zentraler, fundierender Aspekt des Selbstmodells gezeigt haben (Metzinger 2014).
- Es ist klar erkennbar, dass es bei biologischen Systemen eine *evolutionäre Geschichte*

der internen Selbstmodellierung gab. Z.B. scheint das Körpermodell eine entscheidende Rolle bei der Evolution des Werkzeuggebrauchs gespielt zu haben (Maravita/Iriki 2004; Metzinger 2009).

- Sich selbst als eine Ganzheit zu repräsentieren und globale Eigenschaften seiner selbst intern zu repräsentieren, ist insbesondere in den Bereichen sozialer Kognition und Selbsttäuschung (Trivers 2011; von Hippel/Trivers 2011), Raumkognition und bei der globalen Motorkontrolle *adaptiv* und funktional adäquat.

Subjekt

Welche Art von Selbstmodell müsste ein kognitives System besitzen, um durch den Aufbau wechselseitiger Anerkennungsbeziehungen in einem sozialen Kontext seinen Status als *Person* zu etablieren? Was genau wäre, auf der repräsentationalistischen und funktionalistischen Beschreibungsebene, der Schritt vom Selbstmodell zum *Subjektmodell*, zum Modell eines rationalen epistemischen Agenten, der auch die ethische Dimension seiner eigenen Handlungen erfassen kann? In der interdisziplinären Erforschung des phänomenalen Bewusstseins zeigt sich, dass phänomenales Erleben nicht ein einziges Problem ist, sondern dass wir es hier mit einem ganzen Bündel von epistemischen Zielen zu tun haben: Manche sind metatheoretischer Natur, andere rein empirisch; bei manchen handelt es sich eher um begriffliche und philosophische Fragestellungen, bei anderen eher um die Isolierung minimal hinreichender physischer Korrelate durch die kognitive Neurowissenschaft oder um die mathematische Modellierung der funktionalen Feinstruktur solcher Korrelate (Metzinger 1995, 2000). Es gibt jedoch so etwas wie ein *Kernproblem* – eine einzelne theoretische Problematik, welche die Problemlandschaft integriert: Die Subjektivität mentaler Zustände. Was genau könnte es bedeuten, dass ein kognitives System eine Erste-Person-Perspektive (1PP) entwickelt? Welche kognitiven Leistungen können ohne eine 1PP *nicht* erbracht

werden?

Dieses Kernproblem ist etwas, das wir in der Physik, der Chemie und der Biologie nicht finden. Wenn man Bewusstsein als ein Erkenntnisziel der naturwissenschaftlichen Forschung betrachtet, dann zeigt sich, dass die Zustände des phänomenalen Bewusstseins sich von physikalischen, chemischen oder biologischen Zuständen dadurch unterscheiden, dass sie fast immer an eine individuelle 1PP gebunden sind. In der Vergangenheit waren es fast ausschließlich Philosophen, die sich mit dem Problem der Subjektivität beschäftigt haben. Die Selbstmodell-Theorie sagt z.B., dass eine 1PP genau dann entsteht, wenn ein System sich als *epistemischen Agenten* modelliert, wenn es also in phänomenaler Echtzeit die repräsentationalen und agentivischen *Subjekt-Objekt-Beziehungen*, in denen es zur Welt steht, noch einmal ko-repräsentiert und dadurch ein internes Modell der Intentionalitätsbeziehung erzeugt. Wenn es tatsächlich gelänge, Subjektivität und die 1PP zu ‚naturalisieren' – wenn also sozusagen die Subjektivität des Mentalen selbst in ihrem vollen Gehalt mit den empirischen Methoden der Kognitionswissenschaft traktabel würde, wenn die Evolution und Ontogenese bewusster kognitiver Zustände *in ihrer Subjektivität* zum Gegenstand erfolgreicher reduktiver Erklärungen würde – dann wäre dies ein großer Fortschritt. Dazu müssten jedoch die Phänomenologie, die Erkenntnistheorie und die Ontologie der 1PP begrifflich wesentlich genauer differenziert werden als in der Vergangenheit. Dieses Erkenntnisziel würde es jedoch auch erforderlich machen, die Philosophie des Geistes noch stärker und auf einem völlig neuen Niveau inhaltlicher Komplexität mit der empirischen Arbeit in den vielen verschiedenen Disziplinen der Kognitionswissenschaft zu integrieren.

