Thomas Metzinger

The *Subjectivity* of Subjective Experience: A Representationalist Analysis of the First-Person Perspective[1]

Introduction: Finding the Neural Correlates of Consciousness Through a Representationalist Analysis

Why Is a Representationalist Analysis Useful?

In this chapter I will briefly sketch the outline of a representationalist theory of subjective experience. (See also Metzinger 1993, forthcoming). A representationalist theory is one that chooses to analyze its target properties – those aspects of the domain which eventually are to be explained – on a certain level of description: by describing conscious systems as *representational* systems and conscious states as *representational* states, one hopes to achieve progress with regard to the relevant properties. This first background assumption is shared by many philosophers today (see, e.g., Dretske 1995; Lycan 1996; Metzinger 1993; Tye 1995), and one may interpret it as a weak version of Brentano's intentionalism. William Lycan (e.g., 1996: 11) has called it the "hegemony of representation": The explanatory base for *all* mental properties is formed by a definite, exhaustive set of functional and representational

[1] This is an extended and slightly revised version of a book chapter that first appeared in T. Metzinger (2000), ed., *Neural Correlates of Consciousness – Empirical and Conceptual Questions*. Cambridge, AM: MIT Press.

properties of the class of systems under investigation. However, if in this way one also aims at achieving genuine scientific progress by naturalizing classical theoretical problems, these have to be transformed into an empirically tractable form. It is therefore vital that the underlying theory of representational content is itself empirically plausible, open to data-driven conceptual changes, and limited to a specific domain (Churchland 1989; Clark 1989). One such domain are biological nervous systems, a more specific domain would be human brains in waking or dream states. What, then, are the target properties of our domain and how are they to be treated?

Table 20.1 shows the seven most important theoretical problems connected with conscious experience, in terms of phenomenological constrains imposed on any representationalist theory of consciousness.

Table 20.1
What makes a neural representation a phenomenal representation?
The seven most important phenomenological constraints for any representationalist analysis of conscious experience

Phenomenological Constraint →	Representational Content →	Causal Role ("Functional Correlate") →	Neural Correlate
"Ultra-smoothness": homogeneity of sensory primitives ("qualia"; "grain problem")	*Presentational* content (see Metzinger 1997) • First-order properties; i.e. "elementary" features • representational atomicity; i.e. structureless "density" • indicator function • non-conceptual content	Stimulus-correlated information which is • attentionally/volitionally available, but: • *not* cognitively available • "memory constraint" (Raffman 1995); i.e. discrimination without categorization	? Synchronicity of feature detectors contributing to perceptual object, read out by higher-order mechanism?
Transparency	• Immediacy of epistemic contact • Introspective access exhausted by content properties • *Naive realism*	Internal causal history and temporal microstructure of phenomenal states is globally unavailable	? "Glossing over" by temporal coherence of neural responses?
Presence	Temporal internality: *de nunc* character of phenomenal content	Activation within a virtual "window of presence"	? Recurrent activity underlying formation of short-term memory?
Embeddedness into world model	Dynamic integration into highest-order situational context	*Global* availability of information: • attentional availability • volitional availability • cognitive availability	? Global coherence? "Dynamical core"? (see chapter 9, this volume) "Highest-order binding"? (Metzinger 1995)
Convolved holism	Multi-layered, flexible part-whole relationships	"Liquid architecture": Dynamic linkages on different time-scales	? Synchronous activity with multiple time-constants?
Dynamicity	Temporal macrostructure of causal interaction-domain	Unknown aspects of network dynamics	? Which anatomical subset?
Perspectivalness	*Access from "first-person perspective"; "Subjectivity"* (see text)	*Centeredness of world model* (see text)	? (research program needed)

Intended class of systems: *Homo sapiens* in nonpathological waking states.

According to this model, the typical way of generating an empirical research program in interdisciplinary cooperation would consist in moving through one of the rows in the table from left to right. When confronted with a conceptual interpretation of a certain element of subjective experience, the job would first consist in analyzing this element as a form of representational content, generated and utilized by the system in interaction with its environment. The second step would be an attempt to give a functional description of the relevant content-bearing states in the system, by individuating them through their causal role. This opens the functional level of analysis, on which the "functional correlates" of phenomenal experience have to be specified. As soon as the intended class of systems (e.g., humans, macaques, or certain artificial systems) has been delimited, experimental approaches can proceed to isolate the minimal set of basic physical properties (e.g., the *neural* correlates) that the system needs in order to exhibit the target properties by nomological necessity. In this way domain-specific reductive explanations become possible.

Of course, real-world cooperation between disciplines is much more complex. For instance, the rows shown above can also be read from right to left: Differentiated research into the physical correlates of phenomenal states can draw our attention to fine-grained differences in the actual functional role of these correlates. This may eventually lead us to being able to describe our conscious experience in a more fine-grained manner, thereby increasing the amount of information conveyed when speaking about our own introspective experience. In other words, one way of

making progress can consist in increasing the *number* of rows in the left column.[2] There may also be oscillations between neighboring columns, as when philosophers discuss the adequacy of a representationalist conceptualization of experiential content, or when neuroscientists have doubts about a specific functional analysis offered by researchers in classical AI or cognitive science. And real-life science progresses through multiple loops: There are many trajectories through the problem landscape and across levels of description, and one and the same line of inquiry may return to the same vicinity again and again, at a slightly different angle.

For a philosopher, attempting to contribute to a reductive explanation of consciousness is never an ideology or an emotional substitute for religion. If it turns out that there are principled reasons why important features of subjective experience can never be epistemically grasped through a system of interdisciplinary cooperation like the one sketched above, she will be quite happy with this result, because the philosopher will then have achieved what she has always aimed at in the first place: epistemic progress, a growth of knowledge. All she will insist on is that those elements of consciousness purportedly resistant to any reductive approach are described with a maximum of conceptual clarity. So even antireductionists should, if only as a heuristic strategy, follow a reductionist methodology (Walter 1998).

---

[2] This, of course, is a point Paul Churchland has often made: "I suggest, then, that those of us who prize the flux and content of our subjective phenomenological experience need not view the advance of materialist neuroscience with fear and foreboding. Quite the contrary. The genuine arrival of a materialist kinematics and dynamics for psychological states and cognitive processes will constitute not a gloom in which our inner life is suppressed or eclipsed, but rather a dawning, in which its marvelous intricacies are finally *revealed* - most notably, if we apply [it] ourselves, in direct self-conscious introspection." See Churchland 1989: 66.

In this essay I will select only one aspect of phenomenal experience, its *perspectivalness*. This very last aspect mentioned in the table above is crucial for interdisciplinary research programs, because it poses the greatest methodological and epistemological problems. Any serious scientific approach to consciousness will have to rely entirely on objective, third-person criteria for the ascription of phenomenal states to a given system. How, then, could it ever do justice to the essentially subjective, first-person character of the phenomenon? (See Metzinger 1993, 1996; Nagel 1974, 1986) Can the subjectivity of the target phenomenon itself ever be turned into an explanandum for the hard sciences?

Analysanda and Explananda: What Does the *Subjectivity* of Subjective Experience Consist in?

This could work only if, first of all, the concept of a "first-person perspective" is clarified in a way that makes the corresponding phenomenon empirically tractable. Our starting point therefore is an *analysandum*, a certain allegedly self-evident manner of speaking about ourselves and the structure of our conscious experiences. In order for this analysandum to be transformed into a set of experimentally tractable *explananda*, we have to choose a suitable level of description. My first claim would be that whenever we have in the past been speaking about the subjectivity of phenomenal experience in terms of having an experiential "first-person perspective," we have merely been employing a very soft "visuogrammatical" metaphor. Our visual space, as a matter of contingent, trivial fact, possesses a perspectival,

geometrical structure:[3] This is the spatial part of the metaphor, which originates in the folk phenomenology of visual experience.

