The Self-Model Theory of Subjectivity: A Brief Summary with Examples

Thomas Metzinger *
metzinger@uni-mainz.de

ABSTRACT

My epistemic goal in this lecture will be the simplest form of phenomenal self-consciousness: I want to contribute to a deeper understanding of non-conceptual, pre-reflexive layers in conscious selfrepresentation.
Conceptually, I will defend the claim that agency is not part of the metaphysically necessary supervenience-basis for bodily self-consciousness. On the level of new empirical data I want to show how out-of-body experiences (OBEs) and full-body illusions (FBIs) provide an interesting new entry point for investigating the nature of the phenomenal self. I will then proceed to sketch a new research program and advertise a new research target: "Minimal Phenomenal Selfhood", ending with an informal argument for the thesis that agency, phenomenologically as well as functionally, is not necessary condition for self-consciousness.

"SMT": What is the Self-Model Theory of Subjectivity?

The goal of this chapter is to give a very brief summary of the "self-model theory of subjectivity" (SMT). For an accessible introduction I recommend my book *Il tunnel dell'io - Scienza della mente e mito del soggetto* (2010).[1] Here,

* Johannes Gutenberg-Universität (Mainz)

[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.theassc.org/files/assc/2608.pdf>. On the monograph level, 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. The shortest freely available summary can be found in *Scholarpedia* 2(10), 4174 at <www.scholarpedia.org/article/Self_Models>.

I will use 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 self-model theory of subjectivity 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 "ontological self-subsistence" – 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.[2] What does exist is an intermittent process, 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 way you *appear* to yourself, subjectively, consciously. Under SMT, this conscious *experience* of being a self is conceptually 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.[3] This theory, however, first focuses on the minimal representational and functional properties that a naturally evolved information-processing system – such as *Homo sapiens* – has to have in order to satisfy the constraints for realizing these higher-order forms of self-consciousness. Therefore, the first question we will have to answer is this: What, in the case of human beings, are minimally sufficient conditions for the emergence of a conscious self? Later we can ask: Are some of these conditions strictly necessary?

SMT assumes that the properties in question are representational and functional properties. In other words, the phenomenal property that allows us to become a person in the first place – namely, "selfhood" – is analyzed with

[2] See Metzinger 2010.

[3] I analyzed the relation between conceptual and non-conceptual contents of self-consciousness in detail in Metzinger 2003b. 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.

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, selfhood and a subjective, inward perspective – is conceptually described 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: if it should turn out 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 epistemic progress. This type of progress could mean being able to describe, in a much more precise and fine-grained manner and with an 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, but an open-outcome research heuristics: 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.

THE PHENOMENAL SELF-MODEL

What we like to call "the self" in folk-psychological contexts is the phenomenal self: that part of our mental self-representation, which is immediately given in subjective 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 typically 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. Central, but not phenomenologically necessary features of self-

consciousness are:

- the sense of ownership (for individual mental contents, like *my* thoughts or feelings);
- the sense of boundedness (i.e., having a conscious self-world boundary, as in a body image);
- epistemic agency (i.e., the experience of a self in the act of knowing, of the self *as subject*; as in attentional control);

Candidates for phenomenologically necessary features of self-consciousness are:

- phenomenology of existence (i.e., the fundamental "sense of being", of "realness");
- presence (i.e., localization in a temporal order);
- self-localization (i.e., in a spatial frame of reference)
- identification (i.e., "global" ownership for the integrated contents of self-consciousness as a whole; see Blanke and Metzinger 2009).

The first step consists in the introduction of a new theoretical entity: the phenomenal self-model (PSM). It is the representational basis for instantiating the relevant phenomenal properties. One of our key questions was: Which set of minimally sufficient *representational* properties does a system like the human brain 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 typically determined by the system's very own properties. Whenever such a global 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.

