R. Banerjee & B.K. Chakrabarti (Eds.)
Progress in Brain Research, Vol. 168
ISSN 0079-6123

CHAPTER 18

Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples

Thomas Metzinger[1,2,*]

[1]*Philosophisches Seminar, Johannes Gutenberg Universität, D-55099 Mainz, Germany*
[2]*Frankfurt Institute for Advanced Studies, D-60438 Frankfurt am Main, Germany*

Abstract: A concise sketch of the self-model theory of subjectivity (SMT; Metzinger, 2003a), aimed at empirical researchers. Discussion of some candidate mechanisms by which self-awareness could appear in a physically realized information-processing system like the brain, using empirical examples from various scientific disciplines. The paper introduces two core-concepts, the ''phenomenal self-model'' (PSM) and the ''phenomenal model of the intentionality relation'' (PMIR), developing a representationalist analysis of the conscious self and the emergence of a first-person perspective.

Keywords: consciousness; self-consciousness; first-person perspective; ownership; agency; self-model; phenomenal transparency; phantom limbs; robotics; rubber-hand illusion; out-of-body experience; evolution of tool-use

SMT: what is the self-model theory of subjectivity?

The goal of this chapter is to give a brief summary of the ''self-model theory of subjectivity'' (SMT) that is accessible to readers who are not professional philosophers.[1] Here, I will use a series of empirical examples from a number of different disciplines to illustrate some core ideas and to demonstrate the explanatory scope as well as the predictive power of SMT. The SMT is a philosophical theory about what it means to be a self. It is also a theory about what it means to say that mental states are ''subjective'' states and that a certain system has a ''phenomenal first-person perspective.'' One of the ontological claims of this theory is that the self is not a substance in the technical philosophical sense of something that could maintain its existence on its own, even if the body, the brain, or everything else disappeared. It is not an individual entity or a mysterious *thing* in the metaphysical sense. No such things as selves exist in the world: selves and subjects are not part of the irreducible constituents of reality. What does exist is the *experience* of being a self, as well as the diverse and constantly changing contents of self-consciousness. This is what philosophers mean when they talk about the ''phenomenal self'': the

*Corresponding author. Tel.: +49-6131-39-23279;
Fax: +49-6131-39-25141; E-mail: metzinger@uni-mainz.de

[1]A short Précis, which deliberately focuses on the conceptual skeleton and ignores bottom-up constraints, is freely available in an electronic version as Metzinger (2005a), at <www.psyche.cs.monash.edu.au/>. See also the *Scholarpedia* entry on "Self Models" at <www.scholarpedia.org>. On the monograph level, an early German language (and meanwhile outdated) version of this theory can be found in Metzinger (1993); for the most comprehensive formulation of the theory to date, see Metzinger (2003a). The standard procedure to learn more about the theory is to go to Section 8.2 in Metzinger (2003a), find the questions most relevant to one's personal interests and work one's way back, using the pointers given there and the index at the end.

DOI: 10.1016/S0079-6123(07)68018-2

way you *appear* to yourself, subjectively, consciously. Under SMT, this conscious experience of being a self is analyzed as the result of complex information-processing mechanisms and representational processes in the central nervous system. Of course, there are also higher-order, conceptually mediated forms of phenomenal self-consciousness that not only have neuronal, but also *social* correlates.[2] This theory, however, begins by focusing on the minimal representational and functional properties that a naturally evolved information-processing system — such as *Homo sapiens* — has to have in order to later satisfy the constraints for realizing these higher order forms of self-consciousness. As most philosophers today would agree, the real problem lies in first understanding the simplest and most elementary form of our target phenomenon. This is the non-conceptual, prereflective and prelinguistic layer in self-consciousness. Therefore, the first question we will have to answer is this: What are minimally sufficient conditions for the emergence of a conscious self?

The self-model theory assumes that the properties in question are representational and functional brain properties. In other words, the psychological property that allows us to become a person in the first place is analyzed with the help of concepts from *subpersonal* levels of description. In philosophy of mind, this type of approach is sometimes called a "strategy of naturalization": a complex and hard-to-understand phenomenon — such as the emergence of phenomenal consciousness and a subjective, inward perspective — is conceptually analyzed in such a way as to make it empirically tractable. By reformulating classical problems from their own discipline, naturalist philosophers try to open them for interdisciplinary investigations and scientific research programs, for instance in the cognitive and neurosciences. These philosophers do not endorse naturalism and reductionism as part of a scientistic ideology; instead, they see them as a rational research strategy. For instance, if it should turn out — as many people believe (see for instance Nagel, 1986, especially Chapter 4, which is also discussed in Metzinger, 1995a) — that there is something about human self-consciousness that lies outside the reach of the natural sciences *in principle*, they would be satisfied with this finding as well. They would have achieved exactly what they set out to do in the first place: they would now have what philosophers like to call "epistemic progress." This type of progress could mean being able to describe, in a much more precise and fine-grained manner and with a historically unprecedented degree of conceptual clarity, *why exactly* science is unable to provide satisfying answers to certain questions, even in principle. Therefore, the most serious and respectable philosophical anti-naturalists will typically also be the ones who show the profoundest interest in recent empirical findings. Naturalism and reductionism are not ideologies or potential new substitutes for religion. It is exactly the anti-naturalist and exactly the anti-reductionist who will have the strongest ambition to make their philosophical case convincingly, in an empirically informed way.

[2]I analyzed the relation between conceptual and non-conceptual contents of self-consciousness in detail in Metzinger (2003c); Metzinger (2003b) is an earlier German version of this text. A hypothesis on the role of the unconscious self-model in the development of non-conceptually mediated forms of social cognition is formulated in Metzinger and Gallese (2003).

Step one: what exactly is the problem?

What we like to call "the self" in folk-psychological contexts is the phenomenal self: that aspect of self-consciousness that is immediately given in subjective experience, as the content of phenomenal experience. The phenomenal self may well be the most interesting form of phenomenal content. It endows our phenomenal space with two particularly fascinating *structural* features: centeredness and perspectivalness. As long as a phenomenal self exists, our consciousness is centered and bound to what philosophers call a "first-person perspective." States inside this center of consciousness are experienced as *my own* states, because they are endowed with a sense of ownership that is prior to language or conceptual thought. In all of my conscious experiences and actions, I engage in constantly changing relations with the environment and with my own mental

states. I experience myself as being *directed* — towards perceptual objects, other human beings, or the contents of my own mental states and concepts. This process gives rise to a subjective inner perspective. The fact that I have such an inner perspective, in turn, is cognitively available to me.[3] In other words, what probably distinguishes human beings from most other animals is that we not only *have* a subjectively experienced inner perspective, but can also consciously conceptualize ourselves *as beings that have such an inner perspective*.

The first problem, however, is that we are not exactly sure what we mean when we talk about these questions in this way. It is not just that we are unable to define concepts like "I", "self", or "subject". The real problem is that these concepts often do not seem to refer to observable objects in the world. Therefore, the first thing we have to understand is how certain structural features of our inner experience determine the way we *use* these concepts. In order to analyze the logic of ascribing psychological properties to ourselves and to understand what these concepts actually refer to, we must first investigate the representational deep structure of conscious experience itself. Three higher order phenomenal properties are particularly interesting in this context:

- "Mineness": This is a higher order property of particular forms of phenomenal content. It is an immediately given, non-conceptual sense of ownership. Here are some examples of how we try to refer to this phenomenal property in folk-psychological discourse, using everyday language: "Subjectively, *my* leg is always experienced as being a part of *me*"; "*My* thoughts and feelings are always experienced as part of *my own* consciousness"; "My volitional acts are always initiated *by myself*."
- "Selfhood": This experientially untranscendable feeling of being a self is the essence, the phenomenal core property we are looking for. Again, a few brief examples can illustrate how we refer to this highly salient feature of our inner experience from the outside, using linguistic tools: "I am *someone*"; "I experience myself as *identical* across time"; "The contents of my self-consciousness form a coherent *whole*"; "Without having the need to engage in any prior cognitive and reflexive operations I am always intimately familiar with the contents of my self-consciousness."
- "Perspectivalness": In the context discussed here, perspectivalness is the dominant structural feature of phenomenal space as a whole: it is centered in an acting and experiencing subject, a self that engages in constantly changing relationships with itself and the world. Examples include: "My world has a fixed center, and *I* am this center"; "Being conscious means having an *individual first-person perspective*"; "In experiencing persons and objects in the world as well as my own mental states, I am always bound to this inward perspective — I am its origin."

The next step consists in a representational and functional analysis of these target properties. We must ask: What functional and representational properties does an information-processing system have to have in order to instantiate the *phenomenal* property in question? Which of these properties are sufficient, and are any of them strictly necessary? What *exactly* does it mean for such a system to experience the world as well as its own mental states from a first-person perspective? What we need is a consistent conceptual background that is sufficiently flexible to continually integrate new empirical findings and at the same is capable of taking the wealth, the heterogeneity, and the subtlety of phenomenal experience into account. Obviously, this is not an easy task. I will now briefly try to sketch the outlines of such a conceptual framework in the remaining five steps.

[3]For a first introduction to the problem of cognitive self-reference as a potential difficulty for philosophical naturalism, see Baker (1998). See also Metzinger (2003a) (Section 6.4.4) and especially Metzinger (2003c). An interesting and lucid criticism of my own account of the cognitive first-person perspective is Baker (2007).

Step two: the self-model

Step two consists in the introduction of a new theoretical entity: the phenomenal self-model (PSM). It is the most important part of the representational basis for instantiating the relevant phenomenal properties (Cummins, 1983). What is a mental ''representation''? A representational state, for instance in the brain, is a state that has a certain *content*, because it is directed at something in the world. The brain-state is the physical carrier; the content is the meaning of this state. An inner representation is *about* something: having a correct representation implies *reference*. A representational state often functions as a placeholder for something external, the referent; it represents because it ''stands in'' for something else. However, this ''something'' can also be a past event, a potential future outcome, or even a mere possibility — in such cases, we speak of representations as *simulations*. They simulate merely *possible* states of affairs; they represent a possibility, not an actuality. SMT is predominantly a representational theory of consciousness, because it analyzes conscious states as representational states and conscious contents as representational contents.

One of our key questions was: Which set of minimally sufficient *representational* properties does a system have to develop in order to possess the relevant target properties? This is our first, preliminary answer: the system needs a coherent self-representation, a consistent internal model of itself as a whole. In our case, the self-model is an episodically active representational entity whose content is determined by the system's very own properties. Whenever such a self-representation is needed to regulate the system's interactions with the environment, it is transiently activated — for instance in the morning, when we wake up. According to SMT, what happens when you wake up in the morning — when you first *come to yourself* — is that the biological organism, which you are, boots up its PSM: it activates the conscious self-model.

In other words, what we need is a comprehensive theory of the self-model of *Homo sapiens*.[4] Personally, I assume that this will be a predominately neurocomputational theory (see for instance, Churchland, 1989). This means that there is not only a true representational and functional description of the human self-model, but also a true neurobiological description — for instance in terms of being a widely distributed, complex activation pattern in the brain (Damasio, 1999). The PSM is exactly that part of the *mental* self-model that is currently embedded in a highest order integrated structure, the global model of the world (Yates, 1975; Baars, 1988; for a detailed analysis of the criteria for distinguishing different degrees of consciousness, see Metzinger, 2003a, Chapter 3). In other words, certain parts of the self-model can be unconscious and functionally active at the same time. The PSM is a coherent multimodal structure that probably depends on a partially innate, ''hard-wired'' model of the system's spatial properties. (More about this in the second example; see also the fifth section of O'Shaughnessy, 1995 and his use of the concept of a ''*long-term body image*''; and Metzinger, 1993, 1996, 1997; Damasio, 1994, 1999). This type of analysis treats the self-conscious human being as a special type of information-processing system: the subjectively experienced content of the phenomenal self is the representational content of a currently active, dynamic data structure in the system's central nervous system.

Aside from the representational level of description, one can also develop a *functional* analysis of the self-model. Whereas representational states are individuated by their content, a functional state is conceptually characterized by its *causal role*: the causal relationships it bears to input states, output states, and other internal states. An active self-model then can be seen as a subpersonal functional state: a set of causal relations of varying complexity that may or may not be realized at a

[4]The methodological core of psychology — insofar as I may venture this type of metatheoretical observation from my standpoint as a philosophical outsider — can now be analyzed in a fresh and fruitful way. Psychology is *self-model research.* It is the scientific discipline that focuses on the representational content, the functional profile and the neurobiological realization of the human self-model, including its evolutionary history and its necessary social correlates.

given point in time. Since this functional state is realized by a concrete neurobiological state, it plays a certain causal role for the system. For instance, it can be an element in an information-processing account. The perspective of classic cognitive science can help illustrate this point: the self-model is a *transient computational module* that is episodically activated by the system in order to control its interactions with the environment. In other words, what happens when you wake up in the morning, i.e., when the system that you are "comes to itself," is that this transient computational module is activated — the moment of "waking up" is exactly the moment in which this new instrument of intelligent information-processing emerges in your brain. It does so because you now need a conscious self-model in order to achieve sensorimotor integration, generate complex, flexible and adaptive behavior, and attend to and control your body *as a whole*. The development of ever more efficient self-models as a new form of "virtual organ" — and this point should not be overlooked — is also a precondition for the emergence of complex societies. Plastic and ever more complex self-models not only allowed somatosensory, perceptual, and cognitive functions to be continuously optimized, but also made the development of social cognition and cooperative behavior possible. The most prominent example, of course, is the human mirror system, a part of our unconscious self-model that *resonates* with the self-models of other agents in the environment through a complex process of motor-emulation — of "embodied simulation," as Vittorio Gallese (2005) aptly puts it — e.g., whenever we observe goal-directed behavior in our environment. Such mutually coupled self-models, in turn, are the fundamental representational resource for taking another person's perspective, for empathy and the sense of responsibility, but also for metacognitive achievements like the development of a *concept* of self and a *theory of mind* (see for instance, Bischof-Köhler, 1996, 1989; on the possible neurobiological correlates of these basic social skills, which fit very well into the framework sketched above, see Gallese and Goldman, 1998; Metzinger and Gallese, 2003).