Then there is another, more abstract element of our metaphor: Self-ascriptions – for instance, of psychological properties – follow a certain logic; they take place from a grammatical "first-person" perspective. This is the grammatical aspect of the analogy, and prominent philosophers have in the past analyzed the underlying logic a lot (e.g., Castañeda 1966; Chisholm 1981; Lewis 1979; Perry 1993; Shoemaker 1996). However, if you want to understand how this logic of conceptual self-reference could ever come about, what our soft visuogrammatical metaphor actually refers to, you have to go much deeper: We have to understand what the deep representational structure is that enables beings like ourselves to pose philosophical questions of this type in the first place. Why do human beings employ visual metaphors in picking out global properties of their experiential space and in trying to understand the underlying logic of their linguistic abilities, like the self-ascription of psychological properties? What are the necessary and sufficient conditions for *any* representational system, when trying to understand its own mental properties, to run into the problem of the "immunity to error of misidentification," to eventually start wondering about "essential indexicals," "*de-se*-attitudes," or "objective selves"? Which class of representational architectures will inevitably lead all systems, that realize this architecture, into the theoretical landscape sketched at the beginning of this introduction? In order to answer this more fundamental question, on needs to

[3] Although this structure may already contain primitive, self-specifying information. See especially Bermúdez

produce an informative account of the neural and functional correlates of three very special phenomenal properties. To understand subjectivity on the logical, the epistemic or metaphysical level, one has to investigate what the *phenomenal* first-person perspective is.

A Representationalist Analysis of the Phenomenal First-Person Perspective

Three Phenomenal Target Properties

The "phenomenal first-person perspective" is one of the most fascinating natural phenomena we know, and in a certain sense we *are* this phenomenon ourselves: The essence of being a person seems to consist in the potential for being a conscious subject. It consists of three phenomenological target properties that in their conceptual interpretation constitute three different aspects of one and the same problem:

(1) *Mineness*: a higher-order property of *particular* forms of phenomenal content. Here are some typical examples of how we refer to these properties in folk psychological contexts: I experience *my* leg subjectively as always having belonged to *me*; I always experience *my* thoughts, *my* focal attention, and *my* emotions as part of my *own* consciousness; voluntary acts are initiated by *myself*.

(2) *Selfhood* ("prereflexive self-intimacy"): the phenomenal *target* property. Again, let us look at some examples of how we usually attempt to linguistically describe the phenomenal content of the representational states underlying these properties: I am

---

1998: chap.5.

someone; I experience myself as being *identical* through time; the contents of my phenomenal self-consciousness form a coherent *whole*; before initiating any intellectual operations, and independent of them, I am already "directly" acquainted with the contents of my self-consciousness.

What we frequently just call "the self" in a folk-psychological context is the *phenomenal* self: the content of self-consciousness, as given in subjective experience. We are therefore confronted with a higher-order phenomenal property that is constituted by different and constantly changing types of phenomenal content. Let us call this property "selfhood" or, to use a more technical term, "prereflexive self-intimacy" (e.g., Frank 1991). The phenomenal self arguably is the theoretically most interesting form of phenomenal content because, among other reasons, it endows our mental space with two highly interesting structural characteristics: centeredness and perspectivalness. As long as there is a phenomenal self, our conscious model of the world is a functionally centered model and is tied to what in philosophy of mind is called the "first-person perspective."

(3) *Perspectivalness*: a global, *structural* property of phenomenal space as a whole. It possesses an immovable center. According to subjective experience, the overall dynamics within this space is organized around a supramodal point of view. The phenomenal self is this center. And at this stage the conflict between first-person and third-person descriptions of our target properties becomes very obvious. This is what constitutes the philosophical puzzle: I am this center *myself*; to be phenomenally

aware means to possess an inward perspective and to take this perspective on in the subjective experience of the world and of one's own mental states.

So phenomenal subjectivity – as opposed to subjectivity in an epistemological or ontological sense – simply amounts to the fact that under standard conditions, the dynamics of conscious experience unfolds in a space that is centered on a singular, temporally extended experiential self. What now has to be achieved is a representational and a functional analysis of these properties. The pivotal question is What is the minimal set of functional and representational properties that an information-processing system must possess in order to exemplify the *phenomenal* properties under investigation? Which of these low-level properties are necessary, and which are sufficient? What, precisely, does it mean for such a system to take a phenomenal first-person perspective on the world and on its own mental states?

Step 1: What Is a Self-Model?

The first step of my analysis consists in introducing a new theoretical entity: the phenomenal *self-model*. It forms the representational instantiation basis of the phenomenal properties to be explained. The content of the phenomenal self-model, again, is what we often call "the self" in folk-psychological contexts; its content is the content of self-consciousness, bodily, emotional, and cognitive. My claim is that – ontologically speaking – no such things as selves exist in the world. What actually

exists is a special kind of self-models and their contents, and this content makes us believe that we actually do have, or are identical to, a self.

The following can be said of the self-model:

- The self-model is an episodically active representational entity, the content of which is formed solely by properties of the system itself.
- A self-model can be described on multiple, subpersonal levels. For instance, we might describe it as an activation vector or as a trajectory through some suitable state-space. A trivial but important background assumption is that, in our own case, it also possesses a true *neurobiological* description, for instance, as a complex neural activation pattern with a certain temporal fine structure. On a more abstract level the same pattern of physiological activity can also be described as a complex functional state.
- The *phenomenal* self-model is that partition of the presently active mental self-model which is currently embedded in the highest-order integrated structure, the global model of the world (see, e.g., Yates 1975; Baars 1988; see also row 4 of table 20.1). In other words, nonconscious but causally active self-models (or nonconscious subregions of such models) may exist (see also chapter 6 in this volume).
- In our own case the phenomenal self-model is a plastic, multimodal structure that is plausibly based on an innate and "hardwired" model of the spatial properties of the system (e.g., a "long-term body image"; see section "The Central Theoretical Problem on the Functional Level of Description," below, and O'Shaughnessy 1995; Bermúdez 1998; Damasio 1994; Kinsbourne 1995; Metzinger 1993, 2001), while being

functionally rooted in elementary bioregulatory processes; e.g., those systems in the upper brain stem and hypothalamus achieving homeostasis and the stability of the internal chemical milieu (see chapter 7 and Damasio 1999). The content of this underlying form of primitive self-awareness is nonconceptual and subdoxastic.[4]

In order to better understand what a self-model actually is, you can develop a parallel description on the functional level of analysis. An active self-model is the physical realization of a subpersonal functional state. It plays a certain causal role; that is, under an analytical perspective it represents a discrete set of causal relations. It is likely that the neural microevents constituting the relevant causal matrix within the system have to be individuated on a very fine-grained, subsymbolic level including the temporal structure of these events (see Singer 1993, 1994; and especially chapter 8 in this volume). However, just to illustrate the core idea, you could also take a classical cognitivist perspective. Then the self-model could be described as a transient computational module, episodically activated by the system in order to regulate its interaction with the environment.[5]

---

[4] Higher-order, conceptually mediated forms of self-consciousness are always anchored in more primitive forms of noncategorizable, cognitively unavailable forms of content (from which they very likely have developed: see Bermúdez 1998; Metzinger 1993). It is exciting to see how currently the best philosophical theorists working on analytical theories of self-representation and self-consciousness are starting to do justice to the importance of bodily awareness in the constitution of higher-level forms of subjectivity (see, e.g., Bermúdez 1998 or Cassam 1997). However, one should be careful as not to introduce a principled distinction (and hence a new dualism between reified conceptual and nonconceptual forms of content) at this point. Content is not a mysterious type of *thing*, but an abstract property of a highly fluid and complex cognitive dynamics, and the best empirically plausible models of representational content we have at this point clearly suggest the existence of a *continuum* reaching from simple sensory content to more abstract, concept-like forms (see Churchland 1998: 32). I would like to suggest that we will find exactly such a continuum in the case of self-representation as well.

[5] This, of course, is just another way of describing what happens when you wake up in the morning. There is a formal proof that every regulator of a complex system will automatically become a *model* of that system. See Conant and Ashby 1970.