What we need is a comprehensive theory of the self-model of *Homo sapiens*.[4] I assume that this will be a predominately neurocomputational

[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-

theory.[5] 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 *phenomenal* self-model is exactly that part of the *mental* self-model that is currently embedded in a higher-order integrated structure, the globally available model of the world (Yates 1975, Baars 1988).[6] In other words: Certain parts of the self-model can be unconscious and functionally active at the same time. The phenomenal self-model is a coherent multimodal structure that probably depends on a partially innate, "hard-wired" model of the system's spatial properties.[7] This approach 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 data structure in the system's central nervous system. A first-person perspective is a naturally evolved data-*format*, an inner mode of presentation, in which the system is given to itself as a knowing self.

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 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. Again, what happens when you wake up in the morning, i.e., when the system that you are "comes to itself", is that this transient

model, including its evolutionary history and its necessary social correlates.

[5] See, for instance, Churchland 1989.

[6] For a detailed analysis of the criteria for distinguishing different degrees of consciousness, see Metzinger 2003a, chapter 3.

[7] 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*; Damasio 1994, 1999; Metzinger 2003a and Scholarpedia article quoted in footnote 1.

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*. Importantly, the development of ever more efficient self-models as a new form of "virtual organ" 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" (*simulazione "incarnata"*), as Vittorio Gallese (2005) aptly puts it – for example, 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*.[8] 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.[9] The development and activation of this computational module plays a role *for* the system: the functional self-model possesses a true evolutionary description, that is, it was a *weapon* that was invented and continuously optimized in the course of a «cognitive arms race» (Clark 1989, p. 61). Functional properties still determine phenomenal properties, but the functions realized are not merely computational, but also biological proper functions.

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

[8] 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.

[9] See for instance Bieri 1987; Dennett 1987; Dretske 1988, 1998; Lycan 1996; Millikan 1984, 1993.

human self-model.[10] 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*.[11] 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 phenomenal 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 phenomenal self-model 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",

[10] See Ramachandran and Rogers-Ramachandran 1996; a popular account can be found in Ramachandran and Blakeslee 1998, pp. 46ff. The figure was published courtesy of Ramachandran.

[11] Related ideas are discussed by Grush 1997, and 1998, p. 174; see also Ramachandran and Rogers-Ramachandran 1996, p. 378.

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".
"Open your eyes". (Ramachandran 1998, pp. 47-48)[12]

Figure 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.

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 phenomenal self-model. What made the sudden occurrence

[12] For the clinical and experimental details, see Ramachandran and Rogers-Ramachandran 1996.

of kinaesthetic movement sensations in the lost subregion of the self-model possible was the installation of an additional source of feedback, of "virtual information". 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? 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 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 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. 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).[13]

[13] For further details concerning the phenomenological profile see ibid.; for an interesting experimental follow-up study demonstrating the intactness of the phenomenal model of kinaesthetic and postural limb properties, see Brugger *et al.* 2001.

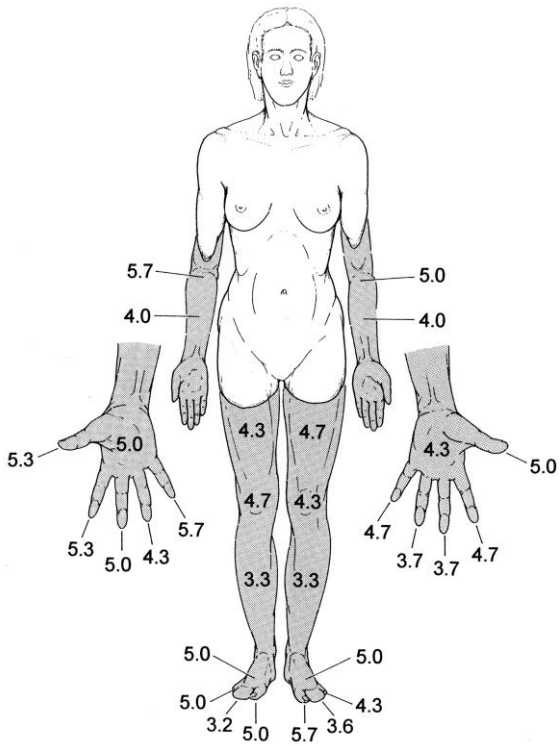


Figure 2: Evidence for an innate component of the PSM? Phantoms (shaded areas) in a subject with limb amelia. The numbers are vividness ratings for the felt presence of different phantom body parts on a 7-point scale from 0 (no awareness) to 6 (most vivid impression). Picture courtesy of Peter Brugger,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

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 – perhaps the global shape of the body? 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 content of this self-representation, namely the PSM postulated by the self-model theory. 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 it a kind of "fixed idea", anchored in an inborn and genetically predetermined *nucleus*?