The obvious fact that the development of our self-model has a long biological, evolutionary, and (a somewhat shorter) social history can now be accounted for by introducing a *teleofunctionalist background assumption*, as it is often called in philosophy of mind (see for instance Millikan, 1984, 1993; Bieri, 1987; Dennett, 1987; Dretske, 1988, 1998; Lycan, 1996). The development and activation of this computational module plays a role *for* the system: the functional self-model possesses a true evolutionary description, i.e., it was a weapon that was invented and continuously optimized in the course of a "cognitive arms race" (Clark, 1989, p. 61). The functional basis for instantiating the phenomenal first-person perspective can be seen as a specific cognitive achievement: the ability to use a *centered* representational space. In other words, phenomenal subjectivity (the development of a subsymbolic, non-conceptual first-person perspective) is a property that is only instantiated when the respective system activates a coherent self-model and integrates it into its global world-model.

The existence of a stable self-model allows for the development of what philosophers call the "perspectivalness of consciousness": the existence of a single, coherent, and temporally stable reality-model that is representationally centered in a single, coherent, and temporally stable phenomenal subject, a model of the system *in the act of experiencing* (see last section). This structural feature of the global representational space then leads to the episodic instantiation of a temporally extended, non-conceptual first-person perspective. If this global representational property is lost, this also changes the phenomenology and leads to the emergence of different neuropsychological deficits or altered states of consciousness. Some readers may have the impression that all of this is extremely abstract. A self-model, however, is not at all abstract — it is entirely concrete. A first, now classic, example will help demonstrate what — among many other things — I actually mean with the concept of a "self-model."

In a series of fascinating experiments, in which he used mirrors to induce synesthesia and kinesthetic illusions in phantom limbs, Indian neuropsychologist Vilayanur Ramachandran

demonstrated the existence of the human self-model (see Ramachandran and Rogers-Ramachandran, 1996; a popular account can be found in Ramachandran and Blakeslee, 1998, 46ff. The figure was published courtesy of Ramachandran). Phantom limbs are subjectively experienced limbs that typically appear after the accidental loss of an arm or a hand or after surgical amputation. In some cases, for instance following a non-traumatic amputation performed by a surgeon, patients have the subjective impression of being able to control and move their phantom limb at will. The neurofunctional correlate of this phenomenal configuration could consist in the fact that motor commands, which are generated in the motor cortex, continue to be monitored by parts of the parietal lobe and — since there is no contradictory feedback from the amputated limb — are subsequently integrated into the part of the self-model that serves as a *motor emulator* (related ideas are discussed by Grush 1997, 1998, p. 174; see also Ramachandran and Rogers-Ramachandran, 1996, p. 378). In other cases, the subjective experience of being able to move and control the phantom limb is lost. These alternative configurations may result from preamputational paralysis following peripheral nerve damage or from prolonged loss of proprioceptive and kinesthetic "feedback" that could confirm the occurrence of movement. On the phenomenological level of description, this may result in a paralyzed phantom limb.

Ramachandran and colleagues constructed a "virtual reality box" by vertically inserting a mirror in a cardboard box from which the lid had been removed. The patient, who had been suffering from a paralyzed phantom limb for many years, was then told to insert both his real arm and his phantom into two holes that had been cut in the front side of the box. Next, the patient was asked to observe his healthy hand in the mirror. On the level of visual input, this generated the illusion of seeing both hands, even though he was actually only seeing the reflection of his healthy hand in the mirror. So, what happened to the content of the PSM when the patient was asked to execute symmetrical hand movements on both sides? This is how Ramachandran describes the typical outcome of the experiment

> I asked Philip to place his right hand on the right side of the mirror in the box and imagine that his left hand (the phantom) was on the left side. "I want you to move your right and left arm simultaneously," I instructed.
>
> "Oh, I can't do that," said Philip. "I can move my right arm but my left arm is frozen. Every morning, when I get up, I try to move my phantom because it's in this funny position and I feel that moving it might help relieve the pain." But, he said looking down at his invisible arm, "I never have been able to generate a flicker of movement in it."
>
> "Okay, Philip, but try anyway."
>
> Philip rotated his body, shifting his shoulder, to "insert" his lifeless phantom into the box. Then he put his right hand on the other side of the mirror and attempted to make synchronous movements. As he gazed into the mirror, he gasped and then cried out, "Oh, my God! Oh, my God, doctor! This is unbelievable. It's mind-boggling!" He was jumping up and down like a kid. "My left arm is plugged in again. It's as if I'm in the past. All these memories from years ago are flooding back into my mind. I can move my arm again. I can feel my elbow moving, my wrist moving. It's all moving again."
>
> After he calmed down a little I said, "Okay, Philip, now close your eyes."
>
> "Oh, my," he said, clearly disappointed. "It's frozen again. I feel my right hand moving, but there's no movement in the phantom."

Fig. 1. Mirror-induced synesthesia. Making part of a hallucinated self available for conscious action control by installing a virtual source of visual feedback. (Picture courtesy of Vilayanur Ramachandran.) (See Color Plate 18.1 in color plate section.)

> "Open your eyes."
>
> (See Ramachandran 1998, 47f. For the clinical and experimental details, see Ramachandran and Rogers-Ramachandran, 1996) (Fig. 1).

By now, it should be clear how these experimental findings illustrate the concept of a "self-model" that I introduced above; what is moving in this experiment *is* the PSM. What made the sudden occurrence of kinesthetic movement sensations in the lost subregion of the self-model possible was the installation of an additional source of feedback, of "virtual information." This immediately created a new functional property, let us call it "availability for selective motor control." By providing access to the visual mode of self-simulation, this made the corresponding information available to volition as well. Now, volitional control once again was possible. This experiment also shows how phenomenal properties are determined by computational and representational properties. Bodily self-consciousness is directly related to brain processes.

Let us directly move on to the next example, while staying with the phenomenology of phantom limbs. How "ghostly" are phantom limbs? Can we measure the "realness" of the conscious self? A recent case study by Brugger and colleagues introduced a vividness rating on a 7-point scale that showed highly consistent judgments across sessions for their subject AZ, a 44-year-old university-educated woman born without forearms and legs. For as long as she remembers, she has experienced mental images of forearms (including fingers) and legs (with feet and first and fifth toes) — but, as the figure below shows, these were not *as* realistic as the content of her non-hallucinatory PSM. Functional magnetic resonance imaging of phantom hand movements showed no activation of the primary sensorimotor areas, but of the premotor and parietal cortex bilaterally. Transcranial magnetic stimulation (TMS) of the sensorimotor cortex consistently elicited phantom sensations in the contralateral fingers and hand. In addition, premotor and parietal stimulation evoked similar phantom sensations, albeit in the absence of motor-evoked potentials in the stump. These data clearly demonstrate how body parts that were never physically developed can be phenomenally simulated in sensory and motor cortical areas. Are they components of an innate body model? Or could they have been "mirrored

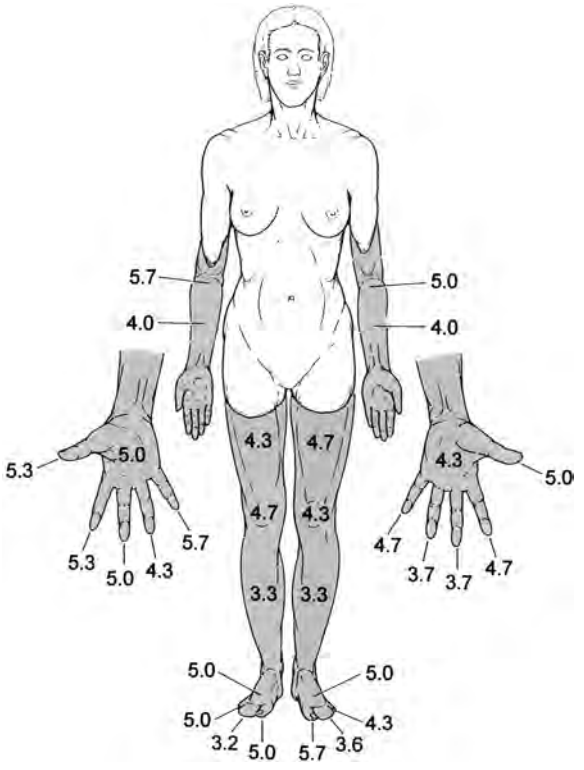

Fig. 2. Evidence for an innate component of the PSM? Phantoms (shaded areas) in a subject with limb amelia. The numbers are vividness ratings for the felt presence of different phantom body parts on a 7-point scale from 0 (no awareness) to 6 (most vivid impression). (Picture courtesy of Peter Brugger, Zürich.)

into'' the patient's self-model through the visual observation of other human beings moving around? As I am a philosopher and not a neuropsychologist, I will refrain from further amateurish speculation at this point (Fig. 2).

However, recent results from research on pain experiences in phantom limbs point to the potential existence of a genetically determined neuromatrix whose activation pattern may form the basis of these rigid parts of the self-model and the more invariant background of bodily self-experience (the ''phylomatrix of the body schema''; see Melzack, 1989; on the concept of a ''neurosignature,'' see Melzack, 1992, p. 93; an important study on phantom limbs following aplasia and early amputation is Melzack et al., 1997). Another interesting empirical result is that more than 20% of children born without an arm or a leg later develop the realistic conscious experience of having a phantom limb. In the context of phenomenal ''realness'' and in terms of the integration of the bodily self-model into the brain's conscious reality model as a whole it may also be interesting to note that, in this case, ''Awareness of her phantom limbs is transiently disrupted only when some object or person invades their felt position or when she sees herself in a mirror.'' (Brugger et al., 2000, p. 6168. For further details concerning the phenomenological profile see ibid; for an interesting experimental follow-up study demonstrating the intactness of the phenomenal model of kinesthetic and postural limb properties, see Brugger et al., 2001).

What do the phenomenologies of Ramachandran's and Brugger's subjects have in common? The transition from stump to phantom limb is *seamless*; subjectively, they are both part of one and the same bodily self, because the quality of ownership is distributed evenly among them. There is no gap or sudden jump in the sense of ownership. The emergence of the bodily self-model is based on a subpersonal, automatic process of *binding features* together, of achieving coherence. But what exactly is it that is being experienced? What is the *content* of experience? Aristotle said that the soul is the *form* of the physical body, which perishes together with it at death (*On The Soul*, II: 412a, 412b–413a). According to Spinoza, the soul is the *idea* that the body develops of itself (*The Ethics*, II: 12 and 13). In more modern terms, we might say that an ''idea'' is simply a mental representation — more precisely a *self*-representation — and that the content of self-consciousness is the introspectively accessible part of this self-representation, namely the PSM postulated by the self-model theory. *Gestalt* properties — like body shape — are *global* properties of an object, and could the self-model then not be a neural mechanism to represent exactly such global properties, a new tool to acquire knowledge about the organism *as a whole*? Plato, however, claimed that some ideas are innate. And this still is an interesting question for today's neuroscience of self-consciousness as well: Does the PSM possess an innate component? Is the conscious body image a

kind of "fixed idea," anchored in an inborn and genetically predetermined *nucleus*?

Let us now turn to example no. 3. It comes from a different scientific discipline altogether, namely from the fascinating new field of evolutionary robotics. It demonstrates a number of further aspects that the conceptual framework of SMT, the self-model theory, predicts and seeks to explain. First, a self-model can be entirely *unconscious*; i.e., it can frequently be seen as the product of an automatic "bottom-up" process of *dynamical self-organization*; second, it is not a "thing" (or a model of a thing) at all, but based on a continuous, ongoing modeling *process*; third, it can exhibit considerable *plasticity* (i.e., it can be modified through learning); and fourth, in its origins it is not based on language or conceptual thought, but very likely on an attempt to organize motor behavior. It is a computational tool to achieve global control. More precisely, a body-model has the function of integrating sensory impressions with motor output in a more intelligent and flexible manner. The unconscious precursor of the PSM clearly was a new form of intelligence.

Bongard et al. (2006) have created an artificial "starfish" that gradually develops an explicit internal self-model. Their four-legged machine uses actuation–sensation relationships to indirectly infer its own structure and then uses this self-model to generate forward locomotion. When part of its leg is removed, it adapts its self-model and generates alternative gaits — it learns to limp. In other words unlike the phantom-limb patients presented in example no. 1 and no. 2 (and like most ordinary patients), it is able to *restructure* its body-representation following the loss of a limb. It can learn. This concept may not only help develop more robust machines and shed light on self-modeling in animals, but is also theoretically interesting, because it demonstrates for the first time that a physical system has the ability, as the authors put it, to "autonomously recover its own topology with little prior knowledge" by constantly optimizing the parameters of its own resulting self-model (Fig. 3a–c).

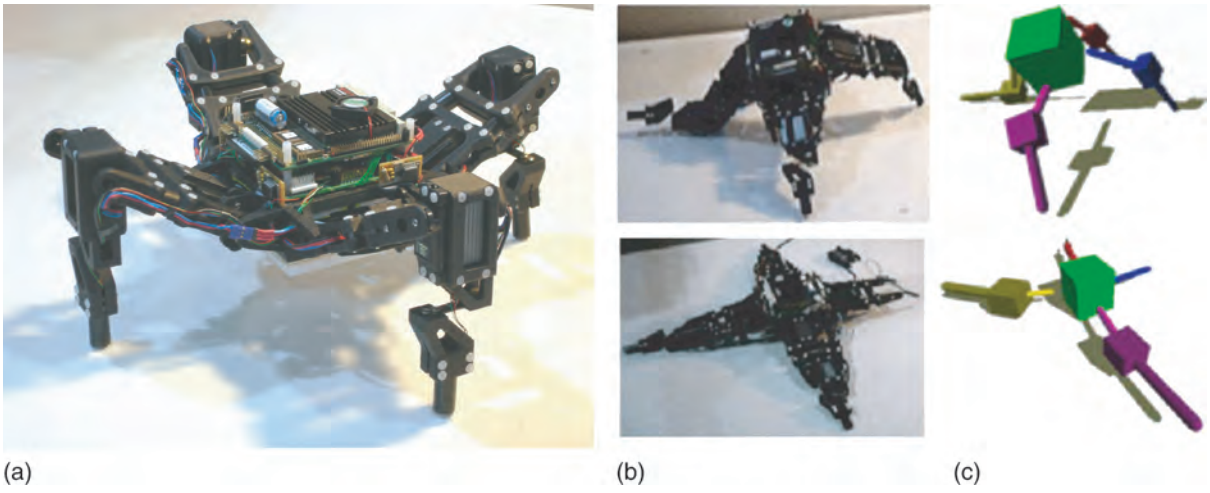

Fig. 3. (a) Starfish, a four-legged physical robot that has eight motorized joints, eight joint angle sensors, and two tilt sensors. (See www.ccsl.mae.cornell.edu/research/selfmodels/morepictures.htm for additional online material.) (b and c) The starfish-robot walks by using an explicit internal self-model that it has autonomously developed and that it continuously optimizes. If he loses a limb, he can adapt his internal self-model. (d) The robot continuously cycles through action execution. (a and b) Self-model synthesis. The robot physically performs an action (a). Initially, this action is random; later, it is the best action found in (c). The robot then generates several self-models to match sensor data collected while performing previous actions (b). It does not know which model is correct. (c) Exploratory action synthesis. The robot generates several possible actions that disambiguate competing self-models. (d) Target behavior synthesis. After several cycles of (a)–(c), the best current model is used to generate locomotion sequences through optimization. (d) The best locomotion sequence is executed by the physical device. (e) (See Color Plate 18.3 in color plate section.)
Figure 3 continued on p. 224.