Literatur

Barresi, John/Martin, Raymond (2011): History as prologue – Western theories of the self. In: Shaun Gallagher (Hg.): *The Oxford Handbook of the Self.* Oxford, 33–?56

Beckermann, Ansgar (2010): Die Rede vom Ich und vom Selbst – Sprachwidrig und philosophisch höchst problematisch. In: Katja Crone/Robert Schnepf/Jürgen Stolzenberg (Hg.): *Über die Seele*. Frankfurt a.M., 458–?473

Bennett, Maxwell/Hacker, Peter (2010): *Die philosophischen Grundlagen der Neurowissenschaften* [*Philosophical Foundations of Neuroscience*, Oxford]. Darmstadt.

Bermúdez, Jose (1998): *The Paradox of Self-consciousness*. Cambridge, MA.

Blanke, Olaf/Arzy, Shahar (2005): The out-of-body experience. In: *The Neuroscientist* 11, 16–24.

Blanke, Olaf/Metzinger, Thomas (2009): Full-body illusions and minimal phenomenal selfhood. In: *Trends in Cognitive Sciences* 13**,** 7–13.

Bongard, Josh/Zykov, Victor/Lipson, Hod (2006): Resilient machines through continuous self-modeling. In: *Science*, 314, 1118.

Botvinick, Matthew/Cohen, Jonathan (1998): Rubber hand 'feels' touch that eyes see. In: *Nature* 391, 756.

Damasio, A. (2011): *Self comes to Mind*. New York.

Frank, Manfred (Hg.) (1991): *Selbstbewußtseinstheorien von Fichte bis Sartre*. Frankfurt a.M.

Frank, Manfred (Hg.) (1994): *Analytische Theorien des Selbstbewußtseins*. Frankfurt a.M.

Gallagher, Shaun (Hg.) (2011): *The Oxford Handbook of the the Self*. Oxford.

Gallup, Gordon (1997): On the rise and fall of self-conception in primates. In: Joan Gay Snodgrass/Robert Thompson (Hg.): *The Self across Psychology*. New York, NY, 4–17.

Gallup, Gordon/Anderson, James/Platek, Steven (2011): Self-recognition. In: Shaun Gallagher (Hg.): *The Oxford Handbook of the Self.* Oxford, 80–110.

Gillihan, Seth/Farah, Martha (2005): Is self special? In: *Psychological Bulletin* 131, 76–97.

Henrich, Dieter (1966): *Fichtes ursprüngliche Einsicht*. Frankfurt a.M.

Johnson-Laird (1983): *Mental Models*. Cambridge.

Kircher, Tilo/Senior, Carl/Phillips, Mary/Benson, Philip/Bullmore, Edward/Brammer, Mick/Simmons, Andrew/Williams, Steven/Bartels, Mathias/David, Anthony (2000): Towards a functional neuroanatomy of self processing. In: *Cognitive Brain Research* 10, 133–144.

Knauff, Markus (2009): A neuro-cognitive theory of deductive relational reasoning with mental models and visual images. In: *Spatial Cognition & Computation* 9, 109–137.

Knoblich, Günther/Elsner, Birgitt/von Aschersleben, Gisa/Metzinger, Thomas (Hg.) (2003): *Self and Action*, Sonderband von *Consciousness & Cognition* (12:4).

Lenggenhager, Bigna/Tadi, Tej/Metzinger, Thomas/Blanke, Olaf (2007): Video ergo sum. In: *Science* 317, 1096–1099.

Maravita, Angelo/Iriki, Atsushi (2004): Tools for the body (schema). In: *Trends in Cognitive Sciences* 8, 79–86.