Let us now turn to example # 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 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. 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.

Josh Bongard, Victor Zykov and Hod Lipson (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 #1 and #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.

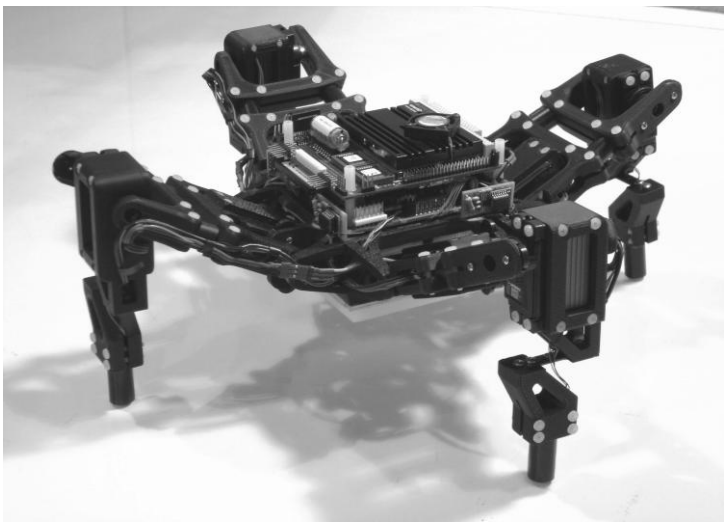
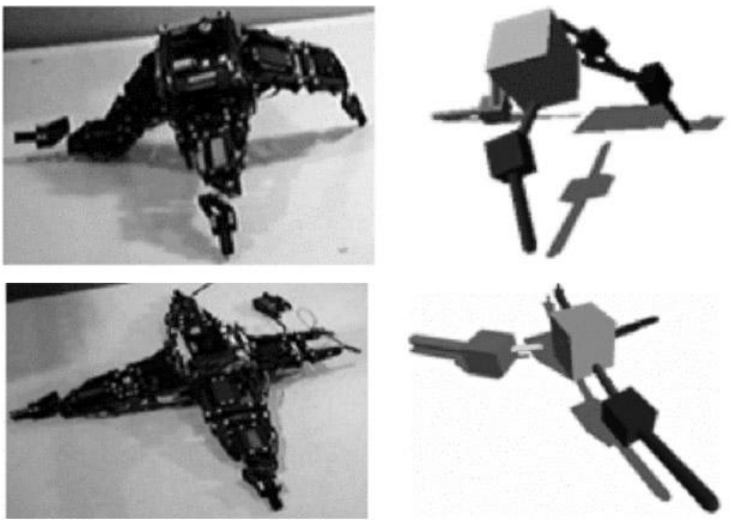


Figure 3a: Starfish, a four-legged physical robot that has eight motorized joints, eight joint angle sensors, and two tilt sensors.



Figures 3b and 3c: 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.

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.