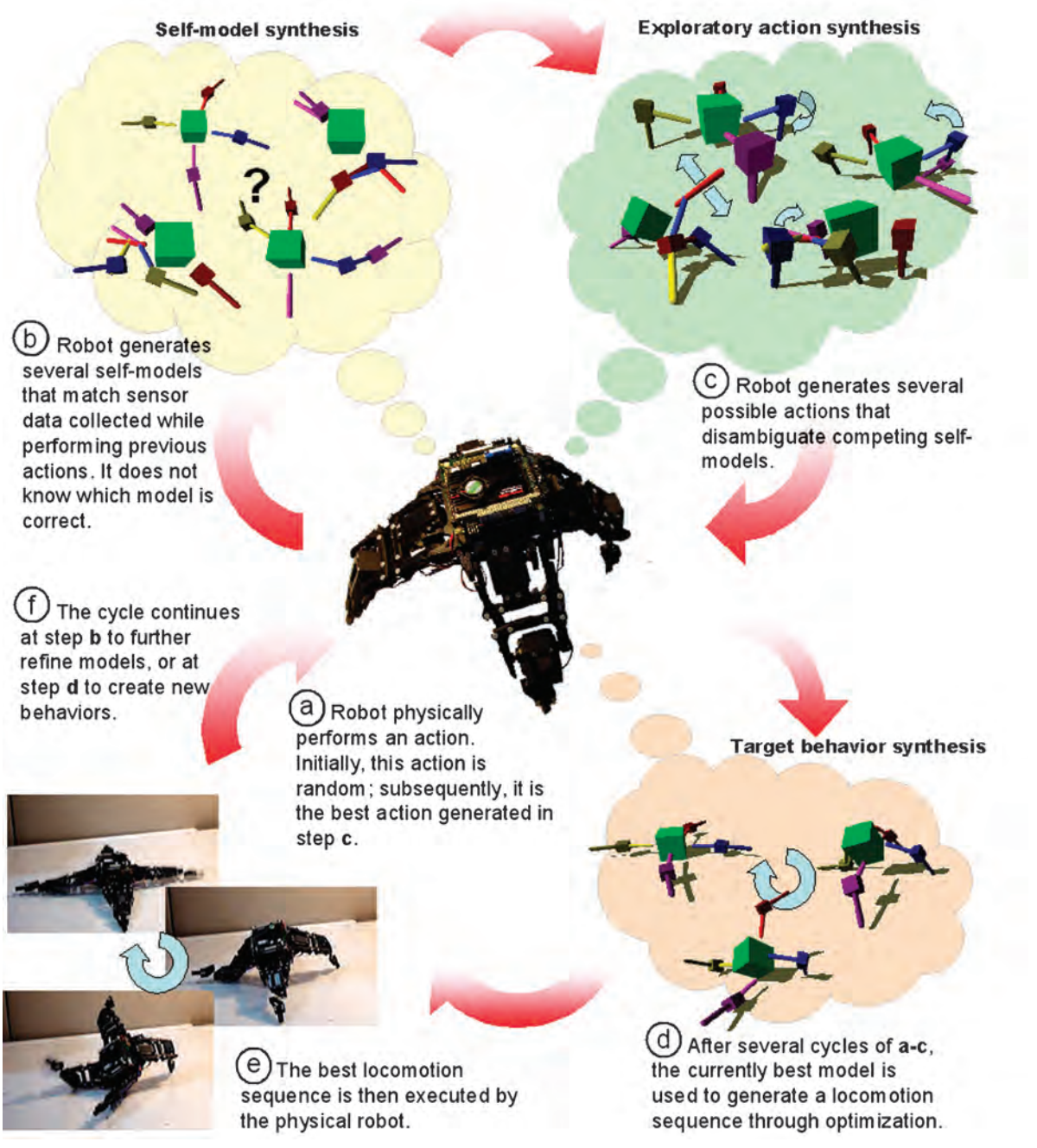

Fig. 3. (*Continued*)

Starfish not only synthesizes an internal self-model, but also uses this self-model to generate intelligent behavior. The next figure gives an overview over this process (Fig. 3d).

As we see, the robot initially performs an arbitrary motor action and records the resulting sensory data. The model synthesis component then synthesizes a set of 15 candidate self-models using stochastic optimization to explain the observed sensory–actuation relationship. The robot then synthesizes an exploratory motor action that causes maximum *disagreement* among the different predictions of these competing self-models. This action is physically carried out, and the 15 candidate self-models are subsequently improved using the new data. When the models converge, the most accurate model is used by the behavior synthesis component to create a desired behavior that can then be executed by the robot. If the robot detects unexpected sensor–motor patterns or an external signal resulting from unanticipated morphological change, it reinitiates the alternating cycle of modeling and exploratory actions to produce new models reflecting this change. The most accurate of these new models is then used to generate compensatory behavior and recover functionality.

Technical details aside — what are the philosophical consequences of example no. 3? First, you do not have to be a living being in order to have a self-model. Non-biological SMT-systems are possible. Second, a self-model can be entirely unconscious, i.e., it does not have to be a PSM. Awareness obviously is a second step (see Metzinger, 1995b, 2000a, for a first overview; Metzinger, 2003a, Section 3.2, for an additional set of ten constraints to be satisfied for conscious experience). Third, a self-model supports planning and fast learning processes in a number of different ways. It clearly makes a system more intelligent. Fourth, it is what I called a virtual model or "virtual organ" above, and one of its major functions consists in appropriating a body by using a global morphological model to control it as a whole. Elsewhere, I have introduced the term "second-order embodiment" for this type of self-control (Metzinger, 2006b). If I may use a metaphor: one of the core ideas is that a self-model allows a physical system to "enslave" its low-level dynamics with the help of a single, integrated, and internal whole-system model, thereby controlling and functionally "owning" it. This is the decisive first step towards becoming an autonomous agent.

Step three: a representationalist analysis of the three target properties

Here, the basic idea is that self-consciousness, first of all, is an *integrative* process: by becoming embedded in the currently active self-model, representational states acquire the higher order property of phenomenal mineness. If this integrative process is disturbed, this results in various neuropsychological syndromes or altered states of consciousness (for case studies, see Chapter 7 in Metzinger, 2003a). Let us take a look at some examples of what happens when phenomenal mineness, the subjective sense of ownership, is selectively lost.

- Florid schizophrenia: Consciously experienced thoughts are no longer *my* thoughts.
- Somatoparaphrenia, unilateral hemi-neglect: My leg is no longer *my* leg.
- Depersonalization, delusions of control: I am a robot, I am turning into a puppet, and volitional acts are no longer *my* volitional acts. (In this case, what philosopher and psychiatrist Karl Jaspers called *Vollzugsbewusstsein*, or "executive consciousness," is selectively lost.)
- Manic disorders: I am the whole world; all events in the world are controlled by *my own* volitional acts.

Subjectively experienced "mineness" is a property of discrete forms of phenomenal content, such as the mental representation of a leg, a thought, or a volitional act. This property, the sense of ownership, is not necessarily connected to these mental representations; i.e., it is not an intrinsic, but a *relational* property. That a thought or a body part is consciously experienced as your own is not an essential, strictly necessary property of the conscious experience of this thought or body part.

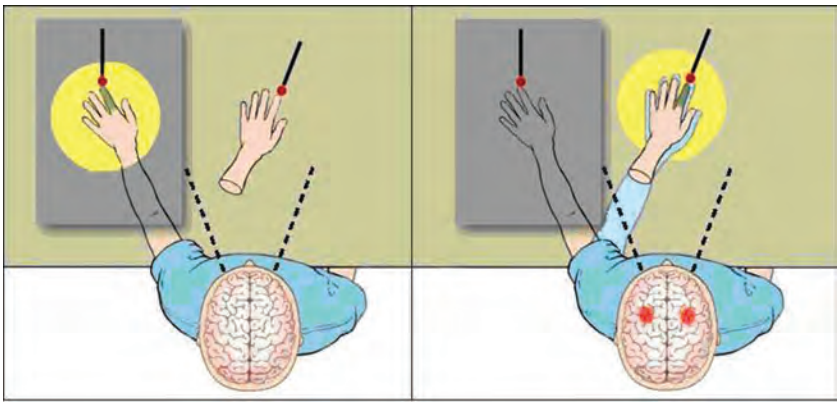

Fig. 4. The rubber-hand illusion. A healthy subject experiences an artificial limb as part of her own body. The subject observes a facsimile of a human hand while one of her own hands is concealed (gray square). Both the artificial rubber hand and the invisible hand are then stroked repeatedly and synchronously with a probe. The yellow and green areas indicate the respective tactile and visual receptive fields for neurons in the premotor cortex. The illustration on the right shows the subject's illusion as the felt strokes (green) are brought into alignment with the seen strokes of the probe (areas of heightened activity in the brain are colored red; the phenomenally experienced, illusory position of the arm is indicated by the blue area). The respective activation of neurons in the premotor cortex is demonstrated by experimental data. (Figure by Litwak illustrations studio 2004.) (See Color Plate 18.4 in color plate section.)

It could have been otherwise, in other phenomenological contexts, mineness disappears. Its distribution over the different elements of a conscious world-model can vary. If the system is no longer able to integrate certain discrete representational contents into its self-model, it is lost. If this analysis is correct, it should be possible, at least in principle, to operationalize this property by searching for an empirically testable metrics for the coherence of the self-model in the respective areas of interest. One could also empirically investigate *how* and in which brain areas a certain type of representational content is integrated into the self-model. Here is a concrete example for what I mean by "mineness," example no. 4 (Fig. 4).

In the rubber-hand illusion (RHI), the sensation of being stroked with a probe is integrated with the corresponding visual perception in such a way that the brain transiently matches a proprioceptive map (of the subject's own-body perception) with a visual map (of what the subject is currently seeing). At the same time, the feeling of "ownership" or phenomenal "mineness" is transferred to the rubber hand. The subject experiences the rubber hand as her *own* hand and feels the strokes *in* this hand. When asked to point to her concealed left hand, her arm movement will automatically swerve in the direction of the rubber hand (Botvinick and Cohen, 1998, p. 756). If one of the fingers of the rubber-hand is "hurt" by being bent backwards into a physiologically impossible position, the subject will also experience her real phenomenal finger as being bent much farther backwards than it is in reality. At the same time, this will also result in a clearly measurable skin conductance response. While only 2 out of 120 subjects reported an actual pain sensation, many subjects drew back their real hands, opened their eyes up widely in surprise, or laughed nervously (Armel and Ramachandran, 2003, p. 1503). Subjects also showed a noticeable reaction when the rubber hand was hit with a hammer. Again, it becomes clear how the phenomenal target property is directly determined by representational and functional brain processes. What we experience as part of our self depends on the respective context and on which information our brain integrates into our currently active self-model (see especially Botvinick and Cohen, 1998, and the neuroimaging study by Botvinick, 2004; Ehrsson et al., 2004). The intriguing question, of course, is this: Could *whole-body* illusions exist as well? The answer is

yes, and we will soon return to this point in example no. 5.

But first, let us take a look at the second target property, at consciously experienced selfhood. Methodologically, it is important to first isolate the simplest form of the target. Phenomenal selfhood corresponds to the existence of a single, coherent, and temporally stable self-model that constitutes the center of the representational state as whole. If this representational module is damaged or disintegrates, or if multiple structures of this type alternate or are simultaneously activated by the system, this will again result in various neuropsychological disturbances or altered states of consciousness

- Ansognosia and anosodiaphoria: Loss of higher order insight into existing deficits, e.g., in cortically blind patients who deny that they are blind (Anton's Syndrome).
- Dissociative Identity Disorder (DID): The system uses different and alternating self-models as a means of coping with extremely traumatic and socially inconsistent situations (for the current diagnostic criteria for DID, see DSM-IV: 300.14).
- *Ich-Störungen*, or identity disorders: A large class of psychiatric disturbances connected to altered forms of experiencing one's own *identity*. Schizophrenia is a classical example, as are Cotard syndrome, reduplicative paramnesia, or delusional misidentifaction (for a discussion on why identity disorders are interesting from a philosophical perspective, see Metzinger, 2004a).

The existence of a stable self-model also almost always gives rise to the ''perspectivalness of consciousness'' in terms of transient subject–object relationships (see step 6 below; see also Nagel, 1986; Metzinger, 1993, 1995a, 2005a, and especially Metzinger, 2006a). This structural feature of the global representational space leads to the episodic instantiation of a temporally extended and non-conceptual first-person perspective. It, too, can be lost.

- Complete depersonalization: Loss of the phenomenal first-person perspective, accompanied by dysphoric states and functional deficits (''dreadful ego-dissolution''; see Dittrich, 1985).
- Mystical experiences: Selfless and non-centered global states, which are experienced viz. described as non-pathological and unthreatening (''oceanic boundary loss,'' ''*The Great View from Nowhere*'').

In order to do justice to the wealth and the diversity of different forms of human experience, one has to acknowledge the existence of certain non-perspectival and selfless forms of conscious experience. Phenomenologically, *non-subjective* consciousness — phenomenal experience that is not tied to a self or an individual first-person perspective — is not only a possibility, but a reality, even if we may find this idea inconceivable. The self-model theory provides the conceptual means to account for these special cases (for additional neurophenomenological case studies, see Metzinger, 2003a, Chapters 4 and 7).