The next step in defining the working concept of a self-model consists in integrating the biological history of our target phenomenon into the explanatory base. One may plausibly assume that it was, very obviously, *adaptive* to have something like a partly veridical, functionally active self-representation. Philosophers call this theoretical move a "teleofunctionalist assumption": The development and the activation of this neurocomputational module plays a role *for* the system. The functional properties of the self-model possess a true evolutionary description; that is, it is a weapon that was invented and optimized in the course of a "cognitive arms race" (a very apt and unromantic metaphor coined by Andy Clark; see Dennett 1987; Lycan 1996; Clark 1989, p. 62). The functional instantiation basis for a phenomenal first-person perspective, then, is a specific cognitive achievement: the capacity to open and employ centered representational spaces, to operate under egocentric world models. This amounts to the central necessary[6] (but not sufficient[7]) condition in ascribing subjective experience to a given system: weak subjectivity$_1$ – phenomenal subjectivity

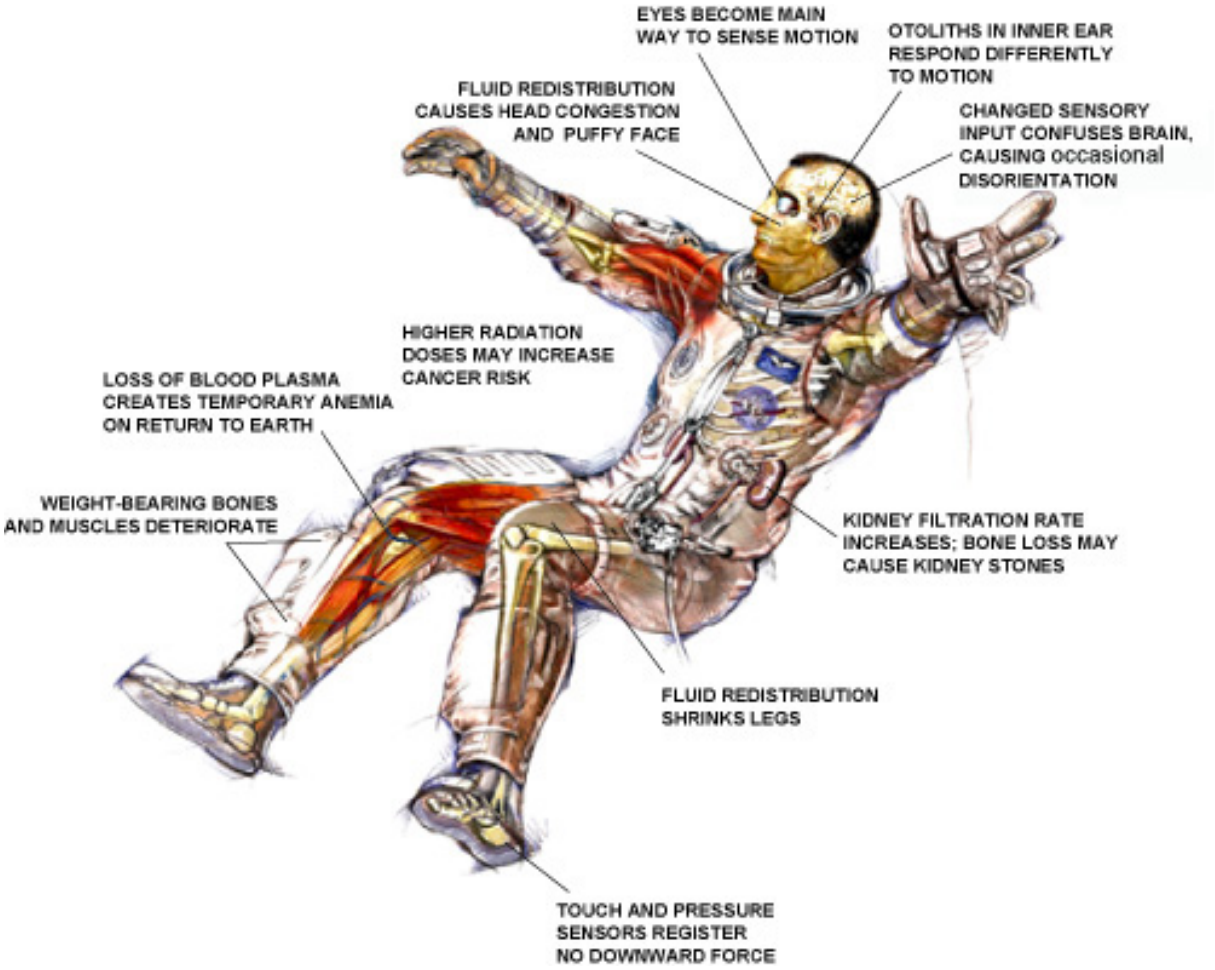
---

[6] In table 20.1 I have tried to give an overview of what I think are the seven most important phenomenological characteristics of subjective states. For each case, a convincing representationalist analysis could constitute another necessary, but not sufficient, ascription criterion for conscious experience in standard cases. However, consciousness is such a rich and complex phenomenon that counter examples can always be found if our domain is sufficiently wide (e.g., considering case studies from clinical neuropsychology, see Metzinger forthcoming). Therefore, my goal in this paper is rather modest: I am not claiming that even if we had all seven criteria (let's call them weak subjectivity 1–7) clearly defined in a conceptually convincing and empirically contentful manner, the conjunction of these criteria would already constitute a *necessary* condition. I am not even claiming that the criterion investigated here is a necessary criterion in a strict analytical sense (see note 6).
The reason for concentrating on perspectivalness and self-representation are in part philosophical and in part methodological (see the section "Step 5"). From a philosophical perspective, for many reasons the subjectivity of our target phenomenon is arguably the most interesting aspect of the problem. If one is interested in generating empirical research programs, then the neural correlates of phenomenal self-modelling are especially interesting, because under standard conditions, they will constitute the most *invariant* aspect of state-space.

[7] Note that in certain contexts this condition may not even be a *necessary* condition for the ascription of phenomenal experience in general: If we allow for noncentered, selfless states of phenomenal experience (e.g., in psychiatric cases like full depersonalization or during some classes of mystical or religious experiences), then we will have to conceptually describe these state classes as aperspectival or *nonsubjective* varieties of conscious experience. As exclusively *internal* representations, such states may still be "subjective" from an epistemological perspective (and in a quite weak sense), but they are nonsubjective on a phenomenological level of description.

(a subsymbolic, nonconceptual first-person perspective) is a property that is instantiated only if the system in question activates a coherent self-model and embeds this into its global model of the world.

Before going on, let us take a short break and look at two examples. They will serve to illustrate the concept of a phenomenal self-model by some anecdotal evidence. Astronauts, after some time in space, tend to lose their phenomenal body axis, the subjective feeling of where the top and where the bottom of their body is. When trying to eat, for instance, this can be uncomfortable. Every astronaut knows how to help his buddy: He briefly touches the sole of his partner's foot, and instantly his phenomenal body image will lock into a subjective "top-bottom-axis" again. Every astronaut also knows how to tease his partner: by tapping his head, thereby reversing the spatial frame of reference.



This shows that the self-model of human beings is a *virtual* model which, if underdetermined by internally generated input (i.e., from gravitational acceleration affecting the *maculae utriculi* and *sacculi* in the vestibular organ), is highly context-dependent: Its content is a possibility and not a reality. Just as the phenomenal properties of external experience are properties of virtual objects, so the properties exemplified in inner experience are those of a virtual subject. Its content is simply the best hypothesis about the current state of the system, given all constraints and information resources currently available. Interestingly – and this seems to be one of the core characteristics of phenomenal experience – this possibility is depicted *as* a reality, as an untranscendable presence (see section “Step 3: Transparency and Naive

Realism," below). The actuality of situated self-awareness is a *virtual* form of actuality.[8]

A second, even more vivid example of what is meant by the concept of a phenomenal self-model is demonstrated in Vilayanur Ramachandran's intriguing experiments on mirror-induced synesthesia and illusory movements in phantom limbs (see Ramachandran and Rogers-Ramachandran 1996; see also Ramachandran and Blakeslee 1998: 46pp). Phantom limbs are subjectively experienced limbs that typically remain after the loss of an arm or a hand as the result of an accident or surgical amputation (we will return to phantom limbs in the section "The Central Problem on the Functional Level of Description," below). In some situations, such as following a nontraumatic surgical amputation, patients are subjectively able to volitionally control and move their phantom limbs. The neurofunctional correlate of this phenomenal configuration could be that – since there is no contradictory feedback from the amputated limb – motor commands originating in motor cortex are still continuously monitored by the parietal lobes and thereby integrated into that part of the self-model which serves as an internal motor *emulator* (see Grush 1997, 1998: 174; Ramachandran and Rogers-Ramachandran 1996: 378; Metzinger 2001, chap. 7; for a recent review see Ramachandran and Hirstein 1998). In other situations, however, subjective motility and control over the phantom limb can be lost. Such a configuration may arise because of a preamputational paralysis due to peripheral

---

[8] I think that "virtual reality" is the best technological metaphor which is currently available as a source for generating new theoretical intuitions. In the context of this paper, heuristically the most interesting concept may

nerve lesions or to a prolonged absence of confirming proprioceptive and kinesthetic feedback. The result on the phenomenal level of representation is a paralyzed phantom limb.