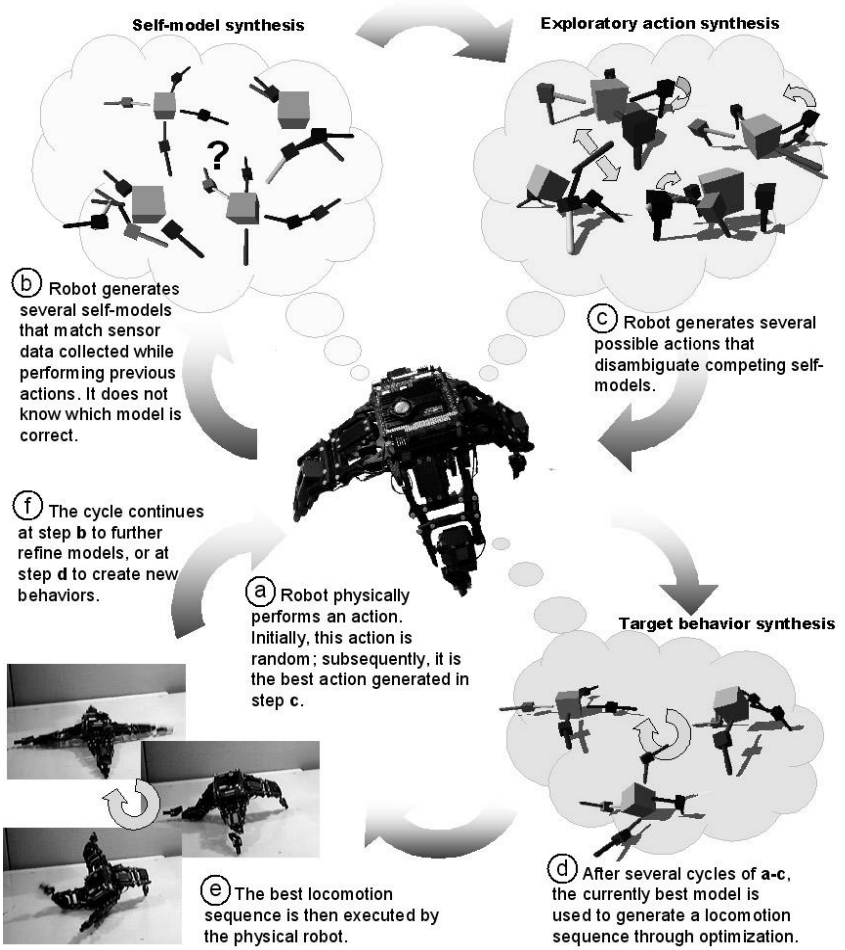


Figure 3d: The robot continuously cycles through action execution. (A and B) Self-model synthesis. The robot physically performs an action (A). Initially, this action is random; later, it is the best action found in (C). The robot then generates several self-models to match sensor data collected while 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) to (C), the best current model is used to generate locomotion sequences through optimization. (E) The best locomotion sequence is executed by the physical device (F).

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 modelling 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 #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, that is, it does not have to be a *phenomenal* self-model (i.e., a PSM). Awareness obviously is a second step.[14] Third, a self-model supports 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.

[14] See Metzinger 1995b, 2000b for a first overview; Metzinger 2003a, section 3.2, for an additional set of ten constraints to be satisfied for conscious experience.

THE SENSE OF OWNERSHIP

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. [15] Subjectively experienced ownership 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; that is, 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. 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.

Here is a concrete example for what I mean by local ownership, example #4:

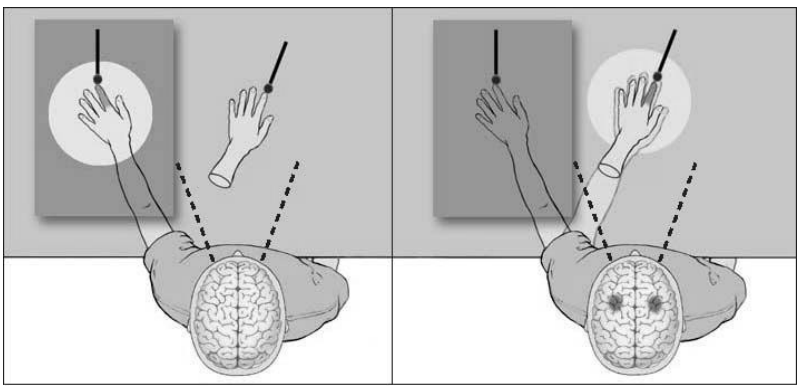


Figure 4*: The rubber-hand illusion.* A healthy subject experiences an artificial limb as part of her own body. The subject observes a facsimile of a human hand while one of her own hands is concealed (grey square). Both the artificial rubber hand and the invisible hand are then stroked repeatedly and 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

[15] For case-studies, see Chapter 7 in Metzinger 2003a.

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.

In the rubber-hand illusion, 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.[16] The philosophically intriguing question, of course, is this: Could *whole-body* illusions exist as well? The answer is yes, but for reasons of space I must refer readers to my book *Il tunnel dell'io* and the references given there.