Example no. 5 will demonstrate this principle in another domain. If we have the necessary conceptual instruments, we can not only take the subtleties and the variability of human experience seriously. We can also develop new interdisciplinary research programs that penetrate into ''taboo zones'' and shed light on phenomena that in the past were only the targets of esoteric folklore and metaphysical ideologies. Could there be an integrated kind of bodily self-consciousness, be it of a mobile body fully available for volitional control or of a paralyzed body that in its entirety is a phenomenal confabulation — in short, a *hallucinated* and a *bodily* self at the same time? Is it conceivable that something like a full-body analog of the rubber-hand-illusion or a ''globalized phantom-limb experience'' — the experience of a *phantom body* — could emerge in a human subject? The answer is, yes. There is a well-known class of phenomenal states in which the experiencing person undergoes the untranscendable and highly realistic conscious experience of leaving his or her physical body, usually in the form of an etheric double, and moving around outside of it. In other

words, there is a class (or at least a strong cluster) of intimately related phenomenal models of reality that are classically characterized and defined by a *visual representation* of one's own body from a perceptually impossible, externalized third-person perspective (e.g., seeing oneself from above, lying on the bed, or on the road) plus a *second representation* of one's own body, typically (but not in all cases) freely hovering or floating in space. This second body-model is the locus of the phenomenal self. It not only forms the "true" focus of one's phenomenal experience, but also functions as an integrated representation of all kinesthetic qualia and all non-visual forms of proprioception. This class of phenomenal states is called the "Out-of-body experience" (OBE). Elsewhere (Metzinger, 2005b, for further references see also Lenggenhager et al., 2007), I have argued that our traditional, folk-phenomenological concept of a "soul" may have its origins in accurate and sincere first-person reports about the experiential content of this specific neurophenomenological state-class.

OBEs frequently occur spontaneously while falling asleep, but also following severe accidents or during surgical operations. At present, it is not clear whether the concept of an OBE possesses a clearly delineated set of necessary and sufficient conditions. Instead, the concept of an OBE may turn out to be a cluster concept constituted by a whole range of diverging (and possibly overlapping) subsets of phenomenological constraints, each forming a set of sufficient, but not necessary, conditions. On the other hand, the OBE clearly is something like a phenomenological *prototype*. There is a common core to the phenomenon, as can be seen from the simple fact that many readers will already have heard about this type of experience in one way or another.

One can offer a representationalist analysis of OBEs by describing them as a class of deviant self-modeling processes. On the level of conscious self-representation, a prototypical feature of this class of deviant PSM seems to be the coexistence of (a) a more or less veridical representation of the bodily self as seen from an external visual perspective, which does *not*, however, function as the center of the global model of reality, and (b) a second self-model, which according to subjective experience largely integrates proprioceptive perceptions — although, interestingly, weight sensations are only integrated to a lesser degree — and possesses special properties of shape and form that may or may not be veridical. Both models of the experiencing system are located within the same spatial frame of reference (that is why they are *out-of*-body-experiences). This frame of reference is an *egocentric* frame of reference. Let us now look at two classical phenomenological descriptions of OBEs, as spontaneously occurring in an ordinary non-pathological context

> I awoke at night — it must have been at about 3 a.m. — and realized that I was completely unable to move. I was absolutely certain I was not dreaming, as I was enjoying full consciousness. Filled with fear about my current condition, I had only one goal, namely to be able to move my body again. I concentrated all my will-power and tried to roll over to one side: Something rolled, but not my body — something that was me, my whole consciousness including all of its sensations. I rolled onto the floor beside the bed. While this happened, I did not feel bodiless, but as if my body consisted of a substance in between the gaseous and the liquid state. To the present day, I have never forgotten the combination of amazement and great surprise that gripped me when I felt myself falling onto the floor, but without the expected thud. Had the movement actually unfolded in my normal physical body, my head would have had to collide with the edge of my bedside table. Lying on the floor, I was overcome by terrible fear and panic. I knew that I possessed a body, and I only had one great desire — to be able to control it again. With a sudden jolt, I regained control, without knowing how I managed to get back into it. (Waelti, 1983, p. 25; English translation TM)

The prevalence of OBEs ranges from 10% in the general population to 25% in students, with extremely high incidences in certain subpopulations like, to name just one example, 42% in schizophrenics (Blackmore, 1986; see also Blackmore, 1982; for an overview and further references see Alvarado, 1986, 2000, p. 18; Irwin, 1985, p. 174). However, it would be false to assume that OBEs typically occur in people suffering from severe psychiatric disorders or neurological deficits. Quite the contrary, most OBE-reports come from ordinary people in everyday life situations. Let us therefore stay with non-pathological situations and look at another paradigmatic example, again reported by Swiss biochemist Ernst Waelti

> I went to bed in a dazed state at 11 p.m. and tried to go to sleep. I was restless and turned over frequently, causing my wife to grumble briefly. Now, I forced myself to lie in bed motionless. For a while I dozed before feeling the need to pull up my hands, which were lying on the blanket, in order to bring them into a more comfortable position. At the same instant, I realized that I was absolutely unable to move and that my body was lying there in some kind of paralysis. Nevertheless, I was able to pull my hands out of my physical hands, as if the latter were just a stiff pair of gloves. The process of detachment started at the fingertips, in a way that could be clearly felt, almost with a perceptible sound, a kind of crackling. It was exactly the movement that I had actually intended to carry out with my physical hands. With this movement, I detached from my body and floated out of it head first. I moved into an upright position, as if I was almost weightless. Nevertheless, I had a body consisting of real limbs. You have certainly seen how elegantly a jellyfish moves through water. I could now move around with the same ease. I lay down horizontally in the air and floated across the bed, like a swimmer, who has pushed himself from the edge of a swimming pool. A delightful feeling of liberation arose within me. But soon, I was seized by the ancient fear common to all living creatures, the fear of losing my physical body. It sufficed to drive me back into my body. (Waelti, 1983, p. 25; English translation TM) (Figs. 5 and 6)

Sleep paralysis is not a necessary precondition for OBEs. They frequently occur during extreme sports, for instance, in high-altitude climbers or marathon runners.

> A Scottish woman wrote that, when she was 32 years old, she had an OBE while training for a marathon. "After running approximately 12–13 miles … I started to feel as if I wasn't looking through my eyes but from somewhere else. … I felt as if something was leaving my body, and although I was still running along looking at the scenery, I was looking at myself running as well. My 'soul' or whatever, was floating somewhere above my body high enough up to see the tops of the trees and the small hills." (Alvarado, 2000, p. 184)

The classic OBE contains two self-models, one visually represented from an external perspective and one forming the center of the phenomenal world from which the first-person perspective originates. What makes the representationalist and functionalist analysis of OBEs difficult and at the same time challenging is the fact that many *related* phenomena exist, e.g., autoscopic phenomena during epileptic seizures in which only the first criterion is fulfilled (for a neurological categorization see Brugger et al., 1997). Devinsky et al. (1989, p. 1080) have differentiated between autoscopy in the form of a complex hallucinatory perception of one's own body as being external with "the subject's consciousness … usually perceived within his body" and a second type, the classic OBE, which includes the feeling of leaving one's body and viewing it from another vantage-point. The incidence of autoscopic seizures is

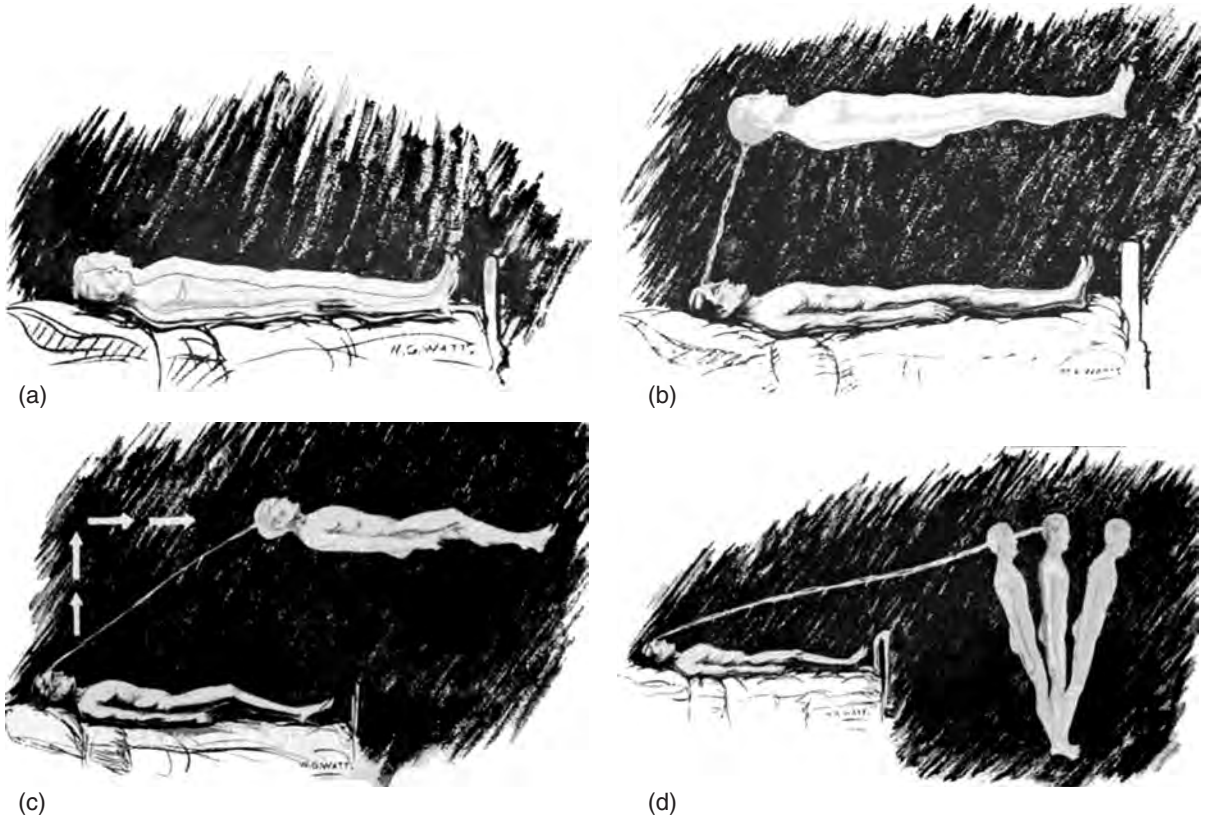

Fig. 5. Kinematics of the PSM during OBE-onset: The classical Muldoon-scheme. From Muldoon S. and Carrington, H. (1929). The Projection of the Astral Body. Rider & Co., London.

Fig. 6. Kinematics of the phenomenal body-image during OBE onset. An alternative, but equally characteristic motion pattern, as described by Swiss biochemist Ernst Waelti (1983).

possibly higher than previously recognized, and the authors found a 6.3% incidence in their patient population (Devinsky et al., 1989, p. 1085). Seizures involving no motor symptoms or loss of consciousness, which may not be recognized by the patient, may actually be more frequent than commonly thought for a case study of a patient who first experienced OBEs for a number of years and only later suffered from generalized seizures (see Vuilleumier et al., 1997, p. 116).

What function could this type of experience have *for* the organism as a whole? Here is a speculative proposal by Devinsky and colleagues

> There are several possible benefits that dissociative phenomena, such as autoscopy, may confer. For example, when a prey is likely to be caught by its predator, feigning death may be of survival value. Also, accounts from

> survivors of near-death experiences in combat or mountaineering suggest that the mental clarity associated with dissociation may allow subjects to perform remarkable rescue manoeuvres that might not otherwise be possible. Therefore, dissociation may be a neural mechanism that allows one to remain calm in the midst of near-death trauma. (Devinsky et al., 1989, p. 1088)

It is not at all inconceivable that there are physically or emotionally stressful situations in which an information-processing system is forced to introduce a ''representational division of labour'' by distributing different representational functions into two or more distinct self-models (for instance in what in the past was called ''multiple personality disorder,'' see Metzinger, 2003a, Section 7.2.4). The OBE may be an instance of transient functional modularization, of a ''purposeful,'' i.e., functionally adequate, separation of levels of representational content in the PSM. For instance, if the system is cut off from somatosensory input or flooded with stressful signals and information threatening the overall integrity of the self-model as such, it may be advantageous to integrate the ongoing conscious representation of higher cognitive functions like attention, conceptual thought, and volitional selection processes into a *separate* model of the self. This may allow for a high degree of integrated processing, i.e., of ''mental clarity,'' by functionally encapsulating and thereby *modularizing* different functions like proprioception, attention, and cognition in order to preserve at least some of these functions in a life-threatening situation. Almost all necessary system-related information is still globally available, and higher order processes like attention and cognition can still operate on this information as it continues to be presented in an integrated manner, but its distribution across specific subregions of phenomenal space as a whole changes dramatically. Only one of the two self-models is truly ''situated'' in the overall scene; only one of them is immediately embodied and virtually self-present in the sense of being integrated into an internally simulated behavioral space.

It has long been known that OBEs not only occur in healthy subjects, but in certain clinical populations (e.g., epileptics) as well. In a recent study, Olaf Blanke and colleagues were able to localize the relevant brain lesion or dysfunction in the temporo–parietal junction (TPJ) in five out of six patients. It was also possible, for the first time, to induce an OBE-type state by direct electrical stimulation. These researchers argue that two separate pathological conditions may be necessary to cause an OBE. First, a disintegration in the self-model or ''personal space'' (brought about by a failure to integrate proprioceptive, tactile, and visual information regarding one's own body) plus an additional, second disintegration between external, ''extrapersonal'' visual space, and the internal frame of reference created by vestibular information. The experience of seeing one's own body in a position that does not coincide with its felt position could therefore be caused by cerebral dysfunction at the TPJ, causing both types of functional disintegration and thereby leading to the representational configuration described above (Figs. 7 and 8).