Metzinger, Thomas (Hg.) (1995): *Bewusstsein – Beiträge aus der Gegenwartsphilosophie*. Paderborn.

Metzinger, Thomas (Hg.) (2000): *Neural Correlates of Consciousness*. Cambridge, MA.

Metzinger, Thomas (2003): Phänomenale Transparenz und kognitive Selbstbezugnahme. In: Ulrike Haas-Spohn (Hg.): *Intentionalität zwischen Subjektivität und Weltbezug*. Paderborn, 411–459.

Metzinger, Thomas ([2]2004a): *Being No One* [2003] Cambridge, MA.

Metzinger, Thomas (2004b): Why are identity-disorders interesting for philosophers? In:

Thomas Schramme/Johannes Thome (Hg.): *Philosophy and Psychiatry*. Berlin, 311– 325.
Metzinger, Thomas (2006): *Being No One* – Eine sehr kurze deutsche Zusammenfassung. In: Thomas Metzinger (Hg.), *Grundkurs Philosophie des Geistes* (Bd. 1). Paderborn, 424–475.
Metzinger, Thomas (2008): Empirical perspectives from the self-model theory of subjectivity. In: *Progress in Brain Research* 168, 215–246.
Metzinger, Thomas (2009): *Der Ego Tunnel*. Berlin.
Metzinger, Thomas (2011). The no-self-alternative. In: Shaun Gallagher (Hg.): *The Oxford Handbook of the Self*. Oxford, 279–296.
Metzinger, Thomas (2013): Why are dreams interesting for philosophers? The example of minimal phenomenal selfhood, plus an agenda for future research. In: *Frontiers in Psychology*, 4. In Jennifer M. Windt (Hg.): *Contrasting Dreaming and Wakefulness: Frontiers in Consciousness Research*.
Metzinger, Thomas (2014): First-order embodiment, second-order embodiment, third-order embodiment: From spatiotemporal self-location to minimal phenomenal selfhood. In Lawrence Shapiro (Hg.): *The Routledge Handbook of Embodied Cognition*. London.
Oeing-Hanoff, Ludger/Verbeke, Gérard/Schrott, Balthasar/Nobis, Herbert/Marquard, Odo/Rothe, Klaus (1974): Geist. In: Joachim Ritter/Karlfried Gründer (Hg.): *Historisches Wörterbuch der Philosophie* (Bd. 3). Basel, 154–204.
Taylor Parker, Sue/Mitchell, Robert/Boccia, Maria (Hg.) (1994): *Self-Awareness in Animals and Humans*. Cambridge.
Trivers, Robert (2011): *The Folly of Fools*. New York, NY.
Tsakiris, Manos (2011): The sense of body ownership. In: Shaun Gallagher (Hg.): *The Oxford Handbook of the Self.* Oxford, 180–203.
Vogeley, Kai/Gallagher, Shaun (2011): Self in the Brain. In: Shaun Gallagher (Hg.): *The Oxford Handbook of the Self.* Oxford, ?111–136?.
Von Hippel, William/Trivers, Robert (2011): The evolution and psychology of self-deception. In: *Behavioral and Brain Sciences* 34, 1–56.
Weisberg, Josh (2005/6): Consciousness constrained. In: *Psyche* 11(5). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.6413&rep=rep1&type=pdf
Wegner, Daniel/Wheatley, Thalia (1999): Apparent mental causation. In: *American Psychologist* 54, 480–492.
Windt, Jennifer (2010): The immersive spatiotemporal hallucination model of dreaming. In: *Phenomenology and the Cognitive Sciences* 9, 295–316.
Windt, Jennifer (2015): *Dreaming: A Conceptual Framework for Philosophy of Mind and Empirical Research*. Cambridge, MA.
Windt, Jennifer/Metzinger, Thomas (2007): The philosophy of dreaming and self-consciousness. In: Deirdre Barrett/Patrick McNamara (Hg.): *The New Science of Dreaming*, Bd. 3. London, 193–247.

Thomas Metzinger