The Bodily Self as a Functional Anchor of Phenomenal Space

The central theoretical problem on the functional level of description can be summed up by the following question: What exactly is the difference between

[16] See especially Botvinick and Cohen 1998, and the neuroimaging study by Ehrsson *et al.* 2004, and Botvinick 2004.

the phenomenal self-model 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 (that is, 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 its 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 phenomenal self-model 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 #2? The weaker degree of phenomenological "vividness" or "realness" in her phantom limbs may reflect exactly the absence of permanent 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).

Transparency and the Naïve-Realistic Self-Misunderstanding

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. But 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 global *system-model*. The decisive philosophical question is this: 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 "non-conceptual self-awareness"? The solution has to do with what philosophers call "phenomenal transparency". [17] Many conscious representational states generated by the system are *transparent*, that is, 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

[17] For a short explanation of the concept of "phenomenal transparency" see Metzinger 2003b.

no or very little knowledge about reality (or your own mind) and 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 cannot 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 microfunctional 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 lack of self-knowledge. Epistemologically speaking, it is an implicit, not an explicit lack of knowledge: As Franz Brentano (1973, pp. 165-166) and Daniel Dennett (1991, p. 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.[18]

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

[18] See also Dretske 1998, pp. 45ff.

content and representational carrier on the level of subjective experience.

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. For example, if there is a conscious body-model in our brain and we have no introspective access to the construction process (the physical vehicle), then we will *identify* with its content. By logical necessity, a phenomenally transparent self-model will create the subjective experience of existence (i.e., the phenomenology of "being real") and the phenomenology of identification (i.e., 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. This commits me to a specific prediction: Were the PSM to lose 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.

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 an epistemic subject, 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. 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 embodiment is a necessary 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. 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 is identical with a specific time-slice in the continuous, dynamical self-organization of coherent activity taking place in a specific 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.

REFERENCES

Armel, K. C., & Ramachandran, V. S. (2003). Projecting sensations to external objects: Evidence from skin conductance response. *Proceedings of the Royal Society of London: Biological, 270*, 1499-1506.

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

Baker, L. R. (1998). The first-person perspective: A test for naturalism. *American Philosophical Quarterly, 35*, 327-346.

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

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

Bermúdez, J. L., Marcel, A., & Eilan, N. (1995) (Eds.). *The Body and the Self.* Cambridge, MA: MIT Press.

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

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

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

Blanke, O., & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences, 13*(1), 7-13.

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

Botvinick, M. (2004). Probing the neural basis of body ownership. *Science, 305*, 782-3.

Botvinick, M., & Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature, 391*, 756.

Brentano, F. (1973). *Psychologie vom empirischen Standpunkt.* Vol. I. Hamburg: Meiner. [1874]

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

Brugger, P., Kollias, S. K., Müri, R. M., Crelier, G., Hepp-Reymond, M.-C., & Regard, M. (2000). Beyond re-membering: Phantoms sensations of congenitally absent limbs. *Proceedings of the Naticonal Academy of Science USA, 97*, 6167-72.

Brugger, P., Regard, M., & Shiffrar, M. (2001). Hand movement observation in a person born without hands: Is body scheme innate? *Journal of Neurology, Neurosurgery, and Psychiatry, 70*, 276.

Churchland, P. M. (1989). *A Neurocomputational Perspective*. London-Cambridge, MA: MIT Press.

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

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

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

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*. London-Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained*. Boston-Toronto-London: Little, Brown & Co.

Devinsky, O., Feldmann, E., Burrowes, K., & Bromfield, E. (1989). Autoscopic phenomena with seizures. *Archives of Neurology, 46*, 1080-8.

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

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

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

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

Ehrsson, H. H., Spence, C., & 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: Oxford University Press.

Gallese, V. (2005). Embodied simulation: from neurons to phenomenal experience. *Phenomenology and the Cognitive Sciences, 4*, 23-38.

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*, 493-501.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology, 10*, 5-25.

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

Iriki, A., Tanaka, M., & Iwamura, Y. (1996). Coding of modified body schema during tool-use by macaque post-central neurons. *Neuroreport, 7*, 2325-2330.