Using evoked potential mapping, these authors also showed that a selective activation of the TPJ takes place 330–400 ms after healthy volunteers mentally imagined themselves being in a position and taking a visual perspective characteristic of an OBE. At the same time, it is possible to impair this mental transformation of the bodily self-model by interfering at this specific location with TMS. In an epileptic patient with OBEs caused by damage at the TPJ, it could be shown that by mimicking the OBE-PSM (i.e., by mentally simulating an OBE like the ones she had experienced before), there was a partial activation of the seizure focus (Blanke et al., 2005). Therefore, there exists an anatomical bridge overlap between these three very similar types of phenomenal mental content.

What is most needed at the current stage is an experimental design that makes OBEs a controllable and repeatable phenomenon in healthy subjects, under laboratory conditions. Achieving this interim goal would be of high relevance, not only from an empirical, but also from a philosophical perspective. Studying the functional fine structure of embodiment by developing a convincing representationalist analysis of phenomenal

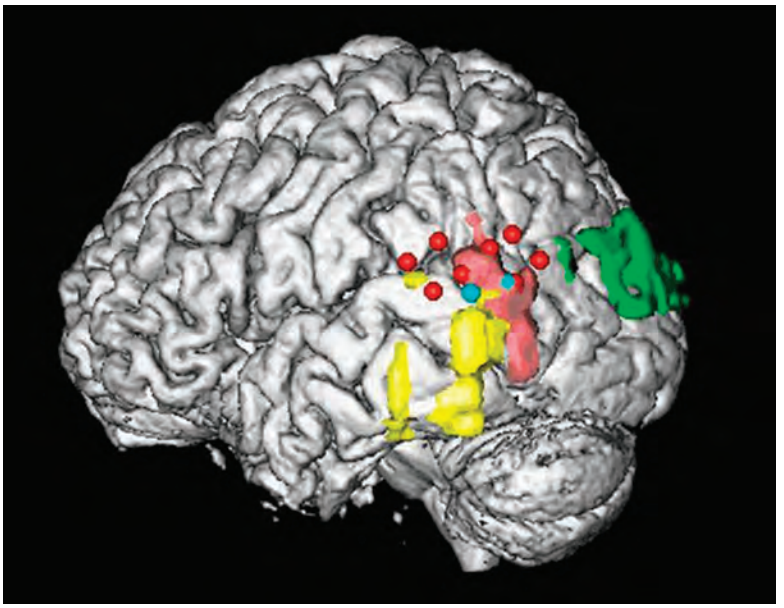

Fig. 7. The figure shows results of a mean lesion overlap analysis in five patients with OBEs. The analysis centers on the TPJ. The blue dots show the locus of electrical cortical stimulation in the patient in whom an OBE-like phenomenal state was artificially induced. (Figure courtesy of Olaf Blanke, cf. Blanke et al., 2004.) (See Color Plate 18.7 in color plate section.)

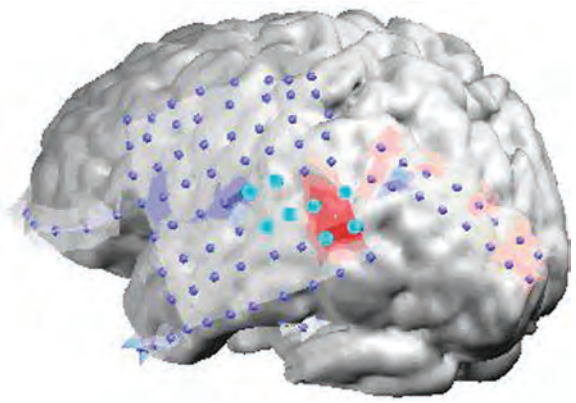

Fig. 8. The next figure shows another set of patient data, in which MRI was performed with implanted electrodes in the left hemisphere. The epileptic focus, where the discharge induced an OBE, is indicated by eight turquoise electrodes at the TPJ. (Figure courtesy of Olaf Blanke, cf. Blanke et al., 2005.) (See Color Plate 18.8 in color plate section.)

disembodiment would certainly shed new light on the issue of non-conceptual self-awareness and the origin of a conscious first-person perspective. In particular, it would be of high theoretical relevance if one could empirically demonstrate the possibility of minimal selfhood *without an agency component*. Let me therefore give you a brief example of my own recent research. Example no. 5 is a study based on interdisciplinary cooperation between neuroscience and philosophy of mind and, specifically, on an experimental design originally developed from philosophical considerations (for details see Lenggenhager et al., 2007).

The classical RHI (example no. 4) only tells us something about the target property of "ownership" (for body parts), but not about "selfhood" (ownership for the *whole* body). To manipulate attribution and localization of the *entire* body and to study selfhood per se we designed an experiment based on clinical data in neurological patients with out-of-body experiences. These data suggest that the spatial unity between self and body may be disrupted leading in some cases to the striking experience that the conscious self is localized at an extracorporeal position. Therefore, the aim of the present experiments was to induce out-of-body experiences in healthy participants in order to investigate the phenomenal target

property of selfhood. We hypothesized that under adequate experimental conditions participants would experience a visually presented body as if it was their own, inducing a drift of the subjectively experienced bodily self to a position outside one's bodily borders. Can one create a whole-body analog of the RHI, an illusion during which healthy participants experience a virtual body as if it were their own and localize their self outside their body boundaries at a different position in space?

We applied virtual reality to examine the possible induction of out-of-body experiences by using multisensory conflict. In the first experiment participants viewed the back of their body filmed from a distance of 2 m and projected onto a 3D-video head-mounted display (HMD; see Fig. 9). The participants' back was stroked during 1 min either synchronously or asynchronously with respect to the virtually seen body. Global self-attribution of the virtual character was measured by a questionnaire that was adapted from the RHI. Global self-localization was measured by passively displacing the blindfolded participants immediately after the stroking and asking them to return to their initial position (Fig. 9).

While being stroked, the subjects were either shown their own back (''own body condition''), the back of a mannequin (''fake body condition''), or an object (''object condition'') being stroked and projected directly (synchronously) or with a time lag (asynchronously) onto a HMD. After being stroked, the subjects were passively displaced and then asked to return to their initial position and fill out a modified ''rubber-hand-questionnaire.'' Results of the questionnaire showed that for the synchronous ''own body'' and ''fake body'' conditions, subjects often felt as if the observed virtual figure were their own body. This impression was less likely to occur in the ''object condition'' and in all of the asynchronous conditions. The synchronous experimental conditions also showed a significantly larger shift towards the projected real or fake body than the asynchronous and control conditions. These data suggest that self-location — due to conflicting visual-somatosensory input — is as prone to misidentification and mislocalization as was previously reported for body parts, as in the RHI.

Illusory self-localization to a position outside one's body shows that bodily self-consciousness and selfhood can be dissociated from an accurate representation of one's physical body position. This differs from the RHI where the aspect of selfhood remained constant and only the attribution and localization of the stimulated hand was manipulated. Does illusory self-localization to a position outside one's body mean that we have experimentally induced full-blown out-of-body experiences? No, this was only a first step. But it is quite clear what the next steps will have to be. Out-of-body experiences are characterized by disembodiment of the self to an extracorporeal location, an extracorporeal visuo-spatial perspective, and seeing of one's own body from this extracorporeal self-location. As the present illusion was neither associated with overt disembodiment nor with a change in visuo-spatial perspective, we argue that we have induced only some aspects of out-of-body experiences or rather the closely related experience of heautoscopy that has also been observed in neurological patients (see original publication for further references).

To give just one example, I believe that an additional necessary condition involved in generating full-blown out-of-body experiences and the complete transfer of selfhood to the illusory body is a transient episode of visual-vestibular disintegration. At least two spatial frames of reference must be functionally dissociated, in order to not only have a ''teleportation-OBE,'' but a realistic exit phenomenology, a gradual motion path through phenomenal space. This general principle should hold for our experimental setup as well as for OBEs in epileptic patients or ''gifted subjects'' in the healthy population. Why is this principle relevant from a theoretical perspective, and why is it difficult to test experimentally? In standard situations, and as opposed to all other conscious model of aspects of reality, the human PSM is anchored in the brain through a continuous flow of self-generated input. There exists a persistent causal link into the physical body itself. In order to understand the SMT better, we must turn to this point now — it explains why our conscious model of reality is a *centered* model of reality.

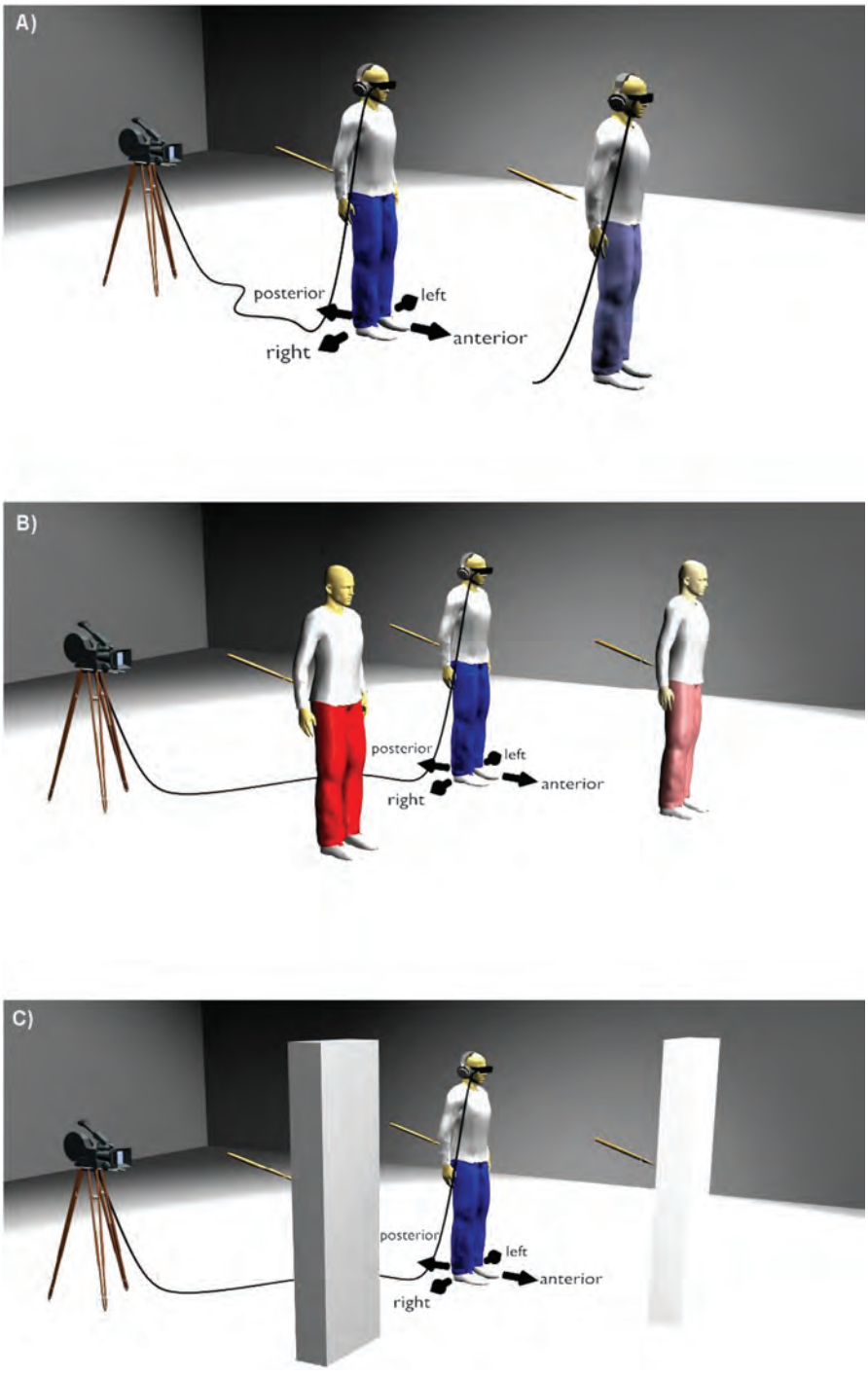

Fig. 9. (A) Participant (in dark blue trousers) sees through a HMD his own virtual body (light blue trousers) in 3D, standing 2 m in front of him and being stroked synchronously or asynchronously at the participant's back. In other conditions (Study II) the participant sees either (B) a virtual fake body (light red trousers) or (C) a virtual non-corporeal object (light gray) being stroked synchronously or asynchronously at the back. Dark colors indicate the actual location of the physical body/object, whereas light colors represent the virtual body/object seen on the HMD. Illustration by Martin Boyer, published in Science. (See Color Plate 18.9 in color plate section.)

Step four: the bodily self as a functional anchor of phenomenal space

Above, I drew attention to the distinction between the representational and the functional analysis of the first-person perspective. The central theoretical problem on the functional level of description can be summed up by the following question: What exactly is the difference between the PSM and the other phenomenal models that are currently active in the system? Is there a characteristic causal mark of the PSM? Which *functional property* is responsible for turning it into the stable center of phenomenal representational space?

This is my first, preliminary, answer. The self-model is the only representational structure that is anchored in a *continuous source of internally generated input* in the brain. Let us call this the "persistent causal link hypothesis." Whenever conscious experience arises (i.e., whenever a stable, integrated model of reality is activated), this continuous source of internal proprioceptive input also exists. The human self-model possesses an enduring causal link in the brain. It has parts, which in turn are realized by *permanent* forms of information processing on *permanent* forms of self-generated input and low-level autoregulation. To put this general point differently, the body, in certain of it aspects, is the only perceptual object from which the brain can never run away. Again, I will not enter into any amateurish empirical speculation here, but offer a number of obvious candidates for sources of high invariance. Basically, there are four different types of internally generated information that during conscious episodes, constitute a persistent functional link between the PSM and its bodily basis in the brain

- Inputs from the vestibular organ: the sense of balance.
- Inputs from the autonomously active, invariant part of the body schema: the continuous "background feeling" in the spatial model of the body, which is independent of external input, e.g., via motion perception.
- Inputs from the visceral sensors, but also from the blood vessels, for instance from the cardiovascular mechanosensors: "gut feelings" and somatovisceral forms of self-presentation.
- Inputs from certain parts of the upper brain stem and hypothalamus: background emotions and moods, which are anchored in the continuous homeostatic self-regulation of the "internal milieu," the biochemical landscape in our blood.