Ramachandran and colleagues constructed a "virtual reality box" by placing a vertical mirror inside a cardboard box with the top of the box removed. Two holes in the front of the box enabled the patient to insert his real and his phantom arm. A patient suffering from a paralyzed phantom limb for many years was then asked to view the reflection of his normal hand in the mirror, thus – on the level of visual input – creating the illusion of observing two hands, when in fact he was seeing only the mirror reflection of the intact hand. What would happen to the content of the phenomenal self-model when the subject was asked to try to make bilateral, mirror-symmetric movements? Ramachandran describes one typical outcome of the experiment:

---

be that of "full immersion." Here, my own ideas converge strongly with those of Antti Revonsuo. See Revonsuo 1995; and chapter 4 in this volume.

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 arms 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, 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 so many 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.”
“Oh, yes, now it’s moving again.”[9]

[9] See Ramachandran and Blakeslee 1998: 47. For clinical and experimental details, see Ramachandran and Rogers-Ramachandran 1996. I am very grateful to Vilayanur Ramchandran for the color photograph inserted above.

By now it should have become clear how these data serve to illustrate the new concept I have introduced: What is moving in this experiment is the phenomenal self-model. The sudden occurrence of kinesthetic qualia in the degraded subspace of the self-model was made possible by installing a second and perfectly superimposed source of "virtual information," restoring, as it were, the visual mode of self-representation and thereby making this information volitionally available. Once again, this experiment nicely demonstrates how phenomenal properties are determined "from below," by functional and representational properties.

Step 2: Representationalist Analysis of the Target Properties

The Phenomenal Property of "Mineness" Now that the basic explanatory concept has been introduced, one can start developing an analysis of the target properties. Let us turn to the first target property, the phenomenal quality of "mineness." On a functional level of description, it is clear that it must be intimately associated with what is very likely the most fundamental partitioning of the neurocomputational state-space underlying conscious experience: the emergence of a self-world border. On the representational level of description there will be a simple and straightforward assumption that generates a large number of testable and more differentiated hypotheses. All representational states, which are being embedded into the currently active phenomenal self-model, gain the additional higher-order

property of phenomenal mineness. Mineness therefore is a prereflexive, nonconceptual sense of ownership (e.g., Martin 1995; Bermúdez 1998). If the underlying integration process is being disturbed, different neuropsychological syndromes or altered states of consciousness result. Here are some examples:

- Consciously experienced thoughts are not *my* thoughts any more: schizophrenia.
- My leg is not *my* leg any more: unilateral hemi-neglect.
- My arm acts without *my* control: alien hand syndrome.
- I am a robot; I am transformed into a puppet; volitional acts are not *my own* volitional acts anymore: depersonalization.[10]
- I *am* the whole world, all events in the world are being initiated and controlled by my own volitional acts: mania.

What could be a more complex phenomenon than conscious experience? An empirically founded philosophical theory will eventually have to do full justice to the depth and complex topology of our phenomenal state-space. Historical experience in science has shown that one of the most promising general research heuristics in understanding a complex phenomenon consists in analyzing borderline cases and impoverished variations of it. Nonstandard situations in complex domains help us in pointing out implicit assumptions and deficits in existing theories, as well as intuitive fallacies. Let us therefore briefly take a closer look at these examples. (I discuss more of such examples in Metzinger 1993, 1996, 2001).

---

[10] Depersonalization is here understood as the loss of phenomenal agenthood, namely, the specific form of phenomenal content that Karl Jaspers called "Vollzugsbewußtsein" (*executive consciousness*).

▪ Schizophrenia is, on the level of phenomenal content, usually characterized by phenomena such as thought insertion, verbal-auditory hallucinations, or delusions of control. Thought insertion can be representationally analyzed as a situation where the content of currently active cognitive states and processes can not be integrated into the self-model, and therefore must be represented as external. Nevertheless, such non-subjective thoughts are encoded as an element of objective reality, and their phenomenal presence (see table 20.1, rows 3 and 4) cannot be transcended by the experiential subject. If, as in hearing voices, internal speech production cannot be representationally integrated into the conscious self, a global phenomenal model of reality will emerge in which external voices are heard. The interesting phenomenon here seems to be the additional "pseudosensory" character of such hallucinations. Possibly they depend on an internal emulator for the generation of coherent speech acts, which, in the schizophrenic, transforms an efference copy of motor speech commands, used as input for an internal model of the ongoing process of speech production, into auditory format (Frith 1996; see also Grush 1997, 1998).

In regard to the third type of deficit classically associated with schizophrenia, according to the model here proposed, delusions of external control arise if the volitional acts preceding external motor behavior of the patient are no longer integrated into a phenomenal self-model. The attribution of such disconnected intentions to an external *person*, visible or invisible, may be a confabulatory reaction of the brain, still desperately trying to maximize the overall coherence of its model of reality. In those cases, where another person is experienced as the cause of one's own

bodily actions, this may simply be the most economical way to still represent such actions as caused by *mental* events (i.e., to preserve a personal-level representation of such events). If a circumscribed neural module for "theory-of-mind" ascriptions should exist, it is well conceivable that such a module suffers from functional dedifferentiation (see Daprati et al. 1997; Frith 1996; for a more detailed philosophical interpretation of empirical data, see chapter 21 in this volume).[11]

A self-model is important in enabling a system to represent itself *to* itself as an agent. If this system possesses the physical resources to activate abstract, allocentric representations of actions and action goals, then it also possesses one of the most important building blocks for language acquisition and social cognition (see chapter 22 in this volume). However, it also faces new computational tasks: It now has to reliably differentiate between own and foreign actions. It has to find a reliable way of integrating only the right subset of currently active goal and action representations into its self-model. Cortical activity in the representation of own and foreign actions clearly overlaps and, interestingly, non-schizophrenic subjects are not able to phenomenally represent the relevant signals generated by their own limb movements (Georgieff and Jeannerod 1998).

Action-relative information seems to be coded in two pathways in systems like ourselves, because the action representation via the phenomenal self-model does not

---

[11] Frith (1996: 1509) points out that attributions of heard voices to external *agents* are more strongly correlated to current delusions than to actually occurring hallucinations, and therefore are in some sense independent from the tendency to phenomenally represent self-generated events as external. McKenna, McKay, Law (1986) offer a case study and a number of considerations concerning the clinical concept of schizophrenia that may be helpful for philosophers.

depend on the same information that is used in generating automatic actions. In other words, the capacity to differentiate between first-person and third-person actions utilizing proprioceptive, endogenous signals is itself not localized on the phenomenal level, is not being carried out through accessing phenomenally available information. What actually enters into phenomenal self-experience may just be the global state of a comparator module (Georgieff and Jeannerod 1998). Therefore, two major classes of hallucinations exist: perceptions without objects and actions (including cognitive actions) without subjects. We do not yet know anything about the detailed mechanisms generating the corresponding classes of phenomenal states (the medial prefrontal area may constitute a major component of their neural correlate; see Frith 1996). However, it is conceivable that a functional dedifferentiation of the mechanism which integrates action representations into the self-model as *intended* or as *imagined* actions, versus leaving them in a more abstract, allocentric status as elements of the external world model (and therefore only as *observed* actions), leads to the pathological phenomenology in question.

The neural correlates of this type of deviant representational dynamics are slowly beginning to emerge. They seem to be related to hyperactivational states in the right inferior parietal lobule (Brodmann area 40) and the cingulate gyrus (e.g., Spence et al. 1997). Interestingly, what in scientific practice leads us from the representational analysis of such uncommon classes of phenomenal states to the delineation of their physical correlates, is imaging their functional correlates. An important aspect of these functional correlates seems to consist in disordered ways of making the

structure of external and internal bodily space available for the system and in deficits of the internal monitoring of ongoing motor acts (Spence et al. 1997).