Kinsbourne, M. (1995). Awareness of one's own body: An attentional theory of its nature, development, and brain basis. In Bermúdez *et al.* 1995.

Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007; submitted). Manipulating embodiment: A full body analogue of the rubber hand illusion.

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

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

Maravita, A., & Iriki, A. (2004). Tools for the body (schema). *Trends in Cognitive Sciences, 8*, 79-86.

Melzack, R. (1989). Phantom limbs, the self and the brain: The D. O. Hebb memorial lecture. *Canadian Psychology, 30*, 1-16.

Melzack, R. (1992). Phantom limbs. *Scientific American, 266*, 90-6.

Melzack, R., Israel, R., Lacroix, R., & Schultz, G. (1997). Phantom limbs in people with congenital limb deficiency or amputation in early childhood. *Brain, 120* (Part 9), 1603-20.

Metzinger, T. (1993). *Subjekt und Selbstmodell. Die Perspektivität phänomenalen Bewusstseins vor dem Hintergrund einer naturalistischen Theorie mentaler Repräsentation.* Paderborn: Mentis.

Metzinger, T. (1995a). *Perspektivische Fakten? Die Naturalisierung des "Blick von nirgendwo".* In G. Meggle & J. Nida-Rümelin (Eds.) (1997), *Analyomen 2 - Perspektiven der Analytischen Philosophie* (pp. 103-110). Berlin-New York: De Gruyter.

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

Metzinger, T. (2000a). *Die Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung für Nicht-Philosophen in fünf Schritten.* In W. Greve (Ed.), *Psychologie des Selbst.* Weinheim: Beltz/Psychologie Verlags Union.

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

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

Metzinger, T. (2003a). *Being No One. The Self-Model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Metzinger, T. (2003b). Phenomenal transparency and cognitive self-reference. *Phenomenology and the Cognitive Sciences, 2*, 353-393. Published online, doi: 10.1023/B:PHEN.0000007366.42918.eb.

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

Metzinger, T. (2004b). *Appearance is not knowledge: The incoherent strawman, content-content confusions and mindless conscious subjects.* Invited commentary for Alva Noë and Evan Thompson: "Are there neural correlates of consciousness?". In a special issue of *Journal of Consciousness Studies, 11*(1), 67-71.

Metzinger, T. (2005a). Précis of "Being No One". *Psyche - An Interdisciplinary Journal of Research on Consciousness*. <www.theassc.org/files/assc/2608.pdf>

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

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

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

Metzinger, T. (2006b). Reply to Gallagher: Different conceptions of embodiment. *Psyche - An Interdisciplinary Journal of Research on Consciousness, 12*(4). <http://psyche.cs.monash.edu.au/symposia/metzinger/reply_to_Gallagher.pdf>

Metzinger, T. (2010). *The No-Self-Alternative*. In S. Gallagher (Ed.), *Oxford Handbook of the Self*. Oxford: Oxford University Press.

Metzinger, T., & Gallese, V. (2003). The emergence of a shared action ontology: Building blocks for a theory. In G. Knoblich, B. Elsner, G. von Aschersleben & T. Metzinger (Eds.), *Self and Action*. Special issue of *Consciousness & Cognition , 12*(4), 549-571.

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

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

Nagel, T. (1986). *Der Blick von nirgendwo*. Frankfurt am Main: Suhrkamp.

Newen, A., & Vogeley, K. (Eds.) (2000). *Das Selbst und seine neurobiologischen Grundlagen*. Paderborn: Mentis.

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

Ramachandran, V. S., & Rogers-Ramachandran, D. (1996). Synaesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society London, B*, 377-86.

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

Revonsuo, A. (1995). Consciousness, dreams, and virtual realities. *Philosophical Psychology, 8*, 35-58.

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

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

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

Vuilleumier, P., Despland, P. A., Assal, G., & Regli, F. (1997). Héautoscopie, exta-se et hallucinations expérientielles d'origine épileptique. *Revue Neurologique, 153*, 115–119.

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

Yates, J. (1985). The content of awareness is a model of the world. *Psychological Review, 92*, 249-84.