Philosophically, it is not as much the neurobiological details that are crucial as the highly plausible assumption that there is a certain part of the human self-model that is characterized by a high degree of stimulus correlation and that depends exclusively on *internally* generated information. This layer of the PSM is directly and permanently anchored in stimuli from the inside of the body. Do you still remember patient AZ from example no. 2? The weaker degree of phenomenological "vividness" or "realness" in her phantom limbs may reflect exactly the absence of permanent bottom-up stimulation that in normal situations is caused by existing physical limbs. In this context, Marcel Kinsbourne has spoken of a "*background 'buzz' of somatosensory input*" (Kinsbourne, 1995, p. 217). To capture the phenomenology involved in this sheer "raw feel of embodiment" on the representationalist level of description I like to distinguish between self-presentation and self-representation.[5] Phenomenologically, the first concept is related to the purely sensory feeling of bodily presence, which so interestingly goes along with a subjective sense of temporal immediacy and the experiential certainty of possessing direct, non-inferential self-knowledge. What exactly is this deepest layer

[5]For an extensive theoretical treatment of the subject and numerous recent empirical results on the body as an anchor of conscious experience, see Damasio (1999). Antonio Damasio uses the term of a *core self*, and elsewhere (Metzinger, 1993, p. 156ff; Metzinger, 2003a, Section 5.4) I introduced the technical concept of "phenomenal self-presentation" (as opposed to self-*re*presentation). On the level of body-representation, self-presentation is what AZ lacks in her phantom limbs, whereas self-*re*presentation is what she actually has — although, as the referent of this representation never existed, this obviously is also a form of *mis*representation.

of the phenomenal self? Why is it the origin of the first-person perspective? My hypothesis is that the constant self-organizing activity of those regions of the bodily self that are independent of external input constitutes the functional *center* of phenomenal representational space.

As our first example of how to understand the concept of a self-model, we used the experiment in which Ramachandran managed to mobilize a paralyzed phantom limb. A self-*presentation* is exactly that part of the phantom limb that remains conscious independently of the occurrence of movement. If *this* part is lost, you also lose the subjective experience of bodily presence — you turn into a "disembodied being."[6] But there may even be other, more general empirical perspectives from which the self-model is necessarily related to the baseline of brain activity per se, as it can be observed in the resting state (see Raichle et al., 2001; Gusnard, 2005).

Step five: autoepistemic closure — transparency and the naïve-realistic self-misunderstanding

Back on the *representational* level of analysis, the central theoretical problem is that one might easily accuse me of mislabeling the actual problem by introducing the concept of a "self-model." First, a self-model, of course, is not a model of a mysterious *thing* that we then call the self. It is a continuous and self-*directed* process tracking global properties of the organism. Second, at least according to certain modal intuitions, there appears to be no necessary connection between the fundamental functional and representational properties on the one hand and the *phenomenal* target properties of "mineness," "prereflexive/preagentive selfhood," and "perspectivalness" on the other hand. All this could easily occur without resulting in a real phenomenal self or a subjective inner perspective; it is conceivable that biological information-processing systems could develop and successfully employ a representational space centered by a self-model *without* also developing self-consciousness. More interestingly, even *given the phenomenal level*, i.e., even in a system that is already conscious, it is not obvious or self-evident that the specific phenomenology of *selfhood* should emerge. What would, by logical necessity, bring about an ego? A "self-model" is by no means a self, but only a representation of the system as a whole — it is no more than a *system-model*. If the functional property of centeredness and the representational property of having a self-model are to lead to the phenomenal property of perspectivalness, the conscious system-model must turn into a phenomenal self. The decisive philosophical question is this: How does the existence of a functionally centered representational space necessarily lead to the emergence of a conscious self and what we commonly call a phenomenal first-person perspective? In other words, how does the system-model turn into a *self*-model?

My answer is that a genuinely conscious self emerges at exactly the moment when the system is no longer able to recognize the self-model it is currently generating *as* a model on the level of conscious experience. So, how does one get from the functional property of "centeredness" and the representational property of "self-modeling" to the phenomenal target property of "prereflexive self-intimacy"? The solution has to do with what philosophers call "phenomenal transparency" (for a short explanation of the concept of "phenomenal transparency," see Metzinger, 2003c; Metzinger, 2003b is the German precursor). The conscious representational states generated by the system are *transparent*, i.e., they no longer represent the very fact that they *are* models on the level of their content. Consequently — and this is a phenomenological metaphor only — the system simply looks right "through" its very own representational structures, as if it were in direct and immediate contact with their content. Please note how this is only a statement about the system's *phenomenology*. It is not a statement about epistemology, about the possession of knowledge: you can be completely deluded and have no or very little knowledge about reality (or your own mind) and

[6]Again, the corresponding phenomenological state classes exist. In Metzinger (1993) and Metzinger (1997), I discussed Oliver Sacks' example of the "disembodied lady". In this context, see also the famous case of Ian Waterman, which is discussed in Metzinger (2003a).

at the same time enjoy the phenomenology of certainty, of knowing that you know. Phenomenal transparency is not *epistemic* transparency, or Descartes' classical — and now falsified — idea that we can not be wrong about the contents of our own mind. Transparency, as defined in this context, is exclusively a property of *conscious* states. Unconscious states are neither transparent nor opaque. Phenomenal transparency also is not directly related to the second technical concept in philosophy, to "referential transparency." Non-linguistic creatures incapable of conceptual thought can have phenomenally transparent states as well. Naïve realism is not a belief or an intellectual attitude, but a feature of phenomenal experience itself.

I have two causal hypotheses about the micro-functional underpinnings and the evolutionary history of transparent phenomenal states. First, in a very small time-window, the neural data structures in question are activated so quickly and reliably that the system is no longer able to recognize them as such, for instance due to the comparatively slow temporal resolution of *metarepresentational* functions. Introspectively, the construction process is invisible. Second, in a much larger explanatory time-window, there apparently was no evolutionary pressure on the respective parts of our functional architecture in the process of natural selection. For biological systems like us, naïve realism was a functionally adequate background assumption. We needed to know "Careful, there is a wolf nearby!" but not "A wolf-representation is active in my brain right now!"

Transparency is a special form of darkness. It is a lack of knowledge. Epistemologically speaking, it is an implicit, not an explicit lack of knowledge. As Franz Brentano ([1874] 1973, 165f) and Daniel Dennett (1991, 359) pointed out, the representation of absence is not the same thing as the absence of representation. In transparent states, there is no representation of earlier processing stages. In the phenomenology of visual awareness, it means not being able to see something. Phenomenal transparency *in general*, however, means that the representational character of the contents of conscious experience itself is not accessible to subjective experience. This analysis can be applied to all of the sensory modalities, especially to the integrated phenomenal model of the world as a whole. Because the very *means* of representation cannot be represented as such, the experiencing system necessarily becomes entangled in naïve realism; it experiences itself as being directly in contact with the contents of its own conscious experience. It is unable to experience the fact that all of its experiences take place in a *medium* — and this is exactly what we mean by the "immediacy" of phenomenal consciousness. In a completely transparent representation, the very mechanisms that lead to its activation as well as the fact that its contents depend on a concrete inner state as a carrier can no longer be recognized by way of introspection. As philosophers like to say: "Only content properties are introspectively accessible, vehicle properties are inaccessible." Therefore, the phenomenology of transparency is the phenomenology of naïve realism.

Many phenomenal representations are transparent because their content and its very existence appear to be fixed in all possible contexts. According to subjective experience, the book you are currently holding in your hands will always stay the same book — no matter how the external perceptual conditions vary. You never have the experience that an "active object emulator" in your brain is currently being integrated into your global reality-model. You simply experience the *content* of the underlying representational process: the *book* as effortlessly given, here and now. The best way to understand the concept of transparency is to distinguish between the vehicle and the content of a representation, between representational carrier and representational content (see also Dretske, 1998, p. 45ff).

The representational carrier of your conscious experience is a particular brain process. This process — that itself is in no way "book-like" — is not consciously experienced; it is transparent in the sense that phenomenologically, you look right through it. What you look *at* is its representational content, the perceptually mediated existence of a book, here and now. In other words, this content is an abstract property of a concrete representational state in your brain. If the representational carrier is a good and reliable instrument for the generation

of knowledge, its transparency allows you to "look right through" it out into the world, at the book in your hands. It makes the information it carries globally available without your having to worry about *how* this actually happens. What is special about most phenomenal representations is that you experience their content as maximally *concrete* and unequivocal, as directly and immediately given even when the object in question — the book in your hands — does not really exist at all, but is only a hallucination. Phenomenal representations appear to be exactly that set of representations for which we cannot distinguish between representational content and representational carrier on the level of subjective experience.

Of course, there are counterexamples, and they may help further illustrate the concept of "transparency." For instance, *opaque* phenomenal representations arise when the information *that* their content is the result of an internal representational process suddenly becomes globally available. If you suddenly discover that the book in your hands does not really exist, the hallucination turns into a pseudohallucination. The information that you are not looking at the world, but rather "at" an active representational state that apparently is not functioning as a reliable instrument for the generation of knowledge at this moment, now also becomes available, and it does so on the level of subjective experience itself. The phenomenal book state becomes opaque. You lose *sensory* transparency. You become aware of the fact that your perceptions are generated by your sensory system and that this system is not always completely reliable. Not only do you now suddenly experience the book as a representation, you also experience it as a *mis*representation.

Let us further assume that you suddenly discover that not only your perception of the book, but all of your philosophical thoughts about the problem of consciousness are taking place in a dream. Then, this dream would turn into a lucid dream (for a discussion of the reasons for regarding lucid dreams as a philosophically relevant class of conscious states, see Metzinger, 2003a, Section 7.2.5; more on the topic can be found in Windt and Metzinger, 2007). The fact that you are currently not experiencing a world, but only a *world-model* would become globally available; now, you could use this information to control your actions, thoughts, and the direction of attention. You would lose *global* transparency. The interesting point, however, is that cognitive availability alone is not sufficient to dissolve the naïve realism of phenomenal experience. You cannot simply "think" yourself out of your phenomenal model of reality by changing your opinions about this model: the transparency of phenomenal representations is cognitively impenetrable; here, phenomenal knowledge is not the same as conceptual/propositional knowledge.

Now, the final step is to apply this insight to the self-model. Here is my key claim— we are systems that are experientially unable to recognize our own subsymbolic self-model *as* a self-model. For this reason, phenomenologically, we operate under the conditions of a "naïve-realistic self-misunderstanding"; we experience ourselves as being in direct and immediate epistemic contact with ourselves. By logical necessity, a phenomenally transparent self-model will create the experience of *being infinitely close to yourself*. The core of the self-model theory is that this is how the basic sense of selfhood arises and how a phenomenal self that is untranscendable for the respective system comes about. The content of non-conceptual self-consciousness is the content of a transparent PSM. It also commits me to a specific prediction: Were the PSM to loose its transparency and become opaque, were the organism as a whole capable of recognizing its current self-model *as* a model, then the phenomenal property of selfhood would disappear. In standard phenomenological configurations, however, the entity that looks at the book in its hands is itself a form of transparent phenomenal content. And this is also true of the "at"-ness inherent in this act of visual attention, of the relation that seems to connect subject and object.

Step six: the PMIR — the phenomenal model of the intentionality relation

Let us take one more step before we close. The experience of selfhood is intimately related not only to the sense of ownership, but also to the

experience of agency; it is not only a question of having a transparent self-model, but also of directedness, of being dynamically related to target objects and goal states. Here are two further examples, this time from yet another academic discipline — experimental neuroscience using macaque monkeys as subjects.

Classical neurology hypothesized about a "body schema," an unconscious, but constantly updated map of body shape and posture in the brain.[7] Recent research shows how Japanese macaque monkeys can be trained to use tools even though they only rarely exhibit tool-use in their natural environment (see Maravita and Iriki, 2004, for a good review). During successful tool-use, changes in specific neural networks in their brains take place — a finding that suggests that the tools are temporarily integrated into their body schema. When a food pellet is dispensed beyond the reach of their hands and they skillfully use a rake to pull it closer, one can observe a change in their bodily self-model in the brain. In fact, it looks as if their conscious model of their hand had been expanded towards the tip of the tool. A more precise way of describing what happens is to say that, on the level of the monkey's conscious model of reality, properties of the hand are now transferred to the distant tip of the tool. We know the same effect in human beings. In our own case, repeated practice can turn the tip of a tool into a part of our own hand, a part that can be used just as "sensitively" and as skillfully as our own fingers.

In other words, recent neuroscientific data clearly support the view that tools not only enable us to extend our reaching space. They show that any successful extension of behavioral space is also mirrored in the neural substrate of the body image in the brain. The brain constructs an "internalized" image of the tool by swiftly assimilating it into the existing body image as a whole. Of course, we do not know if monkeys actually have the conscious experience of ownership or only the unconscious mechanism. But we do know about several similarities between macaques and humans that make this assumption seem plausible. This may be the very beginning of mentally *simulating* yourself as currently being directed at a target object or goal state. And this leads us to second major aspect of selfhood: besides global *ownership* what we need to understand is *agency* — global control.

One exciting aspect of these new data is that they shed light on the evolution of tool-use. A necessary precondition of expanding your space of action and your capabilities by using tools clearly seems to be the ability to integrate them into a preexisting self-model. You can only engage in goal-directed and intelligent tool-use if your brain temporarily represents them as part of your own self. Intelligent tool-use was a major achievement in human evolution. One may plausibly assume that some elementary building block of human tool-use abilities already existed in the brains of our ancestors. Then, due to some not-yet-understood evolutionary pressure, it rapidly expanded into what we see in humans today.[8]

There is a new, rapidly growing field of research in which engineers and neuroscientists work together: brain–machine interfaces (for a brief overview, see Lebedev and Nicolelis, 2006). One application of this general idea consists in driving and controlling artificial limbs or robotic manipulators with the help of ensembles of cortical neurons, allowing a machine to carry out motor commands generated in the brain. The following figure shows an example from the Duke University Center for Neuroengineering, demonstrating the general principle (Fig. 10).

In our context, the perhaps most interesting observation in this experiment (see Carmena et al., 2003, for details) is how the monkey gradually begins to neglect his original arm, which is, after all, a part of his biological body. That is, as he now tries to control feedback in a new kind of motor task and with a different goal-state, optimizing a

[7]The terminology was never entirely clear, but it frequently differentiated between an unconscious "body schema" and a conscious "body image." For a philosophical perspective on the conceptual confusion surrounding both notions, see Gallagher (2005); for an excellent review of the empirical literature, see Maravita (2006).