- Alien hand syndrome (Goldstein 1908; Brion and Jedynak 1972) is characterized by a global experiential state in which the patient typically is well aware of complex, observable movements carried out by the nondominant hand, while at the same time experiencing no corresponding volitional acts. Subjectively (as well as functionally) the arm is "out of control," with a sense of intermanual conflict. On the other hand, many such arm movements clearly seem to be goal-directed actions, although no such goal representation is available either on the phenomenal level in general or on the level of conscious self-representation. Geschwind and his colleagues (1995) offer a case report of a 68-year-old woman suffering from a transient alien hand syndrome caused by a stroke limited only to the middle and posterior portions of the body of the corpus callosum:

> On postoperative day 11, she was noted by nursing staff to have left-sided weakness and difficulty walking. According to her family, she had complained of loss of control of her left hand for the previous three days, as if the hand were performing on its own. She awoke several times with her left hand choking her, and while she was awake, her left hand would unbutton her gown, crush cups on her tray, and fight with the right hand while she was answering the phone. To keep her left hand from doing mischief, she would subdue it with the right hand. She described this unpleasant situation as if someone "from the moon" were controlling her hand. (Geschwind et al. 1995: 803)[12]

In this case the functional correlate of representational shift is likely to have been an interhemispheric motor disconnection, whereas the neural correlate of this functional

[12] I am indebted to Andreas Kleinschmidt for useful advice with regard to relevant literature and possible interpretations of data.

deficit was a rather circumscribed lesion in the midbody of the corpus callosum. On the representational level we see that the triggering events leading to a certain subset of contradictory, but impressively complex and very obviously goal-directed, patterns of motor behavior, cannot be depicted as *my own* volitional acts anymore. In other words, the information about these events taking place within the system cannot be integrated into the phenomenal self-model. It is no longer globally available information for the system, either as a property of the world or as a property of the system itself. Therefore these action-generating events are – from the patient's perspective – not part of *her* phenomenal biography anymore. Only the visually and proprioceptively represented arm movements themselves are endowed with phenomenal subjectivity in the sense of ownership. They are, however, *not* subjective in the sense of phenomenal agenthood. Again, what is missing is a certain integrative capacity: the capacity to integrate a representation of the causal history of certain motor commands into the phenomenal self-model. (Note the implicit parallels to the discussions of schizophrenia in Daprati et al. 1997; Georgieff and Jeannerod 1998). On the functional level this loss is possibly mirrored in the loss of interhemispheric integration of motor and supplementary motor areas (Geschwind et al. 1995: 807).

- Hemi-neglect and other attention disorders seem to present an especially interesting case: You have an active self-representational structure, parts of which cannot be "read out" by higher-order attentional processes. Therefore you do not have a phenomenal self-model of this region in state-space; information already

contained in the system is not available under what the philosopher might want to call an internal context or an “ego-mode of presentation” (Newen 1997:117).

Attention is important in constituting phenomenal content. Bernard Baars and David Chalmers have repeatedly pointed out that “global availability” may be one of the central functional criteria to mark out active phenomenal information (see Baars 1988; Chalmers 1997). I think that given the material from psychophysics and neuropsychology, we need to differentiate this concept into at least three subcategories: *attentional* availability, *cognitive* availability, and *volitional* availability. In alien hand syndrome you seem to have a deficit that destroys volitional availability, but not attentional or cognitive availability. In hemi-neglect, however, attentional unavailability of information contained in an existing self-model can lead to a loss of cognitive availability (as in confabulatory activity and in anosognosia) and volitional availability (as in paralysis). There are of course generalized versions of this type of deviant phenomenal self-modeling.

One also has to do justice to situations in which certain layers of the self-model seem to be extended to the very border of the global model of reality.

- In some cases of mania or during certain religious experiences, the patient (or the mystic) is convinced that *all* events he experiences as taking place in the world are caused by his own volitional acts.

Philosophically speaking, the implicit ontology underlying these states is a Platonistic version of solipsism: All causation is mental causation, and there is only

one causal agent in the world.[13] From a representationalist perspective we clearly

seem to be confronted with a hypertrophy of self-representation: External events are endowed with phenomenal agenthood because their causal history is represented as an internal part of the system itself.[14] One may therefore speculate that the corresponding functional correlate must consist in a dedifferentiation of the integrational mechanism, which embeds some event representations into the currently active self-model while constantly excluding others. To my knowledge, nothing about the neural correlates realizing such a pathological function is known to date.

Let us take a look at the second target property.

The Property of Selfhood Phenomenal selfhood is what makes us an experiential subject. In German the property in question has sometimes been called *präreflexive Selbstvertrautheit* (prereflexive self-intimacy; e.g., Frank 1991). It is a very basic and seemingly spontaneous, effortless way of inner acquaintance, of "being in touch with yourself," a fundamental form of non-conceptual self-knowledge that precedes any higher forms of cognitive self-consciousness. In fact, this basic form of primitive self-awareness is what makes quasi-propositional and conceptually mediated forms of self-consciousness possible, by preventing them from becoming circular and empty (Bermúdez 1995; Metzinger 1993). From a representationalist perspective it clearly must be the result of an ongoing subsymbolic dynamics: the existence of a single, coherent, and temporally extended self-representation forming the center of the global representational state. The resulting *centeredness* of the overall representational

state, however, is a functional property (to which I will come back later). If the representational module I just mentioned is damaged or disintegrates, or if multiple structures of this type alternate in the system or are simultaneously active, different neuropsychological syndromes or altered states result.

Here are some brief examples:

- Anosognosias and anosodiaphorias: loss of higher-order insight into an existing deficit, as in blindness denial (Anton's Syndrome). This extensive and well-documented class of disorders is theoretically relevant, because it falsifies the "epistemic transparency" assumption, under which many classical theories of subjectivity have operated: There exist unnoticed and unnoticeable forms of mis(self-)representation, because large portions of the subsymbolic self-model seem to be cognitively and introspectively impenetrable (see also section "The Central Theoretical Problem on the Representational Level of Discription," below).
- Dissociative identity disorder (DID; for diagnostic features, see *DSM-IV*: 300.14). The system uses different and alternating self-models to functionally adapt to extremely traumatizing or socially inconsistent situations. On the phenomenal level this results in the presence of two or more distinct "identities" or personality states, and sometimes in asymmetric amnesias.

Since I cannot enter an extended discussion of this specific class of phenomenal states here,[15] I only want to draw readers' attention to two important points that demonstrate the explanatory power of the naturalized representationalist approach as opposed to classical philosophical theories of subjectivity. First, it is of course well

conceivable that a system generates a number of different self-models which are functionally incompatible, and therefore modularized. They nevertheless could be internally coherent, each endowed with its own characteristic phenomenal content and behavioral profile. Second, this does not have to be a pathological situation. Operating under different self-models in different situational contexts may be biologically as well as socially adaptive. Don't we all to some extent use multiple personalities to cope efficiently with different parts of our lives?

- *"Ich-Störungen"* [16] are a large class of psychiatric disorders accompanied by changes in the conscious experience of one's own identity. (For useful conceptual discussions and a number of case studies, see, e.g., Halligan and Marschall 1996.)

In these cases, the phenomenal self starts to disintegrate. Schizophrenia and DID are examples, as are depersonalization disorders. Self-models can lack information or become multiplied. They also can lose internal coherence. Phenomenological data from clinical neuropsychology and cognitive neuropsychiatry show that the internal correlation strength between the set of properties which is being depicted in the mode of phenomenal self-representation can vary greatly. If the phenomenal self-model, as I propose, at any given moment is a unified representation that can also dissolve or disintegrate, then there might of course be something like a *gradient of coherence* for this structure. In principle some metric for the internal coherence of a self-model should exist. Once the neural correlate of the phenomenal self-model in humans can be described in sufficient detail, this observation will constitute an interesting target for formal modeling.