[8]See Iriki et al. (1996); Maravita and Iriki (2004).

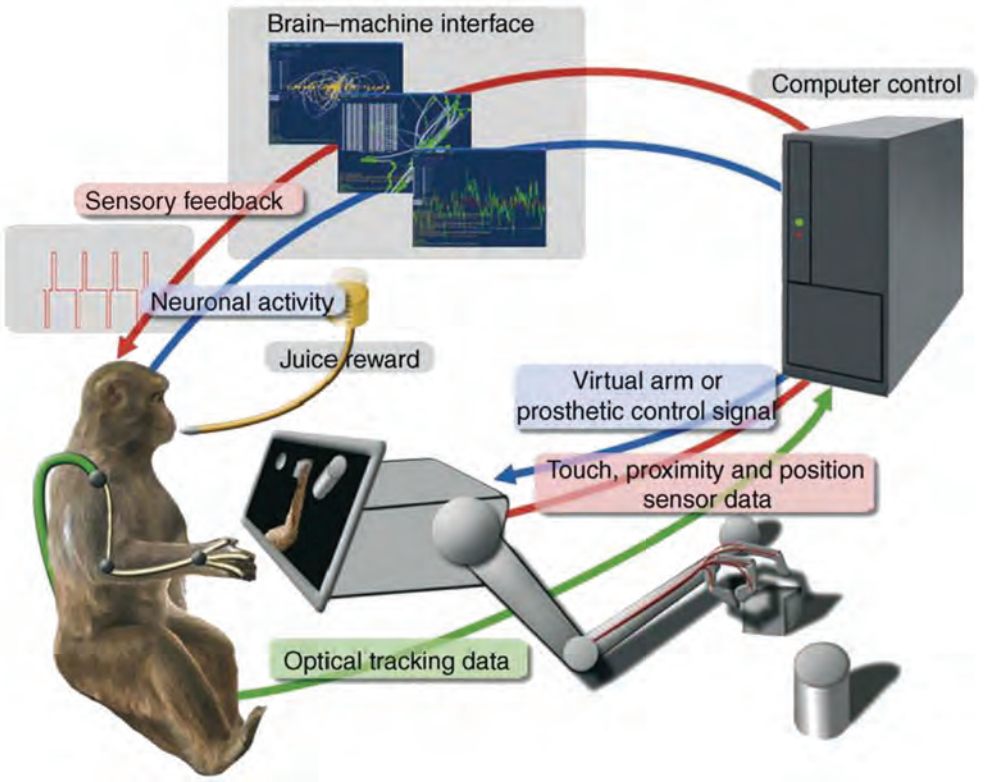

Fig. 10. A BMI with multiple feedback loops that is currently being developed at the Duke University Center for Neuroengineering. A rhesus macaque operates an artificial robotic manipulator that reaches for and grasps different objects. The manipulator is equipped with touch, proximity, and position sensors. Signals from the sensors are delivered to the control computer (right), which processes them and converts them to microstimulation pulses delivered to the sensory areas in the monkey's brain, providing it with feedback information (red loop). A series of microstimulation pulses is illustrated in the inset on the left. Neuronal activity is recorded in multiple brain areas and translated into commands to the actuator via the control computer and multiple decoding algorithms (blue loop). The arm position is monitored using an optical tracking system that tracks the position of several markers mounted on the arm (green loop). The hypothesis is that the continuous operation of this interface would lead to the incorporation of the external actuator into the representation of the body in the brain. (Figure designed by Nathan Fitzsimmons.) (See Color Plate 18.10 in color plate section.)

new set of motor parameters by trying to control a real-world robot arm or even a virtual arm he sees on the screen in front of him, his brain seems to undergo certain changes — the "tuning properties" of neurons change. Here is how Lebedev and Nicolelis (2006, p. 542) describe the effect: "Remarkably, after these animals started to control the actuator directly using their neuronal activity, their limbs eventually stopped moving, while the animals continued to control the actuator by generating proper modulations of their cortical neurons. The most parsimonious interpretation of this finding is that the brain was capable of undergoing a gradual assimilation of the actuator within the same maps that represented the body."

From the perspective of SMT, the self-model theory, the most plausible interpretation is that, once the monkey has successfully embedded an internal representation of this new actuator into his conscious self-model, the representations of his old body parts lose certain functional properties, they transiently becomes less and less available for attentional processing and gradually recede from conscious experience. These examples teach us two further important insights. Very obviously, the PSM is an important part of a *control hierarchy*; it is a means to monitor certain critical aspects of the

process by which the organism generates flexible, adaptive patterns of behavior; second, it is highly plastic in the sense that multiple representations of objects *outside* the body can transiently be integrated into it. This is not only true of rubber-hands, but even more so of tools in the most general sense — extensions of bodily organs which must be successfully controlled in order to generate intelligent, goal-directed behavior. The self-model is the functional window through which the brain can interact with the body as a whole, and vice versa. If the body is augmented by sticks, stones, rakes, or robot arms, the self-model has to be extended. If an integrated representation of body-plus-tool is in existence, the extended system of body-plus-tool can become part of the brain's control hierarchy. After all, how *could* one learn to intelligently use a tool *without* integrating it into the conscious self? The conscious self-model is a virtual organ that allows us to *own* feedback loops, to initiate, sustain, and flexibly adapt control processes. Some elements of the control loop are physical (such as the brain and tools); others are virtual (such as the self-model and goal-state simulation).

In passing, let me briefly emphasize one further point. In human beings (and some other animals as well), it is frequently the behavior or mental state of *another* person that is to be controlled. We "instrumentalize" and "appropriate" each other. Human beings constantly augment themselves not only with sticks, stones, rakes, or robot arms — but also with the brains and bodies of *other* human beings (Metzinger and Gallese, 2003). Clearly, the transition form biological to cultural evolution is intimately connected with the appearance of new and specific functional properties in the primate-PSM. This is one of the most interesting questions for the future: What exactly was the change in the PSM of Homo, as opposed to the PSM of the chimpanzee, which lead to the explosion of culture and the emergence of complex societies? Here, my own speculative working hypothesis would be that it was not complex tool use *per se*, but the ability to take a much larger part of the control hierarchy *offline*, to use it in simulation, while at the same time generating an opaque (i.e., a non-transparent) PSM. It was the ability to consciously represent yourself *as* representing, *as* being directed at a goal state. It was the difference between having a first-person perspective and the mental capacity to explicitly represent this very fact.

Now, let us take a look at the representational architecture underlying the subjective experience of directedness in general. Phenomenologically, a transparent world-model gives rise to a reality. A transparent system-model gives rise to a self that is embedded in this reality. If there is also a transparent model of the transient and constantly changing relations between the perceiving and acting self and the objects and persons in this reality, this results in what I called a "phenomenal first-person perspective" above. A genuine inner perspective arises if only and only if the system represents itself as currently interacting with the world *to itself*, and if it does not recognize this representation *as* a representation. Now, it has a conscious model of the intentionality relation (a PMIR). It represents itself as directed towards certain aspects of the world. Its phenomenal space is a *perspectival* space, and its experiences are *subjective* experiences.

The intentionality relation is primarily an epistemic relation between subject and object. A mental state becomes a carrier of knowledge in virtue of being directed at something other than itself — like an arrow pointing from a person's mind to an object in the real or even just in a possible world. Philosophers say that this type of mental state has *intentional content*. Its content is what the arrow is pointing at. This may be an image, a proposition, or even the goal of an action — as philosophers say, there is "practical intentionality" in terms of your being directed at certain "satisfaction conditions" (e.g., an action goal), and there is "theoretical intentionality" in terms of being directed at the "truth conditions" (e.g., of a sentence). If many of these arrows are consciously available, represented by the brain on the functional level of global availability, this results in a temporally extended first-person perspective. In short, it is one thing to be a biological organism that represents the world, and it is another thing to consciously represent yourself *as representing*, in "real-time" and while this is actually happening. SMT wants to understand the latter case. Now,

there is not only a neurobiologically anchored core self, a self-presentation, but also a dynamic phenomenal simulation of the *self as subject* embedded in the world via constantly changing epistemic relations and agentive interactions. Of course, there is much more to be said about the central notion of a PMIR.[9] But the core idea is as follows: a conscious human being is a system that is capable of dynamically *co-representing* the representational relation while representational acts are taking place, and the instrument it uses for this purpose is the PMIR. The phenomenal model of the intentionality relation (PMIR), is just another naturally evolved virtual organ, just like the PSM. The content of higher order forms of self-consciousness is always relational: the self *in the act of knowing* (Damasio, 1999, p. 168ff), the *currently acting* self. The ability to co-represent this intentional relationship itself while actively constructing it in interacting with a world is what it means to be a subject.

Of course, the way we subjectively experience this subject–object relation is a simplified version of the actual processes — in a sense, it is a functionally adequate confabulation. Once again, evolution favored a simple, but elegant solution. The virtual self-moving through the phenomenal world does not have a brain, a motor system, or sensory organs: certain parts of the environment appear directly in its mind; the perceptual process is experienced as effortless and immediate. Body movements also appear to be caused ''directly.'' Such effects are typical for *our* type of subjective experience and — seen as a neurocomputational strategy — they have the advantage of creating a user-friendly interface. What was defined as ''transparency'' above is a way of describing the *closed* structure of this multimodal, high-dimensional user interface — the brain's user surface. The phenomenal self is the part of this interface that the system uses to experience *itself* as a whole, to represent itself as a thinking self and an agent. This virtual agent ''sees with his eyes'' and ''acts with his hands.'' He does not know that he has a visual or a motor cortex. The PSM is the interface that the system uses to functionally appropriate its own hardware, to control its own low-level dynamics and to become *autonomous*. The intentional arrows connecting this agent to objects and other selves in the currently active reality-model are phenomenal representations of transient subject–object relations — and frequently, they too cannot be recognized *as* representational processes. In standard situations, the consciously experienced first-person perspective is the content of a transparent PMIR.

All this takes place within a phenomenal window of presence. The contents of phenomenal experience not only create a world; they also create a *present* (see Metzinger, 2003a, Section 3.2.2). In a sense, the core of phenomenal consciousness is just the creation of an island of presence within the physical flow of time (see Ruhnau, 1995 and the references given there, especially to the work of Ernst Pöppel). Experiencing means ''being there,'' and this necessarily includes ''being now.'' It means processing information in a very specific way. It means repeatedly and continuously binding discrete events that have already been represented as such into temporal *gestalts*, into a consciously experienced moment. Many recent empirical data clearly demonstrate that in a certain sense, the consciously experienced present is a *remembered* present (see for instance, Edelman, 1989). In this sense, even the phenomenal ''Now'' is a representational construct, a *virtual* present. And this finally helps understand what it means to say that phenomenal space is a virtual space: its content is a *possible* reality.[10] The realism of phenomenal experience arises because it represents a possibility — the best hypothesis there is at a given moment — as an untranscendable reality, or an *actuality*. In other words, the mechanisms creating temporal

[9]Of course, the theory of the PMIR is more complex than I can explain in this brief overview. Apart from Metzinger (2003a), I recommend Section 4 of Metzinger (2005a, p. 26ff) for readers interested in the idea. A more detailed discussion, specifically applied to the representational architecture of conscious volitional acts, can be found in Metzinger (2006a).

[10]My own ideas on this point are very similar to those discussed by Antti Revonsuo: *Virtual reality* is simply the best technological metaphor for phenomenal consciousness we currently have. See Revonsuo (1995, 2000), and especially Revonsuo (2006).

experience and our subjective sense of presence are transparent as well. Then, finally, this point also has to be applied to the special case of self-modeling because the virtual character of both the self-model *and* the window of presence are not available on the level of subjective experience itself, the system they represent turns into a *currently present subject*.

SMT solves the homunculus problem, because we can now see how no "little man in the head" is needed to interpret and "read out" the content of mental representations. It is also maximally parsimonious, as it allows us to account for the emergence of self-consciousness without assuming the existence of a substantial self. Does all this mean that the self is only an illusion? On second glance, the popular concept of the "self-illusion" and the metaphor of "mistaking oneself for one's inner picture of oneself" contain a logical error: *Whose* illusion could this be? Speaking of illusions presupposes someone *having* them. But something that is not an epistemic subject in a strong sense of conceptual/propositional knowledge is simply *unable* to confuse itself with anything else. Truth and falsity, reality and illusion do not exist for biological information-processing systems at the developmental stage in question. So far, we only have a theory of the phenomenology of selfhood, not a theory of self-knowledge. Here, I have only very briefly sketched how a *phenomenal* first-person perspective can be the product of natural evolution. Subjectivity in an *epistemic* sense, an epistemic first-person perspective is yet another step. Of course, the phenomenology of selfhood, of non-conceptual self-consciousness, is the most important precondition for this step, because it is the precondition for genuinely *reflexive*, conceptual self-consciousness. In a way, this is the whole point behind the theory: if we want to take high-level forms of subjectivity and intersubjectivity seriously, we must be modest and careful at the beginning, focusing on their origins on the level of non-conceptual content and self-organizing neural dynamics. And readers will not be surprised that the author of this chapter holds that subjective, first-person *knowledge* is precisely knowledge associated with a specific inner mode of presentation, namely as knowledge *under a PMIR*. Subjectivity in the epistemological sense can be naturalized as well — but only if we can tell a convincing evolutionary and neuroscientific story about how this representational architecture, this highly specific, indexical inner mode of presentation, could actually have developed in a self-organizing physical universe in the first place. Ultimately, and obviously, every single instance of the PSM/PMIR is identical with a specific time-slice in the continuous, dynamical self-organization of coherent activity taking place in an individual biological brain. In this ongoing process on the subpersonal level there is no agent — no evil demon that could count as the *creator* of an illusion. And there is no entity that could count as the *subject* of the illusion either. There is nobody *in* the system who could be mistaken or confused about anything — the homunculus does not exist. On the level of phenomenology, as well as on the level of neurobiology, the conscious self is neither a form of knowledge nor an illusion. It just is what it is.

Acknowledgment

I am greatly indebted to Jennifer M. Windt for help with the English version and to Rahul Banerjee for very stimulating discussion and a number of helpful critical comments.

References

Alvarado, C.S. (1986) Research on spontaneous out-of-body experiences: a review of modern developments, 1960–1984. In: Shapin B. and Coly L. (Eds.), Current Trends in PSI Research. Parapsychology Foundation, New York.

Alvarado, C.S. (2000) Out-of-body experiences. In: Cardeña E., Lynn S.J. and Krippner S. (Eds.), Varieties of Anomalous Experience: Examining the Scientific Evidence. American Psychological Association, Washington, DC.

Armel, K.C. and Ramachandran, V.S. (2003) Projecting sensations to external objects: evidence from skin conductance response. Proc. R. Soc. Lond. Biol., 270: 1499–1506.

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

Baker, L.R. (1998) The first-person perspective: a test for naturalism. Am. Philos. Q., 35: 327–346.

Baker, L.R. (2000) Die Perspektive der ersten Person: Ein Test für den Naturalismus. In: Keil G. and Schnädelbach H. (Eds.), Naturalismus-Philosophische Beiträge. Suhrkamp, Frankfurt am Main.

Baker, L.R. (2007) Naturalism and the first-person perspective. In: Gasser G. (Ed.), How Successful is Naturalism? Publications of the Austrian Ludwig Wittgenstein Society. Ontos Verlag, Frankfurt am Main.

Bieri, P. (1987) Evolution, Erkenntnis und Kognition. In: Lütterfelds W. (Ed.), Transzendentale oder Evolutionäre Erkenntnistheorie? Wissenschaftliche Buchgesellschaft, Darmstadt.

Bischof-Köhler, D. (1989) Spiegelbild und Empathie. Huber, Bern, Nachdruck.

Bischof-Köhler, D. (1996) Ichbewusstsein und Zeitvergegenwärtigung. Zur Phylogenese spezifisch menschlicher Erkenntnisformen. In: Barkhaus A., Mayer M., Roughley N. and Thürnau D. (Eds.), Identität, Leiblichkeit, Normativität. Neue Horizonte anthropologischen Denkens. Suhrkamp, Frankfurt am Main.

Blackmore, S. (1982) Beyond the Body: An Investigation of Out-of-the-Body-Experiences. Granada, London.

Blackmore, S.J. (1986) Spontaneous and deliberate OBEs: a questionnaire survey. J. Soc. Psych. Res., 53: 218–224.

Blanke, O., Landis, T., Spinelli, L. and Seeck, M. (2004) Out-of-body experience and autoscopy of neurological. Brain, 127: 243–258.

Blanke, O., Mohr, C., Michel, C.M., Pascual-Leone, A., Brugger, P., Seeck, M., Landis, T. and Thut, G. (2005) Linking out-of-body experience and self processing to mental own-body imagery at the temporoparietal junction. J. Neurosci., 25(3): 550–557.

Bongard, J., Zykov, V. and Lipson, H. (2006). Resilient machines through continuous self-modeling. Science, 314: 1118. In particular, see also free online support material at http://www.sciencemag.org/cgi/content/full/314/5802/1118/DC1

Botvinick, M. (2004) Probing the neural basis of body ownership. Science, 305: 782–783.

Botvinick, M. and Cohen, J. (1998) Rubber hand ''feel'' touch that eyes see. Nature, 391: 756.

Brentano, F. (1973) [1874]. Psychologie vom empirischen Standpunkt. Erster Band. Meiner, Hamburg.

Brugger, P., Regard, M. and Landis, T. (1997) Illusory reduplication of one's own body: phenomenology and classification of autoscopic phenomena. Cognit. Neuropsychiatry, 2: 19–38.

Brugger, P., Kollias, S.K., Müri, R.M., Crelier, G., Hepp-Reymond, M.-C. and Regard, M. (2000) Beyond remembering: phantoms sensations of congenitally absent limbs. Proc. Natl. Acad. Sci. U.S.A., 97: 6167–6172.

Brugger, P., Regard, M. and Shiffrar, M. (2001) Hand movement observation in a person born without hands: is body scheme innate? J. Neurol. Neurosurg. Psychiatry, 70: p. 276.

Carmena, J.M., Lebedev, M.A., Crist, R.E., O'Doherty, J.E., Santucci, D.M., Dimitrov, D.F., Patil, P.G., Henriquez, C.S. and Nicolelis, M.A.L. (2003) Learning to control a brain–machine interface for reaching and grasping by primates. PLoS Biol., 1: 193–208.

Churchland, P.M. (1989) A Neurocomputational Perspective. MIT Press, Cambridge, MA.

Clark, A. (1989) Microcognition-Philosophy, Cognitive Science, and Parallel Distributed Processing. MIT Press, Cambridge, MA.

Cummins, R. (1983) The Nature of Psychological Explanation. MIT Press, Cambridge, MA.

Damasio, A. (1994) Descartes' Error. Putnam/Grosset, New York.

Damasio, A. (1999) The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Harcourt Brace & Company.

Dennett, D.C. (1987) The Intentional Stance. MIT Press, Cambridge, MA.

Dennett, D.C. (1991) Consciousness explained. Little, Brown and Company, Boston, Toronto and London.

Devinsky, O., Feldmann, E., Burrowes, K. and Bromfield, E. (1989) Autoscopic phenomena with seizures. Arch. Neurol., 46: 1080–1088.

Dittrich, A. (1985) Ätiologie-unabhängige Strukturen veränderter Wachbewusstseinszustände. Enke, Stuttgart.

Dretske, F. (1988) Explaining Behavior: Reasons in a World of Causes. MIT Press, Cambridge, MA.

Dretske, F. (1998) Die Naturalisierung des Geistes. Mentis, Paderborn.

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

Ehrsson, H.H., Spence, C. and Passingham, R.E. (2004) That's my hand! Activity in premotor cortex reflects feeling of ownership of a limb. Science, 305: 875–877.

Gallagher, S. (2005) How the Body Shapes the Mind. Oxford University Press, Oxford.

Gallese, V. (2005) Embodied simulation: from neurons to phenomenal experience. Phenomenol. Cognit. Sci., 4: 23–38.

Gallese, V. and Goldman, A. (1998) Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci., 2: 493–501.

Grush, R. (1997) The architecture of representation. Philos. Psychol., 10: 5–25.

Grush, R. (1998) Wahrnehmung, Vorstellung, und die sensomotorische Schleife. In: Heckmann H.-D. and Esken F. (Eds.), Bewusstsein und Repräsentation. Mentis, Paderborn.

Gusnard, D. (2005) Being a self: considerations from functional imaging. Conscious. Cogn., 14(4): 679–697.

Iriki, A., Tanaka, M. and Iwamura, Y. (1996) Coding of modified body schema during tool-use by macaque postcentral neurons. Neuroreport, 7: 2325–2330.

Irwin, H. (1985) Flight of mind. Scarecrow Press, Metuchen, NJ and London.

Kinsbourne, M. (1995) Awareness of one's own body: an attentional theory of its nature, development, and brain basis. In: Bermúdez J.L., Marcel A. and Eilan N. (Eds.), The Body and the Self. MIT Press, Cambridge, MA.

Lebedev, M.A. and Nicolelis, M.A.L. (2006) Brain–machine interfaces: past, present and future. Trends Neurosci., 29(9): 536–546. (http://www.science-direct.com/science/article/B6T0V-4KFV367-2/2/1c479d63267ad95d2a57070bfd516003)

Lenggenhager, B., Tadi, T., Metzinger, T. and Blanke, O. (2007) Video ergo sum: manipulating bodily self-consciousness. Science, 317(5841).

Lycan, W.G. (1996) Consciousness and Experience. MIT Press, Cambridge, MA.

Maravita, A. (2006) From "body in the brain" to "body in space": sensory and intentional components of body representation. In: Knoblich G., Thornton I., Grosjean M. and Human Shiffrar M. (Eds.), Body Perception from the Inside Out. Oxford University Press, New York.

Maravita, A. and Iriki, A. (2004) Tools for the body (schema). Trends Cogn. Sci., 8: 79–86.

Melzack, R. (1989) Phantom limbs, the self and the brain: The D.O. Hebb memorial lecture. Can. Psychol., 30: 1–16.

Melzack, R. (1992) Phantom limbs. Sci. Am., 266: 90–96.

Melzack, R., Israel, R., Lacroix, R. and Schultz, G. (1997) Phantom limbs in people with congenital limb deficiency or amputation in early childhood. Brain, 120(Pt. 9): 1603–1620.

Metzinger, T. (1993; ^{2}1999). Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewusstseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation. Mentis, Paderborn.

Metzinger, T. (1995a). Perspektivische Fakten? Die Naturalisierung des "Blick von nirgendwo". In: Meggle G. and Nida-Rümelin J. (1997) (Eds.), ANALYOMEN 2: Perspektiven der Analytischen Philosophie. de Gruyter, Berlin, pp. 103–110.

Metzinger, T. (Ed.) (1995b). Conscious Experience. Imprint Academic, Thorverton, UK.

Metzinger, T. (1996) Niemand sein. In: Krämer S. (Ed.), Bewusstsein-Philosophische Positionen. Suhrkamp, Frankfurt am Main.

Metzinger, T. (1997) Ich-Störungen als pathologische Formen mentaler Selbstmodellierung. In: Northoff G. (Ed.), Neuropsychiatrie und Neurophilosophie. Mentis, Paderborn.

Metzinger, T. (2000). The subjectivity of subjective experience: a representationalist analysis of the first-person perspective. In: Metzinger T. (Ed.), Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press, Cambridge, MA. Revised version (2004): Networks, 3–4: 33–64.

Metzinger, T. (2003a; ^{2}2004). Being No One. The Self-Model Theory of Subjectivity. MIT Press, Cambridge, MA.

Metzinger, T. (2003b) Phänomenale Transparenz und kognitive Selbstbezugnahme. In: Haas-Spohn U. (Ed.), Intentionalität zwischen Subjektivität und Weltbezug. Mentis, Paderborn, pp. 411–459.

Metzinger, T. (2003c) Phenomenal transparency and cognitive self-reference. Phenomenol. Cogn. Sci., 2: 353–393. doi: 10.1023/B:PHEN.0000007366.42918.eb

Metzinger, T. (2004a) Why are identity-disorders interesting for philosophers? In: Schramme T. and Thome J. (Eds.), Philosophy and Psychiatry. de Gruyter, Berlin.

Metzinger, T. (2004b). Appearance is not knowledge: the incoherent strawman, content–content confusions and mindless conscious subjects. Invited commentary for Alva Noë und Evan Thompson: "Are there neural correlates of consciousness?" in a special issue of J. Conscious. Stud., 11(1): 67–71.

Metzinger, T. (2005a). Précis of "Being No One". In: PSYCHE: An Interdisciplinary Journal of Research on Consciousness, 11(5): 1–35. URL: www.psyche.cs.monash.edu.au/

Metzinger, T. (2005b) Out-of-body experiences as the origin of the concept of a "soul". Mind Matter, 3(1): 57–84.

Metzinger, T. (2005c) Die Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung in sechs Schritten. In: Herrmann C.S., Pauen M., Rieger J.W. and Schicktanz S. (Eds.), Bewusstsein: Philosophie, Neurowissenschaften, Ethik. UTB/Fink, Stuttgart.

Metzinger, T. (2006a) Conscious volition and mental representation: towards a more fine-grained analysis. In: Sebanz N. and Prinz W. (Eds.), Disorders of Volition. MIT Press, Cambridge, MA, S19–S48.

Metzinger, T. (2006b). Reply to Gallagher: different conceptions of embodiment. In: PSYCHE: An Interdisciplinary Journal of Research on Consciousness, 12(4). URL: www.psyche.cs.monash.edu.au/symposia/metzinger/reply_to_;Gallagher.pdf

Metzinger, T. and Gallese, V. (2003). The emergence of a shared action ontology: building blocks for a theory. In: Knoblich G., Elsner B., von Aschersleben G. and Metzinger T. (Eds.), Self and Action. Special issue of Conscious. Cogn., 12(4): 549–571.

Millikan, R.G. (1984) Language, Thought, and other Biological Categories. MIT Press, Cambridge, MA.

Millikan, R.G. (1993) White Queen Psychology and Other Essays for Alice. MIT Press, Cambridge, MA.

Muldoon, S. and Carrington, H. (1929) The projection of the astral body. Rider and Co., London.

Nagel, T. (1986) The View from Nowhere. Oxford University Press, New York.

Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A. and Shulman, G.L. (2001) A default mode of brain function. Proc. Natl. Acad. Sci. U.S.A., 98: 676–682.

Ramachandran, V.S. and Blakeslee, S. (1998) Phantoms in the Brain. William Morrow and Company, Inc., New York.

Ramachandran, V.S. and Rogers-Ramachandran, D. (1996) Synaesthesia in phantom limbs induced with mirrors. Proc. R. Soc. Lond. B, 377–386.

Revonsuo, A. (1995) Consciousness, dreams, and virtual realities. Philos. Psychol., 8: 35–58.

Revonsuo, A. (2000) Prospects for a scientific research program on consciousness. In: Metzinger T. (Ed.), Neural Correlates of Consciousness: Empirical and Conceptual Questions. MIT Press, Cambridge, MA.

Revonsuo, A. (2006) Inner Presence. MIT Press, Cambridge, MA.

Ruhnau, E. (1995). Time-Gestalt and the observer. In: Metzinger T., (Ed.), Conscious Experience. Imprint Academic, Thorverton.

O'Shaughnessy, B. (1995) Proprioception and the body image. In: Bermúdez J.L., Marcel A. and Eilan N. (Eds.), The Body and the Self. MIT Press, Cambridge, MA.

Vuilleumier, P., Despland, P.A., Assal, G. and Regli, F. (1997) Héautoscopie, exta-se et hallucinations expérientielles d'origine épileptique. Rev. Neurol., 153: 115–119.

Waelti, E. (1983) Der dritte Kreis des Wissens. Ansata, Interlaken.

Windt, J.M. and Metzinger, T. (2007) The philosophy of dreaming and self-consciousness: what happens to the experiential subject during the dream state? In: Barrett D. and McNamara P. (Eds.), The New Science of Dreaming. Praeger Imprint/Greenwood Publishers, Westport, CT.

Yates, J. (1975) The content of awareness is a model of the world. Psychol. Rev., 92: 249–284.