The Property of Perspectivalness The third and last phenomenal target property is "perspectivalness": the existence of a single, coherent, and temporally stable model of reality that is representationally centered on a single, coherent, and temporally extended phenomenal subject. A phenomenal subject is a model of the system *as experiencing.*[17] To analyze perspectivalness, a second theoretical entity has to be introduced: "the phenomenal model of the intentionality-relation," i.e., an ongoing, dynamical representation of the system *as currently interacting with an object-component* (see section "Step 5" and chapter 7, this volume). This structural feature of the global representational space leads to the instantiation of a temporally extended, non-conceptual, first-person perspective. Again, if this global structural feature is lost, phenomenology changes and different neuropsychological disorders or classes of certain altered states emerge. Here are two last examples of situations in which conscious experience seems to remain while no longer being phenomenologically *subjective* experience. In these configurations the first-person perspective has been dissolved:

- Complete depersonalization (see *DSM-IV*: 300.6): loss of the phenomenal first-person perspective, accompanied by dysphoric states and functional deficits (*Angstvolle Ich-Auflösung,* "dreadful ego-dissolution"; see Dittrich 1985).
- Mystical states and religious experiences: selfless and noncentered global states that are later described and "autobiographically experienced" as nonpathological

and nonthreatening (*Ozeanische Selbstentgrenzung*, "oceanic boundary loss"; see Dittrich 1985; "The Great View from Nowhere").

The interesting insight at this point seems to be that there is in fact something like phenomenologically nonsubjective consciousness (see footnote 6). Obviously a full-blown theory of mind will have to do justice to the full spectrum of phenomenal states undergone by human beings. Mystical experiences as well as fully depersonalized pathological situations are important elements of this spectrum. The major epistemological obstacle in turning such states – be they pathological or spiritual – into explananda for neuroscientific research lies in the logical contradiction inherent in all reports from a purportedly *auto*biographical type of memory. The self-contradictory nature of such reports makes them a very doubtful source of information from a methodological perspective. I will now consider the two most important objections to the approach sketched in this paper.

The Central Theoretical Problem on the Functional Level of Description

The obvious criticism at this point is: This analysis does not do full justice to the phenomenology of self-consciousness, and it also cannot help us in understanding why a self-representation can be used as the immovable center of an egocentric world model in a purely functional sense. To take the phenomenology seriously means doing justice to the fact that the first-person perspective is always privileged among all other perspectives, which may be mentally represented in my conscious

space as well.[18] In what way does the phenomenal self-model differ from all other currently active phenomenal models, be they models of objects or models of other persons? Which functional property marks out its special role in the informational architecture of the system, and exactly how does it become the stable *center*, not only of phenomenal but of behavioral space as well?

Here is my answer: The self-model is the only representational structure that is anchored in the brain by a continuous source of *internally* generated input. Whenever conscious experience (i.e., the activation of a stable, integrated model of reality) takes place at all, this continual source of internal, proprioceptive input exists as well. The body is always there, and although its relational properties in space and in movement constantly change, the body is the only coherent perceptual object that *constantly* generates input. If one treats the many different internal sources of information flowing from tactile, visceral, vestibular, and other proprioceptors within the body as one single sensory modality, one arrives at an interesting conclusion: Body perception is unique not only from an epistemological, but also from a functional perspective, in that it has only one singular object ever (see Martin 1995; Bermúdez 1998: chapter. 6). And this elementary body percept will always contain a subvolume, which is unique in that it integrates a region of maximal invariance into phenomenal state-space, a region which is generated from an information flow originating exclusively in internal transducers and which is strictly stimulus-correlated. I call this stimulus-correlated part of the self-model a self-*presentation* (see Metzinger 1993, 1997), because it makes fine-grained information

available to the system that, according to my hypothesis, can be *presented*, but not *re-presented*.[19] The self-model becomes the functional center of representational as well as of behavioral space, because its purely presentational basis contributes to the only perceptual object that is permanently stimulus-correlated. Within that object we have something like a "background 'buzz' of somatosensory input" (Kinsbourne 1995: 217), which on the subcognitive level of phenomenal experience enables us to *feel ourselves* as continuously embodied and as present within a subjective now.

This answer to the functionalist question posed above immediately leads to testable hypotheses, because it makes the step from functional to neural correlates possible. If, e.g., it is really true that the constant activity of that part of the neuromatrix of the spatial model of one's own body, which is independent of external input, becomes the center of experiential space by forming an invariant background of bodily awareness, then this constitutes an empirical hypothesis. Of course, as a philosopher I should now definitely step back and refrain from any dilettante, amateurish speculation. However, let me draw my readers' attention to the fact that new results concerning research on pain experience in phantom limbs may point to the existence of a genetically determined neuromatrix, the input-independent activation pattern of which could form the functional basis of the most invariant partitions in the phenomenal body image ("phylomatrix of the body-schema"; see Melzack 1989, 1990, 1992; Melzak et al. 1997; for a possible more fundamental, but compatible approach see also Damasio 1994, 1999; Damasio and Damasio 1996a, b).

Of course, one might think that the elementary body percept is consolidated in social interactions only after birth, or during earlier motor behavior in the womb.[20] On the other hand, a persistent functional link between regions of primary somatosensory cortex and certain regions in the bodily self-model is proven by direct electrical stimulation during neurosurgical operations under local anesthesia (see Melzack et al. 1997). Of course, sensory body and motor maps are highly plastic and subject to the influence of experience even in the adult organism. And, of course, one has to see that there is probably no such thing as *absolute* invariance or functional rigidity. But there is good evidence for some kind of innate "body prototype," as can, for instance, be seen from the phantom sensations reported by some phocomelic children, who are born without one or more limbs. It seems that these data show that even people born without limbs develop complex bodily self-models which sometimes *include* limbs – even if there never has been a source of input.

Melzack's case studies provide convincing evidence that phantom limbs are experienced by at least 20 percent of congenitally limb-deficient subjects and by 50 percent of those who underwent amputations before the age of six years. If *all* of the congenital cases failed to have phantom experiences, it would be plausible that all self-modeling is only experientially based, but taken together with the fact that some of these patients do not lose the relevant parts of their phenomenal self-model even as adults, these cases seem to constitute plausible evidence that the neural correlate of the spatial model of the self is partly immune to local neuroplasticity in the somatosensory cortex. Melzack also points to numerous cases in which excision of

the somatosensory cortex did not prevent the reappearance of a phantom limb at follow-up. Therefore the neural correlate of self-consciousness (NCSC) must be highly distributed.[21] In an earlier publication, Melzack writes:

> In essence, I postulate that the brain contains a neuromatrix, or network of neurons, that, in addition to responding to sensory stimulation, continuously generates a characteristic pattern of impulses indicating that the body is intact and unequivocally one's own. I call this pattern a neurosignature. If such a matrix operated in the absence of sensory inputs from the periphery of the body, it would create the impression of having a limb even when that limb has been removed (1992: 93).

Again, I do not want to indulge in any amateurish speculation at this point. On the level of conceptual analysis my answer to the first problem is that the self-model is the only active representational structure in the system which is functionally anchored in a continuous, internally generated source of input. If this is correct, there should be many empirical routes to successfully take the step from the functional to the neural correlates of preconceptual self-consciousness.

The Central Theoretical Problem on the Representational Level of Description

Step 3: Transparency and Naive Realism

The antireductionist reply to the theoretical model sketched in this essay is obvious and straightforward. There seems to be no necessary connection between the functional and representational basis properties and the phenomenal target properties of “mineness,” “selfhood,” and “perspectivalness.” Everything described so far could, of course, happen without the instantiation of the phenomenal properties of “mineness,” “selfhood,” and “perspectivalness.” It is conceivable, a

property dualist might argue, that a biological information-processing system opens a centered representational space and then always embeds a model of itself into the model of reality active within this space *without* automatically generating a phenomenal self. An active, dynamical "self-model" still is just a representation of the system; it is a *system* model – not an instance of genuine self-consciousness. In order for the functional property of centeredness to contribute to the phenomenal property of perspectivalness, the model of the system has to become a phenomenal self. From a philosophical point of view the cardinal question is What is needed – by conceptual necessity – to make a phenomenal first-person perspective emerge from a representational space that is already functionally centered? In short, how do you get from the functional property of "centeredness" and the representational property of "self-modeling" to the phenomenal property of "selfhood"?

The answer lies in what one might call the "semantic transparency" of the data structures used by the system. Terminological details[22] aside, the general idea is that the representational vehicles[23] employed by the system are transparent in the sense that they do not contain the information *that* they are models on the level of their content (see Metzinger 1993; Van Gulick 1988a, 1988b). In our present context "transparency" means that: we are systems which are not able to recognize their own representational instruments *as* representational instruments. That is why we "look through" those representational structures, as if we were in direct and immediate contact with their content, with what they represent for us (see also row 2 of table 20.1).

Again, one may move downwards and speculate about certain functional properties of the internal instruments the system uses to represent the world and itself *to* itself. A simple functional hypothesis might say that the respective data structures are activated in such a fast and reliable way that the system itself is not able to recognize

them as such any more (e.g., because of a lower temporal resolution of metarepresentational processes; see, e.g., Metzinger 1995c). This can then be supplemented by a plausible teleofunctionalist assumption: For biological systems like ourselves – who always had to minimize computational load and find simple but viable solutions – naive realism was a functionally adequate "background assumption" to achieve reproductive success. In short, there has been no evolutionary pressure on our representational architecture to overcome the naive realism inherent in semantic transparency. The decisive step of my argument consists in applying this point to the self-model.

Step 4: Autoepistemic Boundedness

Let us take stock. So far we have taken three steps in our investigation: First, the self-model was introduced as a theoretical entity. Second, we make a brief representationalist analysis of the target properties possible. Third, we then introduced an empirically highly plausible assumption regarding the nature of many phenomenal representations, the transparency assumption. We now have two more steps, both of which are decisive. The first consists in applying the transparency assumption to the self-model and thereby solving the homunculus problem.

We are systems that are not able to recognize their subsymbolic self-model *as* a model. For this reason we are permanently operating under the conditions of a "naive-realistic self-misunderstanding": We experience ourselves as constantly being in direct and immediate epistemic contact with ourselves. What we have in the past simply called a „self" is not a non-physical individual, but only the content of an ongoing, dynamical process – the process of transparent self-modeling. Any system

that, because of its functional architecture, is not able to recognize its self-generated subconceptual representation of itself *as* a representation, will inevitably fall into a naive-realistic relationship toward the content of this representation.[24] On the representationalist level of analysis, this clearly seems to be a conceptual necessity. And as an empirical assumption about the way our brain actually works, it is highly plausible. A prereflexive phenomenal self, therefore, emerges if a system operates under a model of reality centered by a transparent self-model.

Step 5: The Phenomenal Model of the Intentionality Relation

The last step consists in applying the transparency constraint to the internal representation of the relation between subject and perceptual object, to the relation between agent and goal. If, for instance, the phenomenal model of one's own perceptual states contains a transparent representation of their causal history, then inevitably convolved global states will result, the content of which can only be truthfully described by the system itself as (e.g.) "I *myself* [= the content of a transparent self-model] am now seeing *this object* [= the content of a transparent object representation], and I am seeing it *with my own eyes*" [= the simple story about immediate sensory perception, which sufficed for the brain's evolutionary purposes]. The phenomenal self is a virtual agent perceiving virtual objects in a virtual world. This agent doesn't know that it possesses a visual cortex, and it does not know what electromagnetic radiation is: It just sees "with its own eyes" – by, as it were, effortlessly directing its visual attention. This virtual agent does not know that it possesses a motor system which, for instance, needs an internal emulator for fast,

goal-driven reaching movements. It just acts “with its own hands.” It doesn’t know what a sensorimotor loop is – it just effortlessly enjoys what researchers in the field of virtual reality design call “full immersion,” which for them is still a distant goal. To achieve this global effect, what is needed is a dynamic and transparent subject-object relation that episodically integrates the self-model and those perceptual objects which cause the changes in its content by telling an internal story about how these changes came about. This story does not have to be the true story; it may well be a greatly simplified internal confabulation that has proven to be functionally adequate.

Based on the arguments given above, I claim that phenomenal subjectivity emerges precisely at this stage: As soon as the system transparently models itself as an epistemic or causal agent, you have a transparent representation of episodic subject-object relations. For philosophers, of course, the new distinction of *phenomenal* intentionality as opposed to unconscious processes bearing intentional content will not be too surprising a move. It certainly is exciting that we presently witness this notion surfacing at the frontier of neuroscientific theory formation as well (see, e.g., Damasio 1994, 1999; Damasio and Damasio 1996a: 172, 1996b: 24; chapter 7, this volume; Delacour 1997: 138; LaBerge 1997: 150, 172).

Why would a concise research program for the neural correlate of self-consciousness (the NCSC) be of highest relevance for understanding phenomenal experience? If all the above is true (or if it at least points in the right direction), then it should prove to be more than heuristically fruitful. The vast majority of phenomenal states are *subjective* states in the way I have just analyzed: Not only are they elements of a

coherent internal model of reality used by the system; not only are they activated within a window of presence; not only does their phenomenal content supervene entirely on internal functional and physical properties; but they are bound into a transparently centered representational space. The maximally salient focus of conscious experience will always be constituted by the object-component of the phenomenal model of the intentionality-relation, with the subject-component, the self-model, providing a source of invariance and stability. If I am correct – and that is what it actually means when one says that such states are subjective states – then a straightforward empirical prediction will follow: Under standard conditions a very large class of phenomenal states should become episodically integrated with the current self-model on a very small time scale, as attention, as volition, as cognition wander around in representational space, selecting ever new object-components for the conscious first-person perspective. Global availability of information means availability for transient, dynamical integration into the currently active self-model, generating a "self in the act of knowing." In other words, the self-model theory of subjectivity can serve to mark out a specific and highly interesting class of neural correlates of consciousness.

And that is why the NCSC is important: Only if we find the neural and functional correlates of the phenomenal self will we be able to discover a more *general* theoretical framework into which all data can fit. Only then will we have a chance to understand what we are actually talking about when we say that phenomenal experience is a *subjective* phenomenon. It is for this reason that I have introduced two

new theoretical entities in this chapter, the notion of a “transparent self-model” and the concept of the “phenomenal model of the intentionality-relation.” Two predictions are associated with them. First, if – all other constraints held constant – the self-model of a conscious system would become fully *opaque*, then the phenomenal target property of experiential “selfhood” would disappear. Second, if the phenomenal model of the intentionality-relation collapses or cannot be sustained in a given conscious system, phenomenal states may exist, but will not be experientially *subjective* states any more, because the phenomenal first-person perspective has disappeared in this system. Intentionality-modeling is a necessary condition for perspectivalness.

In conclusion, let me once again illustrate the central thought of the argument by a metaphor. Interestingly, the point of this metaphor is that it contains a logical mistake: We are systems which were configured by evolution in such a way that they constantly *confuse* themselves with the content of their phenomenal self-model. In other words, we are physical systems that *on the level of phenomenal representation* are not able to differentiate between themselves and the content of their currently active self-model. We know ourselves only under a representation, and we are not able to *subjectively* represent this very fact. The evolutionary advantage of the underlying dynamical process of constantly confusing yourself with your own self-model is obvious: It makes a selfless biological system *egotistic* by generating a very robust self-illusion. Now here is the logical mistake: *Whose* illusion could that be? It makes sense to speak of truth and falsity, of knowledge and illusion, only if you already

have an epistemic agent in the sense of a system possessing conceptualized knowledge in a strong propositional sense. But this is not the case: We have just solved the homunculus problem; there is nobody in there who could be wrong about anything. All you have is a functionally grounded self-modeling system under the condition of a naive-realistic self-misunderstanding. So, if you would really want to carry this metaphor even further, what I have been saying in this paper is that the conscious self is an illusion which is *no one's* illusion.

Acknowledgements

I wish to thank William Banks, Patricia Churchland, Christof Koch, Francis Crick, Antonio Damasio, and Mirko von Elstermann for helpful comments and critical discussions of earlier versions of this paper. Saku Hara has greatly helped me with its revised version.

References

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

Bermúdez, J. L. (1995). Ecological perception and the notion of a nonceptual point of view. In Bermúdez et al. 1995.

Bermúdez, J. L. (1998). *The Paradox of Self-Consciousness*. Cambridge, MA: MIT Press.

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

Block, N., Flanagan, O., and Güzeldere, G., eds. (1997). *Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

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

Brion, S., and Jedynak, C.-P. (1972). Troubles du transfert interhémisphérique (callosal disconnection). A propos de trois observations de tumeurs du corps calleux. Le signe de la main étrangère. *Revue Neurologique* (Paris), 126, 257-266.

Cassam, Q. (1997). *Self and World*. Oxford: Clarendon Press.

Castañeda, H. N. (1966). >He<: A study on the logic of self-consciousness. *Ratio*, 8, 130-157.

Chalmers, D. J. (1997). Availability: The cognitive basis of experience? In Block et al. 1996.

Chisholm, R. M. (1981). *The First Person. An Essay on Reference an Intentionality*. Brighton.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind*, 95, 279-309. Reprinted in Churchland 1989.

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

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *Journal of Philosophy*, 65, 5-32.

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

Conant, R.C., and Ashby, W.R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 2, 89-97.

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.* New York: Hartcourt Brace and Company.

Damasio, A., and Damasio, H. (1996a). Images and subjectivity: Neurobiological trials and tribulations. In R. N. McCauley, ed., *The Churchlands and their Critics.* Cambridge, MA: Blackwell.

Damasio, A., and Damasio, H. (1996b). Making images and creating subjectivity. In R. Llinás and S. Churchland, eds., *The Mind-Brain Continuum*. Cambridge, MA: MIT Press.

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J. and Jeannerod, M. (1997). Looking for the agent: An investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition*, 65: 71-86.

David, A., Kemp, R., Smith, L., and Fahy, T. (1996). Split minds: Multiple personality and schizophrenia. In Halligan and Marshall 1996.

Delacour, J. (1997). Neurobiology of consciousness: An overview. *Behavioural Brain Research*, 85, 127-141.

Dennett, D. C. and Humphrey, N. (1989). Speaking for ourselves: An assessment of multiple personality disorder. *Raritan: A Quarterly Review* 9, no. 1: 68-98.

Diagnostic and Statistical Manual of Mental Disorders: DSM-IV. 4th ed. (1994). Prepared by the Task Force on DSM-IV and other committees and work groups of American Psychiatric Association Washington, DC: American Psychiatric Association.

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

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Frank, M. (1991). *Selbstbewußtsein und Selbsterkenntnis.* Stuttgart: Reclam.

Frith, C. (1996). The role of prefrontal cortex in self-consciousness: The case of auditory hallucinations. *Philosophical Transactions of the Royal Society of London* B351: 1505-1512.

Georgieff, N., and Jeannerod, M. (1998). Beyond consciousness of external reality: A "Who" system for consciousness and action and self-consciousness. *Consciousness and Cognition*, 7, 465-487.

Geschwind, D. H., Iacoboni, M., Mega, M. S., Zaidel, D. W. Clughesy, T. and Zaidel, E. (1995). Alien hand syndrome: Interhemispheric disconnection due to lesion in the midbody of the corpus callosum. *Neurology 45,* 802-808.

Goldstein, K. (1908). Zur Lehre der motorischen Apraxie. *Journal für Psychologie und Neurologie* 11, 169-187.

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

Grush, R. (1998). Wahrnehmung, Vorstellung, und die sensomotorische Schleife. In Heckmann, H.-D. and Esken, F. eds., *Bewußtsein und Repräsentation*. Paderborn: Schöningh.

Halligan, P.W. and Marshall, J.J. eds. (1996). *Method in Madness. Case Studies in Cognitive Neuropsychiatry*. Hove, UK: Psychology Press.

Harman, G. (1990). The intrinsic quality of experience. In J. Tomberlin ed., *Philosophical Perspectives*, vol. 4, *Action Theory and Philosophy of Mind*. Atascadero, CA: Ridgeview . Reprinted in Block et al. 1997.

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

LaBerge, D. (1997). Attention, awareness, and the triangular circuit. *Consciousness and Cognition* 6: 149-181.

Lewis, D. K. (1979). Attitudes *de dicto* and *de se*. *Philosophical Review* 88, 513-542.

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

Malcolm, N. (1988). Subjectivity. *Philosophy* 63: 147-160.

Marcel, A., and Bisiach, E., eds. (1998). *Consciousness in Contemporary Science*. Oxford: Oxford University Press.

Martin, M. G. F. (1995). Bodily Awareness: A sense of ownership. In Bermúdez et al. 1995.

McGinn, C. (1982). *The Character of Mind*. Oxford: Oxford University Press.

McKay, A. P., McKenna, P. J., and Laws, K. (1986). Severe schizophrenia: What is it like? In Halligan and Marshall 1996.

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, 90-96.

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). *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 and J. Nida-Rümelin, eds. (1997), *ANALYOMEN 2 – Perspektiven der Analytischen Philosophie*. Berlin: de Gruyter.

Metzinger, T. (1995b). Phänomenale mentale Modelle. In K. Sachs-Hombach ed., *Bilder im Geiste: Zur kognitiven und erkenntnistheoretischen Funktion piktorialer Repräsentationen*. Amsterdam: Rodopi.

Metzinger, T. (1995c). Faster than thought: Holism, homogeneity and temporal coding. In Metzinger 1995d.

Metzinger, T., ed. (1995d) *Conscious Experience*. Thorverton, UK: Imprint Academic; Paderborn: mentis.

Metzinger, T. (1998). Präsentationaler Gehalt. In Heckmann, H.-D. and Esken, F. eds., *Bewußtsein und Repräsentation*. Paderborn: mentis.

Metzinger, T. (forthcoming) *The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Miller, S. D., and Triggiano, P. J. (1992). The psychophysiological investigation of multiple personality disorder: Review and update. *American Journal of Clinical Hypnosis* 35: 47-61.

Moore, G. E. (1903). The refutation of idealism. *Mind* 12: 433-453.

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

Newen, A. (1997). The logic of indexical thoughts and the metaphysics of the „self.“ In W. Künne, A. Newen, and M. Anduschus, (eds), *Direct Reference, Indexicality and Propositional Attitudes*. Stanford: CSLI.

O’Shaughnessy, B. (1995). Proprioception and the body image. In Bermúdez et al. 1995.

Perry, J. (1993). *The Problem of the Essential Indexical and Other Essays*. Oxford: Oxford University Press.

Raffman, D. (1995). On the persistence of phenomenology. In Metzinger 1995b.

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

Ramachandran, V. S., and Hirstein, B. (1998). The perception of phantom limbs. The D.O. Hebb lecture. *Brain* 121: 1603-1630.

Ramachandran, V. S., and Rogers-Ramachandran, D. (1996). Synaestesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society London* B 263, 377-386.

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

Shoemaker, S. (1990). Qualities and qualia: What's in the mind? *Philosophy and Phenomenological Research Supplement* 50: 109-131.

Shoemaker, S. (1996). *The First-Person Perspective and Other Essays*. Cambridge, UK: Cambridge University Press.

Singer, W. (1993). Synchronization of cortical activity and its putative role in information processing and learning. *Annual Review of Physiology* 55: 349-384.

Singer, W. (1994). Putative functions of temporal correlations in neocortical processing. In Koch and Davis 1994.

Spence, S. A., Brooks, D. J., Hirsch, S. R., Liddle, P. F., Meehan, J., and Grasby, P. M. (1997). A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain* 120: 1997-2011.

Tye, M. (1991). *The Imagery Debate*. Cambridge, MA: MIT Press.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Van Gulick, R. (1988a). Consciousness, intrinsic intentionality, and self-understanding machines. In Marcel and Bisiach 1988.

Van Gulick, R. (1988b). A functionalist plea for self-consciousness. *Philosophical Review* 97: 149-188.

Walter, H. (1998). Emergence and the congnitive neuroscience approach to psychiatry. *Zeitschrift für Naturforschung* 53c: 723-737.

Walter, H. (2000). Emotionales Denken statt kalter Vernunft: Das Konzept des Selbst in der Neurophilosophie der Willensfreiheit. In A. Newen and K. Vogeley eds., *Das Selbst und seine neurobiologischen Grundlagen*. Paderborn: mentis.

Yates, J. (1975). The content of awareness is a model of the world. *Psychological Review* 92: 249-284.